Amsterdam University of Applied Sciences

The diagnostic accuracy of headache measurement instruments
a systematic review and meta-analysis focusing on headaches associated with musculoskeletal symptoms

van der Meer, Hedwig A; Visscher, Corine M; Vredeveld, Tom; Nijhuis van der Sanden, Maria Wg; Hh Engelbert, Raoul; Speksnijder, Caroline M

Link to publication

The diagnostic accuracy of headache measurement instruments: A systematic review and meta-analysis focusing on headaches associated with musculoskeletal symptoms

Hedwig A van der Meer[1,2,3,4,5,6] , Corine M Visscher[2],
Tom Vredeveld[3], Maria WG Nijhuis van der Sanden[4] ,
Raoul HH Engelbert[3,5] and Caroline M Speksnijder[6]

Abstract

Aim: To systematically review the available literature on the diagnostic accuracy of questionnaires and measurement instruments for headaches associated with musculoskeletal symptoms.

Design: Articles were eligible for inclusion when the diagnostic accuracy (sensitivity/specificity) was established for measurement instruments for headaches associated with musculoskeletal symptoms in an adult population. The databases searched were PubMed (1966–2018), Cochrane (1898–2018) and Cinahl (1988–2018). Methodological quality was assessed with the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2) and COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist for criterion validity. When possible, a meta-analysis was performed. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) recommendations were applied to establish the level of evidence per measurement instrument.

Results: From 3450 articles identified, 31 articles were included in this review. Eleven measurement instruments for migraine were identified, of which the ID-Migraine is recommended with a moderate level of evidence and a pooled sensitivity of 0.87 (95% CI: 0.85–0.89) and specificity of 0.75 (95% CI: 0.72–0.78). Six measurement instruments examined both migraine and tension-type headache and only the Headache Screening Questionnaire – Dutch version has a moderate level of evidence with a sensitivity of 0.69 (95% CI 0.55–0.80) and specificity of 0.90 (95% CI 0.77–0.96) for migraine, and a sensitivity of 0.36 (95% CI 0.21–0.54) and specificity of 0.86 (95% CI 0.74–0.92) for tension-type headache. For cervicogenic headache, only the cervical flexion rotation test was identified and had a very low level of evidence with a pooled sensitivity of 0.83 (95% CI 0.72–0.94) and specificity of 0.82 (95% CI 0.73–0.91).

Discussion: The current review is the first to establish an overview of the diagnostic accuracy of measurement instruments for headaches associated with musculoskeletal factors. However, as most measurement instruments were validated in one study, pooling was not always possible. Risk of bias was a serious problem for most studies, decreasing the level of evidence. More research is needed to enhance the level of evidence for existing measurement instruments for multiple headaches.

Keywords

Diagnostics, headache, migraine, tension-type headache

Date received: 29 August 2018; revised: 8 November 2018; 29 November 2018; 13 February 2019; accepted: 25 February 2019

[1]ACHIEVE – Centre of Applied Research, Faculty of Health, Amsterdam University of Applied Sciences, Amsterdam, the Netherlands
[2]Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam, Department of Orofacial Pain and Dysfunction, the Netherlands
[3]Amsterdam University of Applied Sciences, Education of Physical Therapy, Faculty of Health, Amsterdam, the Netherlands
[4]Radboud University Medical Center, Research Institute for Health Sciences, IQ Healthcare, Nijmegen, the Netherlands

[5]University of Amsterdam, Amsterdam University Medical Centers (AUMC), Department of Rehabilitation, Amsterdam Movement Sciences, Amsterdam, the Netherlands
[6]University Medical Center Utrecht, Utrecht University, Department of Oral-Maxillofacial Surgery and Special Dental Care, Utrecht, the Netherlands

Corresponding author:
Hedwig van der Meer, Education of Physical Therapy, Amsterdam University of Applied Sciences, Tafelbergweg 51, 1105 BD Amsterdam, the Netherlands.
Email: h.a.van.der.meer@hva.nl

Introduction

Primary headaches like tension-type headache (TTH) and migraine are associated with various musculoskeletal factors. TTH is, for example, associated with pericranial tenderness, myofascial trigger points and lower muscle coordination of the upper neck flexors (1–4). Furthermore, migraine may be triggered by myofascial trigger points or bruxism (1,5–7). These primary headaches are not caused by musculoskeletal disfunction but are associated with different musculoskeletal symptoms (8). There are several secondary headaches that are actually caused by musculoskeletal problems, such as cervicogenic headache (CGH), headache after whiplash trauma and secondary headache attributed to temporomandibular dysfunction (TMD) (8). The physiotherapist (PT) is a specialist in the musculoskeletal field, and often treats patients with headaches associated with musculoskeletal symptoms. The type of headache must be diagnosed within the physiotherapeutic diagnostic process to choose the proper treatment options and collaborate with medical specialists when needed (9).

The International Headache Society (IHS) published the International Classification of Headache Disorders – 3rd edition (ICHD-3), which contains clear diagnostic criteria for all types of headache (8). Several headache measurement instruments are developed for PTs and other health care professionals to classify different headache types (10–14). The ability of a test to discriminate between the target condition and health or not having the target condition, is called the diagnostic accuracy of the test (15). The diagnostic accuracy is often quantified through measures of sensitivity and specificity (15). Insight into the diagnostic accuracy of these instruments for headaches associated with musculoskeletal symptoms is needed to determine the type of headache. Currently there is, to our knowledge, no overview of diagnostic accuracy of the different headache measurement instruments related to the level of evidence. Therefore, the aim of this study was to systematically review the available literature on the diagnostic accuracy of questionnaires and measurement instruments for headaches associated with musculoskeletal symptoms.

Methods

Protocol and registration

This review has been performed according to the PRISMA statement (17) and registered in PROSPERO (registration number: CRD42017062472). Due to the magnitude of articles found within the original search strategy, there were two review questions created. The focus of the current review is the diagnostic accuracy of measurement instruments for headaches associated with musculoskeletal symptoms. A second review (in preparation) will focus on the clinimetric properties of the instruments that measure other outcomes, based on the International Classification of Functioning, Disability and Health (16); for example, measurement instruments for pain, range of motion, limitations in activity, and quality of life.

Eligibility criteria

Only full text original articles were included concerning the diagnostic accuracy, expressed in sensitivity and specificity, of diagnostic headache tests usable for PTs. Further inclusion criteria were: a) adult patients (≥18 years) and b) patients that experienced headaches associated with musculoskeletal symptoms. These include migraine, TTH, CGH, headache after whiplash and headache attributed to TMD (8,19,20). There was no minimum sample size for inclusion. No restrictions were put on the year of publication. Intervention studies, prediction models and measurement instruments not usable for PTs (e.g. imaging, nerve blocks) (21) were excluded. Only articles in English were included .

