Qual Life Res (2012) 21:659–670
DOI 10.1007/s11136-011-9965-9

Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review

Jasper M. Schellingerhout · Arianne P. Verhagen · Martijn W. Heymans · Bart W. Koes · Henrica C. de Vet · Caroline B. Terwee

Accepted: 17 June 2011 / Published online: 7 July 2011

Abstract
Purpose To critically appraise and compare the measurement properties of the original versions of neck-specific questionnaires.
Methods Bibliographic databases were searched for articles concerning the development or evaluation of the measurement properties of an original version of a self-reported questionnaire, evaluating pain and/or disability, which was specifically developed or adapted for patients with neck pain. The methodological quality of the selected studies and the results of the measurement properties were critically appraised and rated using a checklist, specifically designed for evaluating studies on measurement properties.
Results The search strategy resulted in a total of 3,641 unique hits, of which 25 articles, evaluating 8 different questionnaires, were included in our study. The Neck Disability Index is the most frequently evaluated questionnaire and shows positive results for internal consistency, content validity, structural validity, hypothesis testing, and responsiveness, but a negative result for reliability. The other questionnaires show positive results, but the evidence for each measurement property is mostly limited, and at least 50% of the information on measurement properties per questionnaire is lacking.
Conclusions Our findings imply that studies of high methodological quality are needed to properly assess the measurement properties of the currently available questionnaires. Until high quality studies are available, we recommend using these questionnaires with caution. There is no need for the development of new neck-specific questionnaires until the current questionnaires have been adequately assessed.

Keywords Neck pain · Neck disability · Questionnaire · Pain measurement · Validation studies · Reproducibility of results · Psychometrics · Systematic review

J. M. Schellingerhout (✉) · A. P. Verhagen · B. W. Koes
Department of General Practice, Erasmus Medical Centre, PO Box 2040, 3000 CA Rotterdam, The Netherlands
e-mail: j.schellingerhout@erasmusmc.nl

M. W. Heymans · H. C. de Vet · C. B. Terwee
Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands

M. W. Heymans
Department of Methodology and Applied Biostatistics, VU University, Amsterdam, The Netherlands

Introduction

Several disease-specific questionnaires have been developed to measure pain and/or disability in patients with neck pain (e.g., Neck Disability Index (NDI) and Neck Pain and Disability Scale (NPDS)) [1, 2]. In order to make a rational choice for the use of these questionnaires in clinical research and practice, it is important to assess and compare their measurement properties (e.g., reliability, validity, and responsiveness) [3].

A systematic review, published in 2002, evaluated the measurement properties of several neck-specific questionnaires and showed that, except for the NDI, all questionnaires were lacking psychometric information and that comparison was therefore not possible [4]. Recent reviews show that the amount of studies evaluating measurement properties of neck-specific questionnaires has extended considerably in the past years [5–7]. However, all these reviews lack an adequate instrument to critically appraise

the methodological quality of the included studies. Studies of high methodological quality are needed to guarantee appropriate conclusions about the measurement properties. Recently, the "COnsensus-based Standards for the selection of health status Measurement INstruments" (COSMIN) checklist, an instrument to evaluate the methodological quality of studies on measurement properties of health status questionnaires, has become available [8]. Using the COSMIN checklist, it is now possible to critically appraise and compare the quality of these studies.

A recent review of the cross-cultural adaptations of the McGill Pain Questionnaire showed that pooling of the measurement properties of different language versions results in inconsistent findings regarding the results for measurement properties, caused by differences in cultural context [9]. Since it is likely that the same accounts for the translated questionnaires in our review, we decided to evaluate them in a separate systematic review [10].

The purpose of this study is to critically appraise and compare the measurement properties of the original version of neck-specific questionnaires.

Methods

Search strategy

We searched the following computerized bibliographic databases: Medline (1966 to July 2010), EMbase (1974 to July 2010), CINAHL (1981 to July 2010), and PsycINFO (1806 to July 2010). We used the index terms "neck", "neck pain", and "neck injuries/injury" in combination with "research measurement", "questionnaire", "outcome assessment", "psychometry", "reliability", "validity" and derivatives of these terms. The full search strategy used in each database is available upon request from the authors. Reference lists were screened to identify additional relevant studies.

Selection criteria

A study was included if it was a full text original article (e.g., not an abstract, review or editorial), published in English, concerning the development or evaluation of the measurement properties of an original version of a neck-specific questionnaire. The questionnaire had to be self-reported, evaluating pain and/or disability, and specifically developed or adapted for patients with neck pain.

For inclusion, neck pain had to be the main complaint of the study population. Accompanying complaints (e.g., low back pain or shoulder pain) were no reason for exclusion, as long as the main focus was neck pain. Studies considering study populations with a specific neck disorder (e.g., neurologic disorder, rheumatologic disorder, malignancy, infection, or fracture) were excluded, except for patients with cervical radiculopathy or whiplash-associated disorder (WAD).

Two reviewers (JMS, APV) independently assessed the titles, abstracts, and reference lists of studies retrieved by the literature search. In case of disagreement between the two reviewers, there was discussion to reach consensus. If necessary, a third reviewer (HCV) made the decision regarding inclusion of the article.

Measurement properties

The measurement properties are divided over three domains: reliability, validity, and responsiveness [11]. In addition, the interpretability is described.

Reliability

Reliability is defined as the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g., using different sets of items from the same questionnaire (internal consistency); over time (test–retest); by different persons on the same occasion (inter-rater); or by the same persons on different occasions (intra-rater) [11].