Information sources

The electronic databases PubMed (1966–2018), Cochrane (1898–2018) and Cinahl (1988–2018) were searched for literature. The last search was performed on 25 October 2018. If full texts could not be obtained, the corresponding author was contacted through email to request the full text.

Search

The search strategies included search terms for the construct (e.g. pain, diagnosis), the target population (e.g. migraine, TTH), the instrument (e.g. questionnaire, test) and the methodological PubMed search filter for measurement instruments (21). The search filters for the Cochrane and Cinahl databases were derivatives from the PubMed search filter. The full search strategies for each database can be found in Supplemental material 1. References of retrieved articles were screened for additional relevant studies.

Study selection

Two reviewers (HvdM, CMV) independently assessed titles, abstracts and reference lists of the studies, using the online program Covidence (22). In case of disagreement between the two reviewers, a third reviewer (CMS) made the decision regarding inclusion of the article. After initial screening of the titles and abstracts,

HvdM and CMV read the full texts of included articles and screened these for eligibility. All reviewers are orofacial physiotherapists and researchers in this field.

Data collection process

Two reviewers (HvdM, CMS) independently extracted data from the included articles and registered this in a pre-made, empty Table 1 format. The data extracted were: First author, year of publication, target population, information about the index test (aim, language and name), reference test, study population, diagnostic accuracy (sensitivity/specificity).

Risk of bias in individual studies

The methodological quality of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2) (23,24). This tool assesses the risk of bias within four domains: Patient selection, index test, reference standard, and flow and timing (24). Concerns regarding applicability were also determined for the first three domains (24). Methodological quality of studies regarding the criterion validity was assessed using the COSMIN checklist (25). Criterion validity is defined as the degree to which the scores of an instrument are an adequate reflection of a gold standard (26). Within diagnostic accuracy, criterion validity is an essential measurement property. For criterion validity, box H of the COSMIN was used (25).

Data extraction and assessment of methodological quality were performed by two reviewers independently (HvdM, CMS). HvdM was trained to use the QUADAS-2 tool and CMS was trained by the COSMIN team on quality appraisal and data extraction. The protocol for methodological assessment using the QUADAS-2 tool for this review was made available for the review authors (Supplemental material 2). The protocol for the COSMIN checklist is published elsewhere (25).

Summary measures

Sensitivity and specificity were used as measures of diagnostic accuracy.

Synthesis of results

A best evidence synthesis was performed using the GRADE recommendations for diagnostic accuracy studies with the GRADE pro online software (27). These recommendations provide a step-by-step assessment to determine the certainty of evidence of a diagnostic test, which results in a comprehensive and transparent approach for developing the recommendations for these tests. To determine the impact of the test, both the sensitivity and specificity of the test must be known as well as the prevalence of the target condition (27). Based on the prevalence of the target population, the pre-test probability of the presence of the headache was determined for a population of 1000 people (27) . The test sensitivity and specificity was used to determine how many people would be accurately diagnosed (true positive) or excluded from having the headache (true negative).

A pooled sensitivity and specificity was used for each measurement instrument when there were multiple studies for one measurement tool. The pooled measurements were calculated using the 'rmeta' package for the R statistical software (28). A bivariate model resulting in a summary estimate for sensitivity and specificity together was used, as recommended by the Cochrane Collaboration (29,30). This model takes potential threshold effects and the correlation between sensitivity and specificity into account (29,30). The pooled sensitivity and specificity were used for the GRADE recommendations. When there was only one study for a measurement instrument, the published sensitivity and specificity of that measurement instrument were used. Finally, a summary receiver operating characteristics (S-ROC) curve was created using the 'mada' package for the R statistical software (29,31,32).

Factors determining the quality of evidence according to the GRADE approach are: a) Limitations in study design or execution (risk of bias); b) inconsistency of results; c) indirectness of evidence; d) imprecision; and e) publication bias (27). For limitations, the risk of bias assessment from the QUADAS-2 was used to determine if downgrading of the evidence was needed. When ≥50% of the assessed domains scored a "high" or "unclear" risk of bias, this was considered "serious" and the level of evidence was downgraded by one. When ≥75% of the assessed domains scored a "high" or "unclear" risk of bias, this was considered "very serious" and the level of evidence was downgraded by two. Inconsistency refers to unexplained heterogeneity of the results between multiple studies, after which the level of evidence may be downgraded. The indirectness of evidence was determined by the applicability assessment of the QUADAS-2 tool with the same rules as the risk of bias assessment. In the case where there was only one article studying a measurement tool, the evidence was downgraded for imprecision. All steps of the synthesis of results are depicted in Figure 1.

Risk of bias across studies

Methods to detect publication bias are not very reliable in diagnostic accuracy studies (30). As diagnostic accuracy studies have sensitivity and specificity values as

Table 1. Study characteristics of included articles stratified for the target populations of patients with migraine, migraine and tension-type headache and cervicogenic headache.