Reliability contains the following measurement properties:

- *Internal consistency:* The interrelatedness among the items in a questionnaire, expressed by Cronbach's α or Kuder-Richardson Formula 20 (KR-20) [8, 11].
- *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, expressed by the standard error of measurement (SEM) [11, 12]. The SEM can be converted into the smallest detectable change (SDC) [12]. Changes exceeding the SDC can be labeled as change beyond measurement error [12]. Another approach is to calculate the limits of agreement (LoA) [13]. To determine the adequacy of measurement error, the smallest detectable change and/or limits of agreement is related to the minimal important change (MIC) [14]. As measurement error is expressed in the units of measurements, it is impossible to give one value for adequacy. However, it is important that the measurement error (i.e., noise, expressed as SDC or limits of agreement) is not larger than the signal (i.e., MIC) that one wants to assess.
- *Reliability:* The proportion of the total variance in the measurements, which is due to "true" differences between patients [11]. This aspect is reflected by the

intraclass correlation coefficient (ICC) or Cohen's Kappa [3, 11].

Validity

Validity is the extent to which a questionnaire measures the construct it is supposed to measure and contains the following measurement properties [11]:

- *Content validity:* The degree to which the content of a questionnaire is an adequate reflection of the construct to be measured [11]. Important aspects are whether all items are relevant for the construct, aim, and target population and if no important items are missing (comprehensiveness) [15].
- *Criterion validity:* The extent to which scores on an instrument are an adequate reflection of a gold standard [11]. Since a real gold standard for health status questionnaires is not available [15], we will not evaluate criterion validity.
- *Construct validity* is divided into three aspects:

- *Structural validity:* The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured [11]. Factor analysis should be performed to confirm the number of subscales present in a questionnaire [15].
- *Hypothesis testing:* The degree to which a particular measure relates to other measures in a way one would expect if it is validly measuring the supposed construct, i.e., in accordance with predefined hypotheses about the correlation or differences between the measures [11].
- *Cross-cultural validity:* The degree to which the performance of the items on a translated or culturally adapted instrument is an adequate reflection of the performance of the items of the original version of the instrument [11]. The cross-cultural validity of neck specificity questionnaire is addressed in a separate systematic review [10].

Responsiveness

Responsiveness is the ability of an instrument to detect change over time in the construct to be measured [11]. Responsiveness is considered an aspect of validity, in a longitudinal context [15]. Therefore, the same standards apply as for validity: the correlation between change scores of two measures should be in accordance with predefined hypotheses [15]. Another approach is to determine the area under the receiver operator characteristic curve (AUC) [15].

Interpretability

Interpretability is the degree to which one can assign qualitative meaning to quantitative scores [11]. This means that investigators should provide information about clinically meaningful differences in scores between subgroups, floor and ceiling effects, and the MIC [15]. Interpretability is not a measurement property, but an important characteristic of a measurement instrument [11].

Quality assessment

To determine whether the results of the included studies can be trusted, the methodological quality of the studies was assessed. This step was carried out using the COSMIN checklist [8]. The COSMIN checklist consists of nine boxes with 5–18 items concerning methodological standards for how each measurement property should be assessed. Each item was scored on a 4-point rating scale (i.e., "poor", "fair", "good", or "excellent"), which is an additional feature of the COSMIN checklist (see http://www.cosmin.nl). An overall score for the methodological quality of a study was determined for each measurement property separately, by taking the lowest rating of any of the items in a box. The methodological quality of a study was evaluated per measurement property.

Data extraction and assessment of (methodological) quality were performed by two reviewers (JMS, CBT) independently. In case of disagreement between the two reviewers, there was discussion in order to reach consensus. If necessary, a third reviewer (HCV) made the decision.

Best evidence synthesis: levels of evidence

To summarize all the evidence on the measurement properties of the different questionnaires, we synthesised the different studies by combining their results, taking the number and methodological quality of the studies and the consistency of their results into account. The possible overall rating for a measurement property is "positive", "indeterminate", or "negative", accompanied by levels of evidence, similarly as was proposed by the Cochrane Back Review Group (see Table 1) [16, 17].

To assess whether the results of the measurement properties were positive, negative, or indeterminate, we used criteria based on Terwee et al. (see Table 2) [18].

Results

The search strategy resulted in a total of 3,641 unique hits, of which 119 articles were selected based on their title and

Table 1 Levels of evidence for the overall quality of the measurement property [17]

Level	Rating	Criteria
Strong	+++ or −−−	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or −−	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or −	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodological quality

[..] reference number, + positive result, − negative result

abstract. The full text of these 119 articles was evaluated, which resulted in exclusion of another 68 articles. Reference checking did not result in additional included articles. Twenty-six articles concerned translated versions of neck-specific questionnaires, which were evaluated in a separate systematic review and therefore excluded [10]. Finally, 25 articles, evaluating 8 different questionnaires, were included in our study (see Fig. 1). All original versions were developed in English, except for the Copenhagen Neck Functional Disability Scale (CNFDS), which was originally developed in Danish. The general characteristics of these studies are presented in Table 3. Two studies evaluated measurement properties for different populations and are therefore mentioned twice in Table 3 [19, 20].

The methodological quality of the studies is presented in Table 4 for each questionnaire and measurement property. The synthesis of results per questionnaire and their accompanying level of evidence are presented in Table 5.

Below we will discuss the results per questionnaire. The results from studies of poor methodological quality are not mentioned [21–24].

Neck Disability Index (NDI)

The NDI was designed to measure activities of daily living (ADL) in patients with neck pain and was derived from the Oswestry low back pain Disability Index (ODI) [1, 25]. The 10 items have 6 response categories (range 0–5, total score range 0–50) [1]. We did not find studies evaluating the average time needed to fill out the English version of the NDI.

Exploratory factor analysis shows that there is moderate evidence that the NDI has a 1-factor structure [26], but there is also limited evidence that it is not unidimensional [27]. Both studies evaluating internal consistency assume a 1-factor structure, which resulted in a Cronbach α of 0.87–0.92 [26, 28]. The result of the only methodologically sound study evaluating measurement error is indeterminate [29], because information is needed on the MIC for judging the measurement error. A value for the MIC cannot be provided yet, as the estimates for the MIC are too diverse (i.e., 3.5, 7.5, and 9.5 on a 0–50 scale) [29–31]. There is limited evidence that reliability of the NDI is inadequate (ICC = 0.50) [29]. There is limited positive evidence for the content validity of the NDI [1]. Hypothesis testing shows that NDI has a positive correlation with instruments measuring pain and/or physical functioning (r = 0.53–0.70) [1, 26, 32, 33]. There is moderate positive evidence for responsiveness of the NDI (AUC = 0.79) [30]. Two studies of lower methodological quality confirm this positive finding [29, 34], and one study of lower quality reports a negative result (AUC = 0.57) [31]. Regarding interpretability, no floor or ceiling effects have been detected [1, 21, 28, 33], and differences in score between subgroups (e.g., same work status vs. altered work status) have been reported [30, 33].

Neck Pain and Disability Scale (NPDS)

The NPDS was designed to measure pain and disability in patients with neck pain and was developed using the Million Visual Analogue Scale as a template [2, 35]. It consists of 20 items, and each item is scored on a 10-cm visual analogue scale. Each item is converted to a score from 0 to 5 (total score range 0–100). We did not find studies evaluating the average time needed to fill out the English version of the NPDS.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, or content validity of the NPDS. Exploratory factor analysis shows a 4-factor structure for the NPDS [2]. There is limited positive evidence for hypothesis testing (r = 0.52–0.78) and responsiveness (r = 0.59) [2, 19]. No floor or ceiling effects have been detected [2, 21], and differences in scores between subgroups (neck pain vs. no pain vs. lower back pain) have been reported [2]. There is no information regarding the MIC.

Neck Bournemouth Questionnaire (NBQ)

The NBQ was designed to measure pain, physical functioning, social functioning, and psychologic functioning in patients with nonspecific neck pain and was developed using the Bournemouth Questionnaire for back pain as a template [36, 37]. It consists of 7 items, each scored on a

Table 2 Quality criteria for measurement properties [18]

Property	Rating	Quality criteria
Reliability		
Internal consistency	+	(Sub)scale unidimensional AND Cronbach's alpha(s) $\geq$ 0.70
	?	Dimensionality not known OR Cronbach's alpha not determined
	−	(Sub)scale not unidimensional OR Cronbach's alpha(s) $<$ 0.70
Measurement error	+	MIC $>$ SDC OR MIC outside the LOA
	?	MIC not defined
	−	MIC $\leq$ SDC OR MIC equals or inside LOA
Reliability	+	ICC/weighted Kappa $\geq$ 0.70 OR Pearson's $r \geq$ 0.80
	?	Neither ICC/weighted Kappa, nor Pearson's r determined
	−	ICC/weighted Kappa $<$ 0.70 OR Pearson's $r <$ 0.80
Validity		
Content validity	+	The target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete
	?	No target population involvement
	−	The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
Construct validity		
Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	−	Factors explain $<$ 50% of the variance
Hypothesis testing	+	(Correlation with an instrument measuring the same construct $\geq$ 0.50 OR at least 75% of the results are in accordance with the hypotheses) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	−	Correlation with an instrument measuring the same construct $<$ 0.50 OR $<$ 75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Responsiveness		
Responsiveness	+	(Correlation with an instrument measuring the same construct $\geq$ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC $\geq$ 0.70) AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	−	Correlation with an instrument measuring the same construct $<$ 0.50 OR $<$ 75% of the results are in accordance with the hypotheses OR AUC $<$ 0.70 OR correlation with related constructs is lower than with unrelated constructs

[..] reference number, *MIC* minimal important change, *SDC* smallest detectable change, *LOA* limits of agreement, *ICC* intraclass correlation coefficient, *AUC* area under the curve

+ positive rating, ? indeterminate rating, − negative rating

0–10 numerical scale (total score range: 0–70) [36]. We did not find studies evaluating the average time needed to fill out the English version of the NBQ.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, or structural validity of the NBQ. There is limited positive evidence for hypothesis testing ($r = 0.63$) and responsiveness ($r_{\text{items}} = 0.42$–0.82) [36]. Floor or ceiling effects, differences in scores between subgroups, and the MIC have not been studied.

Northwick Park Neck Pain Questionnaire (NPQ)

The NPQ was designed to measure the influence of non-specific neck pain on daily activities and was developed using the ODI as a template [25, 38]. Each of the nine items consists of five ordinal responses (scores 0–4), and the total (percentage) score is calculated by the following formula: (total score/maximum possible score) $\times$ 100% [38]. We did not find studies evaluating the average time needed to fill out the English version of the NPQ.