Measurement instrument	Author, year	Index test		Reference test	Population		Diagnostic accuracy	
		Language of index test	Aim of index test		N (%F)	Age: mean ± SD	Sensitivity	Specificity
Target Population: Migraine								
3-Question Screen	†Cady, 2004 (10)	English	Triage	ICHD (2)	3014 (85.2)	40.0 ± –	0.78*	0.27*
	Pryse-Phillips, 2002 (59)	English	Triage	Neurologist	476 (81.9)	40.4 ± –	0.86	0.73
Diagnostic screen	†Wahab, 2016 (41)	English	Triage	ICHD-3 (15)	1513 (50.1)	23.3 ± 2.5	0.66	0.98
	Michel, 1993 (II) (37)	English	Triage	Neurologist	160 (83.3)	39.9 ± 0.7	0.44	0.93
ID- Migraine	Brighina, 2006 (44)	Italian	Triage	ICHD-II (4)	222 (73.4)	37.8 ± 11.0	0.95	0.72
	de Mattos, 2017 (45)	Portuguese	Triage	ICHD-II (4)	232 (82.0)	48.9 ± 11.2	0.92	0.60
	Ertas, 2008 (46)	Turkish	Triage	ICHD-II (4)	2625 (58.5)	43.3–47.3 ± 16–18	0.80–0.88	0.74–0.76
	†Gil-Gouveia, 2009 (47)	Portuguese	Triage	ICHD-II (4)	142 (82.8)	39.2 ± 13.9	0.94	0.60
	†Karli, 2007 (49)	Turkish	Triage	ICHD-II (4)	3682 (62.9)	45.2 ± 17.0	0.92	0.63
	Kim, 2006 (50)	Korean	Triage	ICHD-II (4)	176 (81.2)	30.7 ± 9.3	0.58	0.98
	Lipton, 2003 (12)	English	Triage	ICHD (2)	451 (75.6)	39.3 ± 10.1	0.81	0.75
	†Lipton, 2016 (34)	English	Triage	ICHD-3 (15)	111 (82.9)	46.2 ± 13.4	0.81	0.89
	†Siva, 2008 (40)	Turkish	Triage	ICHD-II (4)	227 (65.6)	31.9 ± 5.9	0.71	0.79
MSMDQ	Rueda-Sanchez, 2004 (38)	Spanish	Unclear	Neurologist	170 ( )	–	0.38	0.99
Migraine Assessment Tool	Marcus, 2004 (35)	English	Triage	Neurologist	80 (88.8)	33.7 ± 9.9	0.89	0.79
Migraine Screen Questionnaire	Láinez, 2005 (11)	English	Triage	ICHD (2)	140 (73.0)	39.2 ± 13.0	0.93	0.81
	Láinez, 2010 (51)	English	Triage	ICHD-II (4)	9670 (61.9)	48.9 ± 17.2	0.82	0.97
Migraine-specific questionnaire	Kallela, 2001 (48)	Finnish	Triage	ICHD (2)	94 (71.3)	44.6 ± 18.0	0.99	0.96
Migraine-4	Walters, 2015 (42)	English	Triage	ICHD-3 (15)	1829 (71.5)	19.1 ± 2.1	0.94	0.92
Modified Algorithm for IHS Migraine	Michel, 1993 (I) (36)	English	Replacement	Neurologist	267 (70.3)	–	0.95–0.98	0.53 – 0.78
Screening items	Wang, 2008 (43)	–	Triage	ICHD-II (4)	755 (71.0)	37 ± 15	0.89*	0.67*
Structured Migraine Interview Questionnaire	Shaik, 2015 (39)	Malay	Unclear	ICHD-II (4)	157 (100)	26.8 ± 8.3	0.97	0.63
Target population: Migraine and tension-type headache								
Computerized Headache Assessment Test	Maizels, 2007 (54)	English	Replacement	Headache nurse	117 (–)	–	M: 0.83–1.00 TTH: 1.00	– –
German language questionnaire	†Fritsche, 2007 (13)	German	Replacement	ICHD-II (4)	278 (51.1)	43.9 ± –	M: 0.73 TTH: 0.85	M: 0.96 TTH: 0.98
	†Yoon, 2008 (56)	German	Replacement	ICHD-II (4)	193 (68.4)	45.4 ± 12.4	M: 0.85 T: 0.60	M: 0.85 T: 0.88

(continued)

Table I. Continued.

Measurement instrument	Author, year	Index test			Population		Diagnostic accuracy	
		Language of index test	Aim of index test	Reference test	N (%F)	Age: mean ± SD	Sensitivity	Specificity
Headache Screening Questionnaire – Dutch version	van der Meer, 2017 (14)	Dutch	Triage	ICHD-3 (15)	105 (78.1)	40.3 ± 14.5	M: 0.69 PM: 0.89 TTH: 0.36 PTTH: 0.92	M: 0.90 PM: 0.54 TTH: 0.86 PTTH: 0.48
Headache questions	Hagen, 2010 (53)	Norwegian	Unclear	ICHD-II (4)	297 (49.0)	52.3 ±–	M:0.49–0.67 TTH: 0.96 CT: 0.64	M: 0.91–0.95 TTH: 0.69 CTTH: 1.00
Self-administered headache questionnaire	Rasmussen, 1991 (55)	Danish	Replacement	Neurologist	713 (–)	–	M: 0.51 TTH: 0.43	M: 0.92 TTH: 0.96
Structured Headache Questionnaire	el-Sherbiny, 2017 (52)	Arabic	Unclear	ICHD-3 (15)	232 (72.8)	41.2 ± 10.9	M: 0.86 CM: 0.71 TTH: 0.93 CTTH: 0.70	M: 0.94 CM: 0.98 TTH: 0.93 CTTH: 0.96
Target population: Cervicogenic headache								
Cervical Flexion-Rotation Test (CFRT)	†Hall, 2010 (57)	n/a	Unclear	Sjaastad criteria (32)	60 (63.3)	30–35 ± 6.5–10.9	0.70	0.70
	†Ogince, 2007 (58)	n/a	Unclear	Sjaastad criteria (32)	58 (65.5)	37–46 ±–	0.91	0.91

*Not given in article, therefore calculated based on the published 2 × 2 table.
†Articles included in meta-analysis as shown as in Table 3.
MSMDQ: Michel's Standardized Migraine Diagnosis Questionnaire; –: missing data; F: female; SD: standard deviation; M: migraine; CM: chronic migraine; PM: probable migraine; TTH: tension-type headache; CTTH: chronic tension-type headache; PTTH: probable tension-type headache; n/a: not applicable.
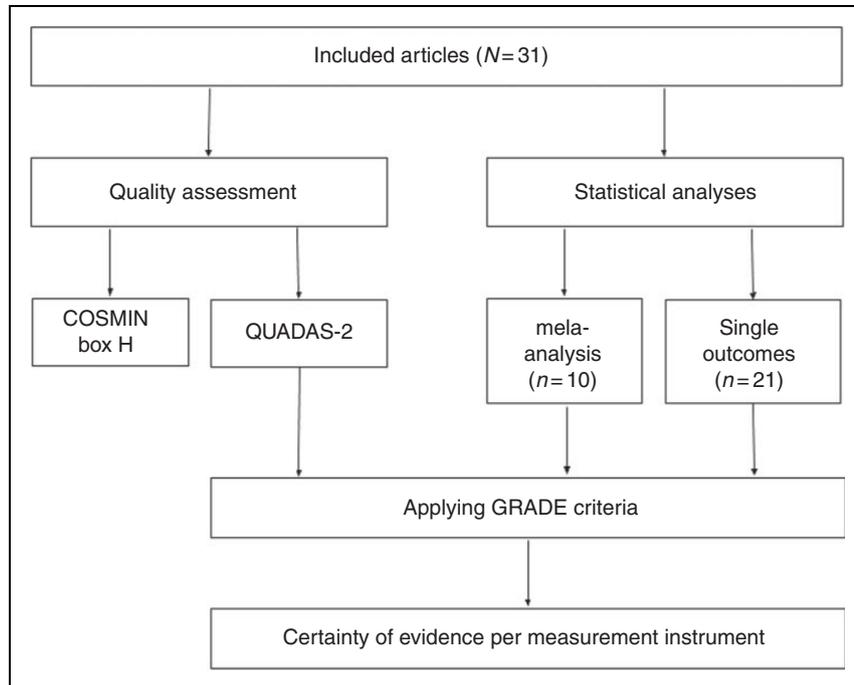
Figure 1. Flow of steps after article inclusion.

outcome measures rather than a stated null hypothesis with a *p*-value, it is unlikely for publication bias to be associated with statistical nonsignificance (33). Therefore, no publication bias assessment was applied in this review.