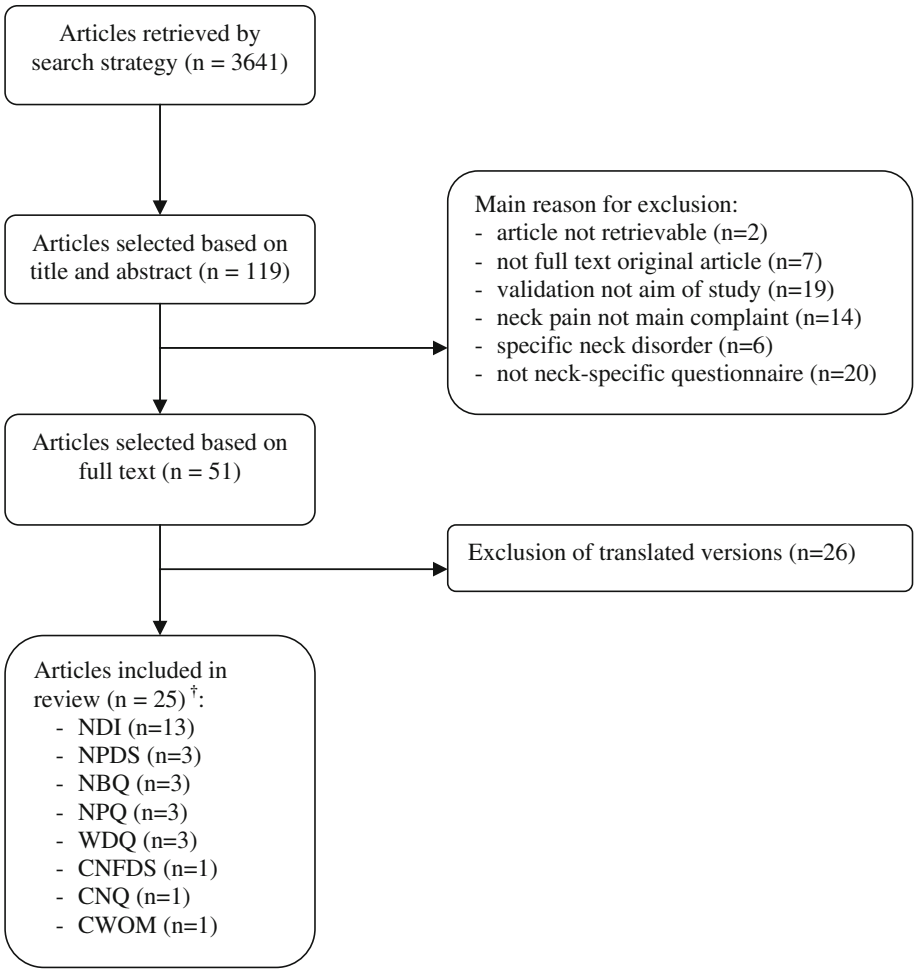

Fig. 1 Flowchart search and selection
† The sum of the different questionnaires is higher than 25, because some studies evaluate more than one questionnaire

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, or structural validity of the NPQ. There is a positive correlation ($r = 0.56$) between the NPQ and problem elicitation technique (PET) [32]. There is moderate positive evidence for responsiveness ($r = 0.60$) [38, 39]. No floor or ceiling effects have been detected [38, 39]. Differences in scores between subgroups have not been evaluated. The MIC is unclear, because the single study on this property does not quantify it [39].

Whiplash Disability Questionnaire (WDQ)

The WDQ was designed to measure disability in patients with WAD and was derived from the NDI [1, 40]. It consists of 13 items, each scored on a 0–10 numerical scale (total score range: 0–130) [40]. We did not find studies evaluating the average time needed to fill out the English version of the WDQ.

There were no methodologically sound studies evaluating the measurement error, reliability, content validity, or hypothesis testing of the WDQ. Exploratory factor analysis shows that the WDQ probably has a 1-factor structure [40]. There is moderate positive evidence for internal consistency (Cronbach's $\alpha = 0.95$–0.96) [40, 41]. These high values indicate that the WDQ might contain redundant items. There is limited positive evidence for responsiveness ($r = 0.67$) [42]. No floor or ceiling effects have been detected [40–42], information on other aspects of interpretability is lacking.

Copenhagen Neck Functional Disability Scale (CNFDS)

The CNFDS was designed by a group of experts in the field of neck pain to measure disability in patients with neck pain [20]. It consists of 15 items with three possible ordinal responses per item (score 0–2). The total score ranges from

Table 3 Characteristics of the included studies

Study	Population	Country	Setting
Bolton et al. [36]	Non-specific neck pain	England	Chiropractor
Bolton et al. [37]	Non-specific neck pain	England	Chiropractor
Chan Ci En et al. [21]	>3 month nontraumatic neck pain	Australia	Physiotherapist
Chok et al. [22]	Neck pain	Singapore	Physiotherapist
Cleland et al. [29]	Non-specific neck pain	USA	Physiotherapist
Cleland et al. [31]	Cervical radiculopathy	USA	Physiotherapist
Ferrari et al. [41]	Motor vehicle collision victims	Canada	Primary care
Gay et al. [23]	Chronic, uncomplicated neck pain	USA	Physiotherapist
Goolkasian et al. [19]†	Mechanical neck pain	USA	Orthopedist
Goolkasian et al. [19]†	Chronic mechanical neck pain	USA	Orthopedist
Hains et al. [26]	Neck pain	Canada	Chiropractor
Hoving et al. [32]	WAD	Australia	Physiotherapist/GP/rheumatology
Jordan et al. [20]†	Chronic mechanical neck pain	Denmark	Primary care
Jordan et al. [20]†	Chronic mechanical neck pain	Denmark	Physiotherapist
Leak et al. [38]	Mechanical neck pain	England	Rheumatologist
Pinfold et al. [40]	WAD	Australia	Physiotherapist
Rebbeck et al. [45]	WAD	Australia	Primary care/insurance cohort
Riddle et al. [33]	Non-specific neck pain	USA	Physiotherapist
Sim et al. [39]	Non-specific neck pain	England	Physiotherapist
Stewart et al. [34]	>3 month whiplash	Australia	Physiotherapist
Stratford et al. [28]	Neck pain of suspected musculoskeletal origin	Canada/USA	Physiotherapist
van der Velde et al. [27]	Mechanical neck pain	USA	General population/chiropractor
Vernon et al. [1]	WAD or chronic nontraumatic neck complaints	England	Chiropractor
Wheeler et al. [2]	Mechanical neck pain	USA	Orthopedist
White et al. [43]	Chronic mechanical neck pain	England	Physiotherapist/rheumatologist
Willis et al. [42]	WAD	Australia	Physiotherapist
Young et al. [30]	Mechanical neck pain	USA	Physiotherapist