## Results

Study selection

The search in all three databases resulted in 4129 articles, which were imported in Covidence (22). After removing duplicates and assessment of eligibility on title/abstract, 150 articles remained to be assessed full text. Of these, 52 articles were excluded based on the inclusion and exclusion criteria (Supplemental material 3) and 67 articles assessed other clinimetric outcome measures than diagnostic accuracy. These 67 articles will be included in the second review regarding clinimetric outcome measures based on the ICF. This resulted in 31 articles to be included in the current review. The complete flowchart of the study selection can be found in Figure 2. No authors were contacted to obtain the full texts of any study.

Study characteristics

The included headaches associated with musculoskeletal symptoms in this review are migraine, TTH and CGH. No measurement instruments were found that

studied the diagnostic accuracy for instruments related to secondary headache attributed to TMD or headache attributed to whiplash injury. Table 1 shows the study characteristics of the 31 included studies, stratified by target population of the measurement instrument. From the 31 studies, 22 articles had migraine as the target population (10–12,34–51). Seven articles had both migraine and TTH as target population (13,14,52–56), and two articles examined patients with CGH (57,58). In total, 28,246 people were included in the 31 studies. Of the included population, 64% were female, though three articles did not describe the gender distribution (38,54,55). Mean age varied from 19 (42) to 52 years (53).

For migraine, 11 different measurement instruments were studied (10–12,34–37,40–43,44–51,59). ID-Migraine was the most studied measurement instrument, with nine studies in five languages (12,34,40,44–47,49,50). Eight of these instruments were screening instruments, one was a replacement test for the diagnostic process, and for two instruments the aim of the test was unclear. Out of the seven studies for both migraine and TTH, only two articles looked at the same questionnaire (13,56). From the seven instruments, one was a screening test, three were replacement tests, and the aim of two was unclear. Both studies on CGH researched the cervical flexion-rotation test (CFRT) (57,58). The aim of the CFRT compared to the ICHD-3 criteria for cervicogenic headache is unclear.
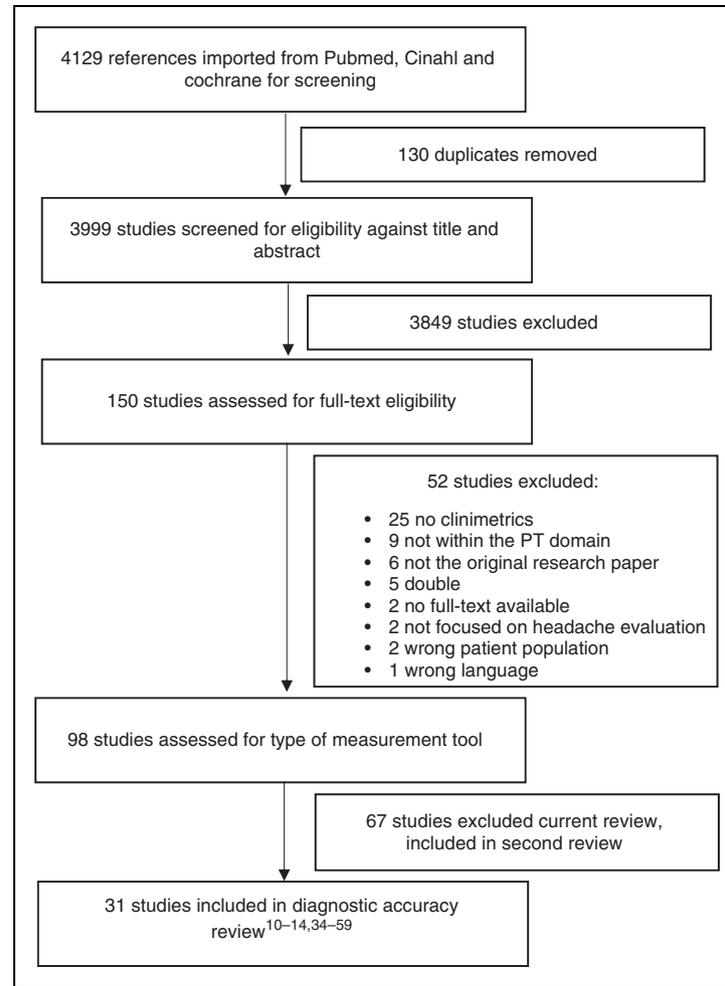
Figure 2. Study flow diagram.

Risk of bias within studies

The risk of bias was assessed for patient selection, index test, reference standard and flow and timing. The summarized assessment of the QUADAS-2 can be found in Table 2. The complete assessment, including reasons for the given scores, can be found in Supplemental material 4. Only one study received a low risk of bias on all domains (43). Twenty-two articles received a "high" risk of bias on ≥1 domain (10–14,35,37, 39–41,43,45–50,55–59). The remaining articles received an "unclear" risk of bias on ≥1 domain (12,35,37, 41,50–53). Risk of bias for the index test and the reference standard was generally scored unclear, because there was uncertainty if the index test was conducted and interpreted without knowledge of the results of the reference standard.

The clinimetric evaluation of the criterion validity was established with the COSMIN Box H. One study scored excellent (14), one good (35), 21 fair (11,12,34,36–48,50–53,57) and the remaining eight

scored poor (10,13,50,55–57,59). Of the studies scoring poor, all but two (54,55) also scored a high risk of bias on ≥2 domains (10,12,13,50,55,57,59).

Migraine measurement instruments

Results of individual studies. The sensitivity of the measurement instruments for migraine ranged from 0.38 (38) to 0.99 (48) (see Table 1). Only three studies had a sensitivity below 0.70 (38,41,50) and eight studies found a sensitivity of 0.90 or higher (11,39,42,44, 45,47–49). Half of these studies with a high sensitivity were researching the ID-Migraine (44,45,47,49). Specificity ranged from 0.27 (10) to 0.99 (37). Six studies found a specificity of 0.70 or lower (10,39,43,45, 47,49), and a specificity above 0.90 was found in six other studies (38,41,42,48,50,51). Eleven studies had both sensitivity and specificity above 0.70 (11,12,34, 35,40,42,44,46,48,51,59), of which two studies had both above 0.90 (42,48).