[..] reference number, *GP* general practitioner, *WAD* whiplash associated disorder

† Study is mentioned twice, because they evaluated a questionnaire in two different populations

0 to 30 [20]. The average time needed to fill out the Danish version of the CNFDS is 10 min [20].

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, structural validity, or responsiveness of the CNFDS. The CNFDS correlates with a Numerical Rating Scale (NRS) for pain ($r = 0.64$) [20]. There are no comparisons available between the CNFDS and other instruments measuring pain or disability. No ceiling effect has been detected [20]; there is no information available on floor effects, differences in scores between subgroups, or the MIC.

Core Neck Questionnaire (CNQ)

The CNQ was designed to measure outcomes of care in patients with nonspecific neck pain and was developed using the core outcome measure for back pain (COM) as a template [43, 44]. The CNQ consists of seven items, scored from 1 to 5, which are added up to a total score [43]. We did not find studies evaluating the average time needed to fill out the English version of the CNQ.

There were no methodologically sound studies evaluating the internal consistency, measurement error, content validity, structural validity, or responsiveness of the CNFDS. The reliability of the total score of the CNQ has not been studied, but four of the six items have an ICC > 0.70 [43]. There was a positive correlation of the CNQ with the NDI ($r > 0.60$) [43]. No floor or ceiling effects have been detected [43]; there is no information on other aspects of interpretability.

Core Whiplash Outcome Measure (CWOM)

The CWOM was designed to measure relevant health outcomes in patients with whiplash associated disorder (WAD) [45]. The COM was used as a template to develop

Table 4 Methodological quality of each study per measurement property and questionnaire

Study	Internal consistency	Measurement error	Reliability	Content validity	Structural validity	Hypotheses testing	Responsiveness
NDI							
Chan Ci En et al. [21]				Poor		Poor	
Chok et al. [22]			Poor				Poor
Cleland et al. [29]		Fair	Fair				Fair
Cleland et al. [31]		Poor	Poor				Fair
Gay et al. [23]	Poor	Poor				Poor	Poor
Hains et al. [26]	Excellent				Good	Good	
Hoving et al. [32]				Poor		Fair	
Riddle et al. [33]						Good	Poor
Stewart et al. [34]							Fair
Stratford et al. [26]	Fair	Poor	Poor				Poor
van der Velde et al. [27]	Fair				Fair	Poor	
Vernon et al. [1]	Poor		Poor	Fair		Fair	Poor
Young et al. [30]			Poor				Good
NPDS							
Chan Ci En et al. [21]				Poor		Poor	
Goolkasian et al.-1 [19]			Poor				
Goolkasian et al.-2 [19]			Poor				Fair
Wheeler et al. [2]	Poor			Poor	Fair	Fair	
NBQ							
Bolton et al. [36]							Poor
Bolton et al. [37]	Poor	Poor	Poor			Fair	Fair
Gay et al. [23]	Poor	Poor				Poor	Poor
NPQ							
Hoving et al. [32]				Poor		Fair	
Leak et al. [38]	Poor	Poor	Poor	Poor			Fair
Sim et al. [39]	Poor						Fair
WDQ							
Ferrari et al. [41]	Fair					Poor	
Pinfold et al. [40]	Good			Poor	Fair		
Willis et al. [42]		Poor	Poor				Fair
CNFDS							
Jordan et al.-1 [20]	Poor		Poor	Poor			
Jordan et al.-2 [20]				Poor		Fair	Poor
CNQ							
White et al. [43]		Fair		Poor		Fair	
CWQ							
Rebbeck et al. [45]	Poor				Fair		Fair

[..] reference number

the CWOM [44, 45]. The CWOM consists of 5 items, each scored on a 1–5 scale (total score range: 5–25) [45]. We did not find studies evaluating the average time needed to fill out the English version of the CWOM.

There were no methodologically sound studies evaluating the internal consistency, measurement error, reliability, content validity, or structural validity of the CWOM. There is limited positive evidence for correlation

Table 5 Quality of measurement properties per questionnaire

Questionnaire	Internal consistency	Measurement error	Reliability	Content validity	Structural validity	Hypothesis testing	Responsiveness
NDI	+++	?	−	+	++	+++	++
NPDS	?	na	?	?	+	+	+
NBQ	?	?	?	na	na	+	+
NPQ	?	?	?	?	na	+	++
WDQ	++	?	?	?	+	?	+
CNFDS	?	na	?	?	na	+	?
CNQ	na	na	+	?	na	+	na
CWOM	?	na	na	na	na	+	+

+++ or −−− strong evidence positive/negative result, ++ or −− moderate evidence positive/negative result, + or − limited evidence positive/negative result, ± conflicting evidence, ? unknown, due to poor methodological quality, *na* no information available

with instruments measuring pain and/or physical functioning ($r = 0.65$–0.82) and for responsiveness (AUC = 0.73–0.81) [45]. The scores for different stages of whiplash have been reported [45], but other aspects of interpretability are not mentioned.

Discussion

Eight different questionnaires have been developed to measure pain and/or disability in patients with neck pain. All original versions are in English, except for the CNFDS, which was developed in Danish. The NDI is the most frequently evaluated questionnaire and its measurement properties seem adequate, except for reliability. The other questionnaires show positive results, but the evidence is mostly limited and at least half of the information on measurement properties per questionnaire is lacking. Therefore, the results should be treated with caution.

The COSMIN checklist has recently been developed and is based on consensus between experts in the field of health status questionnaires [8]. The COSMIN checklist facilitates a separate judgment of the methodological quality of the included studies and their results. This is in line with the methodology of systematic reviews of clinical trials [16]. The inter-rater agreement of the COSMIN checklist is adequate [46]. The inter-rater reliability for many COSMIN items is poor, which is suggested to be due to interpretation of checklist items [46]. To minimize differences between reviewers (JMS, CBT, and HCV) in interpretation of checklist items, decisions were made in advance on how to score the different items.

The criteria in Table 1 are based on the levels of evidence as previously proposed by the Cochrane Back Review Group [17]. The criteria are originally meant for systematic reviews of clinical trials, but we believe that they are also applicable for reviews on measurement properties of health status questionnaires.

Exclusion of non-English papers may introduce selection bias. However, the leading journals, and as a consequence the most important studies, are published in English. So, research performed in populations with a different native language is generally still published in English. This is illustrated by the large number of articles we retrieved regarding translations of neck-specific questionnaires (see Fig. 1). In these papers, we did not find a reference to an original version of a neck-specific questionnaire that was not included in our systematic review. This makes us confident that chances are small that we have missed any original versions of neck-specific questionnaires.

The different studies showed similar methodological shortcomings. A small sample size, for example, frequently led to indeterminate results. We do not discuss these flaws in detail here but elaborate on this subject in a separate publication [47].

A problem we encountered during the rating of "hypothesis testing" and "responsiveness" was that most studies do not formulate hypotheses regarding expected correlations in advance. Moreover, none of the development studies specified the supposed underlying constructs of the questionnaire. Therefore, it is difficult to judge content validity, which is one of the most important measurement properties. We dealt with this problem by reaching agreement about what we thought were the supposed underlying constructs, based on the items in the questionnaire, before we rated the studies.

The assumption that pooling of results from original and translated versions could result in inconsistent findings regarding the results for measurement properties is confirmed in our systematic review of translated versions of neck-specific questionnaires [10]. A poor translation process and/or lack of cross-cultural validation seem to affect the measurement properties of the questionnaire, particularly the validity (i.e., structural validity and hypothesis testing) [10]. This is not surprising, as the importance and/or meaning of

questionnaire items (e.g., driving, depressed mood) may depend on setting and context. So, a simple translation of the original questionnaire is not sufficient and might affect the measurement of the underlying constructs [10].

Since the review in 2002, 17 of the 25 included studies in our review were published, and four new neck-specific questionnaires have been developed [4, 36, 40, 43, 45]. These studies added new information, but due to their poor to fair methodological quality, a substantial amount of uncertainty about the quality of the measurement properties remains.

The quality of the measurement properties of several neck-specific questionnaires was recently evaluated in a best evidence synthesis, which showed positive results for the NDI, NPDS, NBQ, NPQ, CNFDS, and WDQ [5]. However, these results were partially based on methodologically flawed studies and this study contained only a small part of the manuscripts included in our study.

A state-of-the-art review evaluating the NDI reported that its reliability, internal consistency, factor structure (i.e., unidimensional scale), construct validity, and responsiveness are well described and of very high quality [7], which is not completely in agreement with our findings. Possible explanations for the discrepancies are that the study reporting the negative result for reliability was published after the search of the state-of-the-art review ended and that they did not critically appraise the methodological quality or results of the included studies [7, 29]. A more recent systematic review evaluating the NDI reports a good internal consistency, acceptable reliability, good construct validity and responsiveness, and inconsistent results regarding the structural validity of the NDI [6]. The differences with our findings are probably attributable to the fact that they did not take the methodological quality of the included studies into account [6].

It is difficult to determine the content validity of the different neck-specific questionnaires, because almost all retrieved studies on this subject were of poor methodological quality. Furthermore, the underlying constructs were not clear. However, a recent content analysis showed that correspondence between the symptoms expressed by neck pain patients and the content of the questionnaires was low, mainly due to lack of patient involvement in development of the questionnaire [48]. The importance of content validity for a questionnaire makes it desirable that this measurement property is evaluated in a high quality study for each questionnaire. The results from these studies will show which questionnaires are suitable for neck pain patients and whether development of a new neck-specific questionnaire is necessary.

The most frequently studied measurement property is responsiveness. This is not surprising, since these questionnaires are often used as an outcome measure. However, except for the NDI and NPQ, there is only limited positive evidence for responsiveness.