Table 2. Methodological quality assessment with QUADAS-2 and clinimetric evaluation of the criterion validity with the COSMIN checklist Box H.

Measurement instrument	Study	Risk of Bias				Applicability concerns			COSMIN Box H
		1a. Patient selection	2a. Index test	3a. Reference standard	4. Flow and timing	1b. Patient selection	2b. Index test	3b. Reference standard	
Target population: Migraine									
3-Question Screen	Cady, 2004 (10)	High	Unclear	Unclear	High	Low	Low	Low	Poor
	Pryse-Phillips, 2002 (59)	High	Unclear	High	High	Low	Low	Low	Poor
	Wahab, 2016 (41)	Unclear	Unclear	Unclear	Low	Low	Low	Low	Fair
Diagnostic Screen	Michel, 1993 (37)	Unclear	Unclear	Unclear	High	Low	Low	Low	Fair
ID-Migraine	Brighina, 2006 (44)	Low	Low	Low	Low	Low	Low	Low	Fair
	de Mattos, 2017 (45)	High	Low	Low	Unclear	Low	Low	Low	Fair
	Ertas, 2008 (46)	High	Low	Unclear	Low	Low	Low	Low	Fair
	Gil-Gouveia, 2009 (47)	High	Low	Low	Low	Low	Low	Low	Fair
	Karli, 2007 (49)	High	Unclear	Unclear	High	Low	Low	Low	Poor
	Kim, 2006 (50)	Unclear	Low	Unclear	Low	Low	Low	Low	Fair
	Lipton, 2003 (12)	Unclear	Unclear	Low	Unclear	Low	Low	Low	Fair
	Lipton, 2016 (34)	High	Unclear	Unclear	Unclear	Low	Low	Low	Fair
	Siva, 2008 (40)	High	Low	Low	Unclear	Low	Low	Low	Fair
MSMDQ	Rueda-Sánchez, 2004 (38)	Low	Unclear	Unclear	High	Low	Low	Low	Fair
MAT	Marcus, 2004 (35)	Low	Low	Unclear	Low	Low	Low	Low	Good
Migraine Screen Questionnaire	Láinez, 2010 (51)	Low	Low	Unclear	Low	Low	Low	Low	Fair
	Láinez, 2005 (11)	High	High	Low	Unclear	Low	Low	Low	Fair
MSQ	Kallela, 2001 (48)	Low	Low	Unclear	High	Low	Low	Low	Fair
Migraine-4	Walters, 2015 (42)	Low	Unclear	Unclear	High	Low	Low	Low	Fair
MA-HIS-M	Michel, 1993 (36)	Unclear	Low	Low	Unclear	Low	Low	Low	Fair
Screening items	Wang, 2008 (43)	Unclear	Unclear	Unclear	High	Low	Low	Low	Fair
SMIQ	Shaik, 2015 (39)	High	Unclear	Unclear	Low	Low	Low	Low	Fair
Target population: Migraine and tension-type headache									
CHAT	Maizels, 2007 (54)	High	Low	Unclear	Unclear	Low	Low	Low	Poor
German Language Questionnaire	Fritsche, 2007 (13)	High	Low	Low	High	Low	Low	Low	Poor
	Yoon, 2008 (56)	High	Low	Unclear	High	Low	Low	Low	Poor
HSQ-DV	van der Meer, 2017 (14)	Low	Low	Low	High	Low	Low	Low	Excellent
Headache questions	Hagen, 2010 (53)	Low	Unclear	Unclear	Unclear	Low	Low	Low	Fair
SAHQ	Rasmussen, 1991 (55)	Low	Low	High	Unclear	Low	Low	Low	Poor
SHQ	El-Sherbiny, 2017 (52)	Unclear	Low	Unclear	Unclear	Low	Low	Low	Fair
Target population: cervicogenic headache									
Cervical Flexion-Rotation Test	Hall, 2010 (57)	High	Low	Unclear	Unclear	Low	Low	Low	Fair
	Ogince, 2005 (58)	High	Unclear	Unclear	High	Low	Low	Low	Poor

MSMDQ: Michel's Standardized Migraine Diagnosis Questionnaire; MAT: Migraine Assessment Questionnaire; MSQ: Migraine-specific questionnaire; Ib: Patient selection; MA-IHS-M; Modified Algorithm for IHS Migraine; SMIQ: Structured Migraine Interview Questionnaire; CHAT: Computerized Headache Assessment Test; HSQ-DV: Headache Screening Questionnaire – Dutch Version; SAHQ: Self-Administered Headache Questionnaire; SHQ: Structured Headache Questionnaire. An extended version of this table including explanation of judgement can be found in Appendix 4.

Table 3. Pooled sensitivity and specificity of the 3-Question screen, ID-Migraine, German language questionnaire and Cervical Flexion-Rotation Test.

Measurement instrument	Target population	Number of studies; author, year	Pooled sensitivity (95% CI)	Pooled specificity (95% CI)
3-Question screen	Migraine	2; Cady, 2004 (10) Wahab, 2016 (41)	0.73 (0.71–0.75)	0.93 (0.9–0.94)
ID-Migraine	Migraine	4; Lipton, 2016 (34) Siva, 2008 (40) Gil-Gouveia, 2009 (47) Karli, 2007 (49)	0.87 (0.85–0.89)	0.75 (0.72–0.78)
German language questionnaire	Migraine TTH	2; Fritsche, 2007 (13) Yoon, 2008 (56)	0.69 (0.63–0.75) 0.81 (0.75–0.87)	0.90 (0.86–0.94) 0.96 (0.94–0.98)
Cervical Flexion-Rotation Test	Cervicogenic headache	2; Hall, 2010 (57) Ogince, 2007 (58)	0.83 (0.72–0.94)	0.82 (0.73–0.91)

N: number; CI: confidence interval; TTH: tension-type headache.

Synthesis of results. For two measurement instruments, the sensitivity and specificity could be pooled. For the 3-question Screen the pooled sensitivity was 0.73 and specificity was 0.93 (Table 3) based on two (10,41) out of three studies, due to missing data in one article (59). The pooled sensitivity for the ID-Migraine was 0.87 and specificity was 0.75 (Table 3, Figures 3(a) and 3(b)). The results were based on four studies (34,40,47,49) as the other five studies (12,44–46,50) did not have sufficient data available to perform the analyses.

There was a very low level of evidence for six measurement instruments for migraine related to the GRADE recommendations: Diagnostic Screen (37), Michel's Standardized Migraine Diagnosis Questionnaire (38), Migraine Specific Questionnaire (48), Migraine-4 (42), Modified Algorithm for IHS Migraine (36), Screening Items (43), and the Structured Migraine Interview Questionnaire (see Table 4) (39). For two measurement instruments, there was a low level of evidence: The 3-question Screen (10,41) and the Migraine Screen Questionnaire (11,51). There was a moderate level of evidence for the ID-Migraine (34,40,47,49) and also for the Migraine Assessment Tool (35).

Combined migraine and TTH measurement instruments

Results of individual studies. The aim of the index tests differed between the included seven articles, where four were 'replacement' tests (13,54–56), one a 'triage' test (14) and two aims were unclear (52,53). Three articles established the diagnostic accuracy for several migraine and TTH ICHD diagnoses aside from the "standard" diagnoses, including chronic migraine, chronic TTH, probable migraine, and probable TTH (14,52,53). For migraine, the sensitivity ranged from 0.49 (53) to 1.00 (54) and the specificity ranged from 0.85 (56) to 0.96 (13). For chronic migraine, the sensitivity and specificity were 0.71 and 0.98 respectively (52). Probable migraine had a sensitivity of 0.89 and a specificity of 0.54 (14). The sensitivity for TTH ranged from 0.36 (14) to 1.00 (54) and the specificity range was 0.69 (53) to 0.98 (13). One study did not establish the specificity results from their test (54). Chronic TTH was tested in two studies, for which the sensitivity was 0.64 (53) to 0.70 (52) and the specificity 0.96 (52) to 1.00 (53). The test for probable TTH had a sensitivity of 0.92 and a specificity of 0.48 (14).

For migraine, chronic migraine, and probable migraine (13,14,52,54,56) five studies had a sensitivity above 0.70, which was also found for TTH, chronic TTH, and probable TTH in five studies (see Table 1) (13,14,52–54). All six studies that reported specificity, had a specificity of 0.70 or higher for migraine, chronic migraine, and probable migraine and for TTH chronic TTH, and probable TTH (13,14,52,53,55,56).

Synthesis of results. One instrument, the German Language Questionnaire, was supported by two studies (13,56). The pooled sensitivity and specificity for migraine were 0.69 and 0.90 respectively (Table 3, Figure 3(c)). For TTH, the pooled sensitivity and specificity were 0.81 and 0.96 respectively (Table 3, Figure 3(d)). The five other measurement instruments (14,52–55) were supported by one study and therefore downgraded for imprecision (see also Table 5).

There was a very low level of evidence for the Computerized Headache Assessment Test (CHAT) (54), the use of Headache Questions (53) and the Structured Headache Questionnaire (52). The German Language Questionnaire (13,54) and the Self-Administered Headache Questionnaire (55) are both supported with a low level of evidence. Only the Headache Screening Questionnaire (HSQ)– Dutch
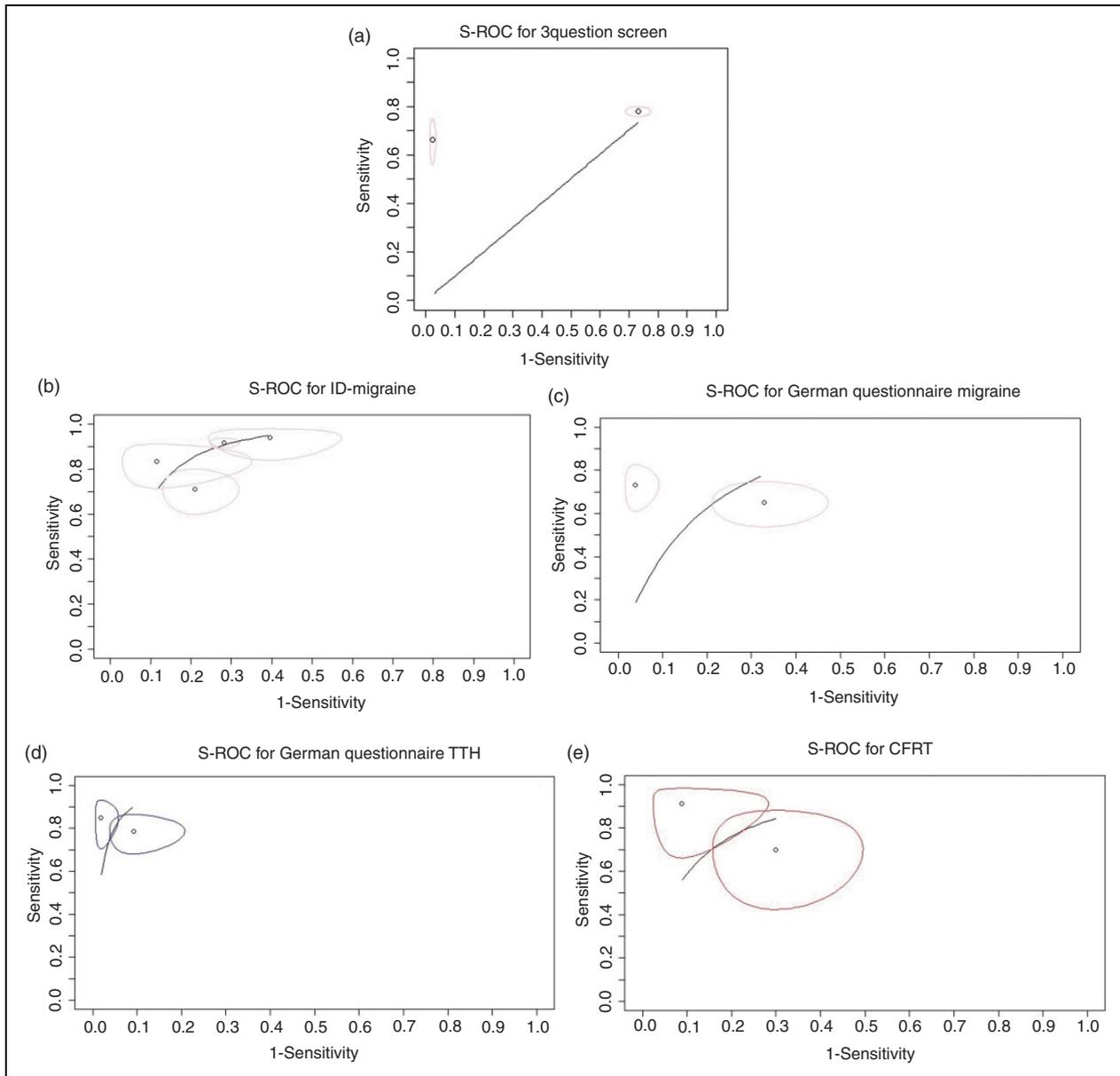
Figure 3. (a) Summary Receiver Operating Characteristics (S-ROC) curves for pooled sensitivity and specificity of the 3-question screen; (b) S-ROC curves for pooled sensitivity and specificity of the ID-migraine; (c) S-ROC curves for pooled sensitivity and specificity of the German questionnaire for migraine; (d) S-ROC curves for pooled sensitivity and specificity of the German questionnaire for tension-type headache; (e) S-ROC curves for pooled sensitivity and specificity of the cervical flexion rotation test.