For clinical practice and research, we advise to use the original version of neck-specific questionnaires with caution: the majority of the results are positive, but the evidence is mostly limited and for each questionnaire, except for the NDI, at least half of the information regarding measurement properties is lacking. Provisionally, we recommend using the NDI, because it is the questionnaire for which the most information is available and the results are mostly positive. However, research is needed to clarify its underlying constructs, measurement error, reliability, and to improve the interpretation of its scores.

No clinician should make decisions regarding management of neck pain patients solely on unvalidated instruments. However, neck-specific questionnaires can provide a broader and deeper understanding of the impact of neck pain on the individual patients.

For future research, we recommend performing high quality studies to evaluate the unknown measurement properties, especially content validity, and provide strong evidence for the other measurement properties. It seems advisable to refrain from developing new neck-specific questionnaires until high quality studies evaluating the measurement properties of current questionnaires show shortcomings that make it necessary to develop a new questionnaire.

Conclusion

A lot of information regarding the measurement properties of the original version of the different neck-specific questionnaires is still lacking or of poor methodological quality. The available evidence on the measurement properties is mostly limited. The NDI is the most frequently evaluated questionnaire, and its measurement properties seem adequate, except for reliability and the fact that there is information lacking regarding its underlying constructs and measurement error.

Our findings do not mean that the current questionnaires are poor but imply that studies of high methodological quality are needed to properly assess their measurement properties. It is recommendable to use the COSMIN checklist when designing these studies. There is no need for the development of new neck-specific questionnaires until the measurement properties of the current questionnaires have been adequately assessed.

Acknowledgments All authors declare that they have no conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject of the manuscript. No financial support was received.

References

1. Vernon, H., & Mior, S. (1991). The Neck Disability Index: a study of reliability and validity. *Journal of Manipulative and Physiological Therapeutics, 14*(7), 409–415.
2. Wheeler, A. H., Goolkasian, P., Baird, A. C., & Darden, B. V. (1999). Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. *Spine, 24*(13), 1290–1294.
3. Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.
4. Pietrobon, R., Coeytaux, R. R., Carey, T. S., Richardson, W. J., & DeVellis, R. F. (2002). Standard scales for measurement of functional outcome for cervical pain or dysfunction: A systematic review. *Spine, 27*(5), 515–522.
5. Nordin, M., Carragee, E. J., Hogg-Johnson, S., Weiner, S. S., Hurwitz, E. L., Peloso, P. M., et al. (2008). Assessment of neck pain and its associated disorders: Results of the bone and joint decade 2000–2010 task force on neck pain and its associated disorders. *Spine, 33*(4 Suppl), S101–122.
6. MacDermid, J. C., Walton, D. M., Avery, S., Blanchard, A., Etruw, E., McAlpine, C., et al. (2009). Measurement properties of the Neck Disability Index: A systematic review. *Journal of Orthopaedic and Sports Physical Therapy, 39*(5), 400–417.
7. Vernon, H. (2008). The Neck Disability Index: State-of-the-art, 1991–2008. *Journal of Manipulative and Physiological Therapeutics, 31*(7), 491–502.
8. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549.
9. Menezes Costa Lda, C., Maher, C. G., McAuley, J. H., & Costa, L. O. (2009). Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. *Journal of Clinical Epidemiology, 62*(9), 934–943.
10. Schellingerhout, J. M., Heymans, M. W., Verhagen, A. P., de Vet, H. C., Koes, B. W., & Terwee, C. B. (2011). Measurement properties of translations of disease-specific questionnaires in patients with neck pain: A systematic review. *BMC Medical Research Methodology, 11*, 87.
11. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745.
12. de Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033–1039.
13. Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*(8476), 307–310.
14. Terwee, C. B., Roorda, L. D., Knol, D. L., De Boer, M. R., & De Vet, H. C. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology, 62*(10), 1062–1067.
15. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2009). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10*, 22.
16. Furlan, A. D., Pennick, V., Bombardier, C., van Tulder, M., & Editorial Board, C. B. R. G. (2009). 2009 updated method guidelines for systematic reviews in the Cochrane back review group. *Spine, 34*(18), 1929–1941.
17. van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., Board, Editorial., & k Review, G. (2003). Updated method guidelines for systematic reviews in the Cochrane collaboration back review group. *Spine, 28*(12), 1290–1299.
18. Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42.
19. Goolkasian, P., Wheeler, A. H., & Gretz, S. S. (2002). The neck pain and disability scale: Test–retest reliability and construct validity. *Clinical Journal of Pain, 18*(4), 245–250.
20. Jordan, A., Manniche, C., Mosdal, C., & Hindsberger, C. (1998). The Copenhagen neck functional disability scale: A study of reliability and validity. *Journal of Manipulative and Physiological Therapeutics, 21*(8), 520–527.
21. Chan Ci En, M., Clair, D. A., & Edmondston, S. J. (2009). Validity of the Neck Disability Index and neck pain and disability scale for measuring disability associated with chronic, non-traumatic neck pain. *Manual Therapy, 14*(4), 433–438.
22. Chok, B., & Gomez, E. (2000). The reliability and application of the Neck Disability Index in physiotherapy. *Physiotherapy Singapore, 3*, 16–19.
23. Gay, R. E., Madson, T. J., & Cieslak, K. R. (2007). Comparison of the Neck Disability Index and the Neck Bournemouth Questionnaire in a sample of patients with chronic uncomplicated neck pain. *Journal of Manipulative and Physiological Therapeutics, 30*(4), 259–262.
24. Bolton, J. E. (2004). Sensitivity and specificity of outcome measures in patients with neck pain: Detecting clinically significant improvement. *Spine, 29*(21), 2410–2417. discussion 2418.
25. Fairbank, J. C., Couper, J., Davies, J. B., & O'Brien, J. P. (1980). The Oswestry low back pain disability questionnaire. *Physiotherapy, 66*(8), 271–273.
26. Hains, F., Waalen, J., & Mior, S. (1998). Psychometric properties of the Neck Disability Index. *Journal of Manipulative and Physiological Therapeutics, 21*(2), 75–80.
27. van der Velde, G., Beaton, D., Hogg-Johnston, S., Hurwitz, E., & Tennant, A. (2009). Rasch analysis provides new insights into the measurement properties of the Neck Disability Index. *Arthritis and Rheumatism, 61*(4), 544–551.
28. Stratford, P., Riddle, D., Binkley, J., Spadoni, G., Westaway, M., & Padfield, B. (1999). Using the Neck Disability Index to make decisions concerning individual patients. *Physiotherapy Canada, 51*(2), 107.
29. Cleland, J. A., Childs, J. D., & Whitman, J. M. (2008). Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. *Archives of Physical Medicine and Rehabilitation, 89*(1), 69–74.
30. Young, B. A., Walker, M. J., Strunce, J. B., Boyles, R. E., Whitman, J. M., & Childs, J. D. (2009). Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *Spine Journal, 9*(10), 802–808.
31. Cleland, J. A., Fritz, J. M., Whitman, J. M., & Palmer, J. A. (2006). The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. *Spine, 31*(5), 598–602.