Version was found to have a moderate level of evidence (14).

Cervicogenic headache measurement instruments

Results of individual studies. The two included studies for CGH established the diagnostic accuracy of the Cervical Flexion-Rotation Test (CFRT) (57,58). Both sensitivity and specificity ranged from 0.70 (57) to 0.91 (58).

Synthesis of results. The pooled sensitivity was 0.83 and the pooled specificity was 0.82 (Table 3, Figure 3(e)). Based on the GRADE recommendations (Table 6), there is a low level of evidence for the use of the CFRT for patients with cervicogenic headache (57,58).

Discussion

Within this review, for migraine alone 11 tools were identified (10–12, 34–37,40–51,59), for the combination

Table 4. GRADE recommendations for measurement instruments for target population Migraine, stratified per measurement instrument.

Measurement instrument	Sensitivity (95% CI) / Specificity (95% CI)	Outcome	Number of studies (number of patients)	Study design	Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Pre-test probability of 14.7%*	Test accuracy CoE
3-Question Screen (10,41,59)	0.73 (0.71–0.75)‡	TP	Two studies 2539 patients	Cross-sectional (cohort type accuracy study)	Serious‡	Not serious	Serious	Not serious	None	107 (104 to 110)	⊕⊕◯◯ Low
		FN								40 (37 to 43)	
	0.93 (0.92–0.94)‡‡	TN	Two studies 1988 patients		Serious‡	Not serious	Serious	Not serious	None	793 (785 to 802)	⊕⊕◯◯ Low
		FP								60 (51 to 68)	
Diagnostic Screen (37)	0.44 (0.35–0.53)	TP	One study 125 patients	Cross-sectional (cohort type accuracy study)	Very Serious*	Not serious	Not serious	Serious×	None	65 (51 to 78)	⊕◯◯◯ Very low
		FN								82 (69 to 96)	
	0.93 (0.85–1.00)	TN	One study 41 patients		Very serious*	Not serious	Not serious	Serious×	None	793 (725 to 530)	⊕◯◯◯ Very low
		FP								60 (0 to 128)	
ID-Migraine (34,40,47,49)	0.87 (0.85–0.89)‡	TP	Four studies 1257 patients	Cross-sectional (cohort type accuracy study)	Serious‡	Not serious	Not serious	Not serious	None	128 (125 to 131)	⊕⊕⊕◯ Moderate
		FN								19 (16 to 22)	
	0.75 (0.72–0.78)‡‡	TN	Four studies 1109 patients		Serious‡	Not serious	Not serious	Not serious	None	640 (614 to 665)	⊕⊕⊕◯ Moderate
		FP								213 (188 to 239)	
Michel's Standardized Migraine Diagnosis Questionnaire[38]	0.38 (0.26–0.52)	TP	One study ? patients	Cross-sectional (cohort type accuracy study)	Very serious*	Serious	Not serious	Serious×	None	56 (38 to 76)	⊕◯◯◯ Very low
		FN								91 (71 to 109)	
	0.99 (0.95–1.00)	TN	One study ? patients		Very serious*	Serious	Not serious	Serious×	None	844 (810 to 853)	⊕◯◯◯ Very low
		FP								9 (0 to 43)	
Migraine Assessment Tool (35)	0.89 (0.80–0.98)‡	TP	One study 46 patients	Cross-sectional (cohort type accuracy study)	Not serious	Not serious	Not serious	Serious×	None	131 (118 to 144)	⊕⊕⊕◯ Moderate
		FN								16 (3 to 29)	
	0.79 (0.65-0.93)‡	TN	One study 34 patients		Not serious	Not serious	Not serious	Serious×	None	674 (554 to 793)	⊕⊕⊕◯ Moderate
		FP								179 (60 to 299)	
Migraine Screen Questionnaire (11,51)	0.82–0.93	TP	Two studies ? patients	Cross-sectional (cohort type accuracy study)	Serious‡	Serious b	Not serious	Not serious	None	121 to 137	⊕⊕◯◯ Low
		FN								10 to 26	
	0.81–0.97	TN	Two studies ? patients		Serious‡	Serious b	Not serious	Not serious	None	691 to 827	⊕⊕◯◯ Low
		FP								26 to 162	
Migraine Specific Questionnaire (48)	0.99 (0.97–1.00)‡‡	TP	One study 69 patients	Cross-sectional (cohort type accuracy study)	Serious‡	Serious	Not serious	Serious×	None	146 (143 to 147)	⊕◯◯◯ Very low
		FN								1 (0 to 4)	
	0.96 (0.88–1.00)‡	TN	One study 25 patients		Serious‡	Serious	Not serious	Serious×	None	819 (751 to 853)	⊕◯◯◯ Very low
		FP								34 (0 to 102)	
Migraine-4 (42)	0.94 (0.87–0.98)	TP	One study ? patients	Cross-sectional (cohort type accuracy study)	Very serious*	Not serious	Not serious	Serious×	None	138 (128 to 144)	⊕◯◯◯ Very low
		FN								9 (3 to 19)	
	0.92 (0.90–0.94)	TN	One study ? patients		Very serious*	Not serious	Not serious	Serious×	None	785 (768 to 802)	⊕◯◯◯ Very low
		FP								68 (51 to 85)	

(continued)

Table 4. Continued.