32. Hoving, J. L., O'Leary, E. F., Niere, K. R., Green, S., & Buchbinder, R. (2003). Validity of the neck disability index, Northwick Park neck pain questionnaire, and problem elicitation technique for measuring disability associated with whiplash-associated disorders. *Pain, 102*(3), 273–281.
33. Riddle, D. L., & Stratford, P. W. (1998). Use of generic versus region-specific functional status measures on patients with cervical spine disorders. *Physical Therapy, 78*(9), 951–963.
34. Stewart, M., Maher, C. G., Refshauge, K. M., Bogduk, N., & Nicholas, M. (2007). Responsiveness of pain and disability measures for chronic whiplash. *Spine, 32*(5), 580–585.
35. Million, R., Nilsen, K. H., Jayson, M. I., & Baker, R. D. (1981). Evaluation of low back pain and assessment of lumbar corsets with and without back supports. *Annals of the Rheumatic Diseases, 40*(5), 449–454.
36. Bolton, J. E., & Humphreys, B. K. (2002). The Bournemouth Questionnaire: A short-form comprehensive outcome measure. II. Psychometric properties in neck pain patients. *Journal of Manipulative and Physiological Therapeutics, 25*(3), 141–148.
37. Bolton, J. E., & Breen, A. C. (1999). The Bournemouth Questionnaire: A short-form comprehensive outcome measure. I. Psychometric properties in back pain patients. *Journal of Manipulative and Physiological Therapeutics, 22*(8), 503–510.
38. Leak, A. M., Cooper, J., Dyer, S., Williams, K. A., Turner-Stokes, L., & Frank, A. O. (1994). The Northwick Park Neck Pain Questionnaire, devised to measure neck pain and disability. *British Journal of Rheumatology, 33*(5), 469–474.
39. Sim, J., Jordan, K., Lewis, M., Hill, J., Hay, E. M., & Dziedzic, K. (2006). Sensitivity to change and internal consistency of the Northwick Park Neck Pain Questionnaire and derivation of a minimal clinically important difference. *Clinical Journal of Pain, 22*(9), 820–826.
40. Pinfold, M., Niere, K. R., O'Leary, E. F., Hoving, J. L., Green, S., & Buchbinder, R. (2004). Validity and internal consistency of a Whiplash-specific disability measure. *Spine, 29*(3), 263–268.
41. Ferrari, R., Russell, A., & Kelly, A. J. (2006). Assessing whiplash recovery—the Whiplash Disability Questionnaire. *Australian Family Physician, 35*(8), 653–654.
42. Willis, C., Niere, K. R., Hoving, J. L., Green, S., O'Leary, E. F., & Buchbinder, R. (2004). Reproducibility and responsiveness of the Whiplash Disability Questionnaire. *Pain, 110*(3), 681–688.
43. White, P., Lewith, G., & Prescott, P. (2004). The core outcomes for neck pain: Validation of a new outcome measure. *Spine, 29*(17), 1923–1930.
44. Deyo, R. A., Battie, M., Beurskens, A. J., Bombardier, C., Croft, P., Koes, B., et al. (1998). Outcome measures for low back pain research. A proposal for standardized use. *Spine, 23*(18), 2003–2013.
45. Rebbeck, T. J., Refshauge, K. M., Maher, C. G., & Stewart, M. (2007). Evaluation of the core outcome measure in whiplash. *Spine, 32*(6), 696–702.
46. Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). Inter-rater agreement and reliability of the COSMIN (consensus-based standards for the selection of health status measurement instruments) checklist. *BMC Medical Research Methodology*, *10*, 82.
47. Terwee, C. B., Schellingerhout, J. M., Verhagen, A. P., de Vet, H. C., & Koes, B. W. (2011). Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: A systematic review. *Journal of Manipulative and Physiological Therapeutics*, *34*(4), 261–272.
48. Wiitavaara, B., Bjorklund, M., Brulin, C., & Djupsjobacka, M. (2009). How well do questionnaires on symptoms in neck-shoulder disorders capture the experiences of those who suffer from neck-shoulder disorders? A content analysis of questionnaires and interviews. *BMC Musculoskeletal Disorders, 10*, 30.