Measurement instrument	Sensitivity (95% CI)	Specificity (95% CI)	Outcome	Number of studies (number of patients)	Study design	Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Pre-test probability of 14.7%*	Test accuracy CoE
Modified Algorithm for IHS Migraine (36)	0.95–0.98		TP	One study 126 patients	Cross-sectional (cohort type accuracy study)	Serious±	Serious	Serious	Serious×	None	144 to 144	⊕○○○ Very low
			FN								3 to 7	
		0.53–0.78	TN	One study 141 patients		Serious±	Serious	Serious	Serious×	None	452 to 665	
			FP								188 to 401	
Screening Items (43)	0.89 (0.86–0.92)‡		TP	One study 363 patients	Cross-sectional (cohort type accuracy study)	Very serious¥	Not serious	Not serious	Serious×	None	131 (126 to 135)	⊕○○○ Very low
			FN								16 (12 to 21)	
		0.67 (0.63–0.72)‡	TN	One study 392 patients		Very serious¥	Not serious	Not serious	Serious×	None	572 (537 to 614)	
			FP								281 (239 to 316)	
Structured Migraine Interview Questionnaire (39)	0.97 (0.94–1.00)‡		TP	One study 100 patients	Cross-sectional (cohort type accuracy study)	Very serious¥	Not serious	Not serious	Serious×	None	143 (138 to 147)	⊕○○○ Very low
			FN								4 (0 to 9)	
		0.63 (0.50–0.76)‡	TN	One study 57 patients		Very serious¥	Not serious	Not serious	Serious×	None	542 (427 to 648)	
			FP								316 (205 to 426)	

*Prevalence in the general population of 14.7% is used (65). CoE: certainty of evidence.

±"Unclear" or "high" risk of bias on ≥50 < 75% of the domains on QUADAS-2.

¥"Unclear" or "high" risk of bias on ≥75% of the domains on QUADAS-2.

×Results based on the outcome of one single study.

‡95% confidence interval (CI) calculated by reviewers.

Table 5. GRADE recommendations for measurement instruments for target populations Migraine and Tension-Type Headache, stratified per measurement instrument.

Measurement instrument	Target population	Sensitivity (95% CI) Specificity (95% CI)	Outcome	№ of studies № of patients	Study design	Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Effect per 1.000 patients tested* Pre-test probability of 14.7%*/62.6%**	Test accuracy CoE
Computerized Headache Assessment Test (CHAT) (54)	Migraine	0.98‡ (0.93–1.00)	TP	One study 41 patients	Cross-sectional (cohort type accuracy study)	Very serious*	Serious	Not serious	Serious×	None	144 (137 to 147)	⊕○○○ Very low
			FN								3 (0 to 10)	
		1.00‡ (1.00–1.00)	TN	One study 76 patients		Very serious*	Not serious	Not serious	Serious×	None	853 (853 to 853)	⊕○○○ Very low
			FP								0 (0 to 0)	
	TTH	1.00‡ (1.00–1.00)	TP	One study 14 patients		Very serious*	Serious	Not serious	Serious×	None	626 (626 to 626)	⊕○○○ Very low
			FN								0 (0 to 0)	
		1.00‡ (1.00–1.00)	TN	One study 14 patients		Very serious*	Very serious	Not serious	Serious×	None	374 (374 to 374)	⊕○○○ Very low
			FP								0 (0 to 0)	
German Language Questionnaire (13,56)	Migraine	0.69‡ (0.63–0.75)	TP	Two studies 217 patients	Cross-sectional (cohort type accuracy study)	Serious‡	Serious	Not serious	Not serious	None	101 (81 to 118)	⊕⊕○○ Low
			FN								46 (29 to 66)	
		0.90‡ (0.86–0.94)	TN	Two studies 254 patients		Serious‡	Serious	Not serious	Not serious	None	768 (657 to 819)	⊕⊕○○ Low
			FP								85 (34 to 196)	
	TTH	0.81‡ (0.75–0.87)	TP	Two studies 177 patients		Serious‡	Serious	Not serious	Not serious	None	507 (470 to 545)	⊕⊕○○ Low
			FN								119 (81 to 156)	
		0.96‡ (0.94–0.98)	**TN**	Two studies 294 patients		Serious‡	Serious	Not serious	Not serious	None	359 (352 to 367)	⊕⊕○○ Low
			FP								15 (7 to 22)	
Headache Screening Questionnaire – Dutch Version (14)	Migraine	0.69 (0.55–0.80)	TP	One study 55 patients	Cross-sectional (cohort type accuracy study)	Not serious	Not serious	Not serious	Serious×	None	101 (81 to 118)	⊕⊕⊕○ Moderate
			FN								46 (29 to 66)	
		0.90 (0.77–0.96)	TN	One study 50 patients		Not serious	Not serious	Not serious	Serious×	None	768 (657 to 819)	⊕⊕⊕○ Moderate
			FP								85 (34 to 196)	
	TTH	0.36 (0.21–0.54)	TP	One study 36 patients		Not serious	Not serious	Not serious	Serious×	None	225 (131 to 338)	⊕⊕⊕○ Moderate
			FN								401 (288 to 495)	
		0.86 (0.74–0.92)	TN	One study 69 patients		Not serious	Not serious	Not serious	Serious×	None	322 (277 to 344)	⊕⊕⊕○ Moderate
			FP								52 (30 to 97)	
Headache Questions (53)	Migraine	0.49 (–)†	TP	One study ? patients	Cross-sectional (cohort type accuracy study)	Very serious*	Not serious	Serious	Serious×	None	72 (- to -)	⊕○○○ Very low
			FN								75 (- to -)	
		0.91 (–)†	TN	One study ? patients		Very serious*	Not serious	Serious	Serious×	None	776 (- to -)	⊕○○○ Very low
			FP								77 (- to -)	
	TTH	0.96 (0.94–0.98)	TP	One study ? patients		Very serious*	Not serious	Not serious	Serious×	None	601 (588 to 613)	⊕○○○ Very low
			FN								25 (13 to 38)	
		0.69 (0.63–0.75)	TN	One study ? patients		Very serious*	Not serious	Not serious	Serious×	None	258 (236 to 281)	⊕○○○ Very low
			FP								116 (93 to 138)	

(continued)

Table 5. Continued.

Measurement instrument	Target population	Sensitivity (95% CI) / Specificity (95% CI)	Outcome	№ of studies / № of patients	Study design	Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Effect per 1.000 patients tested* — Pre-test probability of 14.7%*/62.6%**	Test accuracy CoE
Self-administered Headache Questionnaire (55)	Migraine	0.51‡ (0.41–0.61)	TP	One study 93 patients	Cross-sectional (cohort type accuracy study)	Serious‡	Not serious	Not serious	Serious×	None	75 (60 to 90)	⊕⊕○○ Low
			FN								72 (57 to 87)	
		0.92‡ (0.90–0.94)	TN	One study 619 patients		Serious‡	Not serious	Not serious	Serious×	None	785 (768 to 802)	⊕⊕○○ Low
			FP								68 (51 to 85)	
	TTH	0.43‡ (0.39–0.47)	TP	One study 468 patients		Serious‡	Not serious	Not serious	Serious×	None	269 (244 to 294)	⊕⊕○○ Low
			FN								357 (332 to 382)	
		0.96‡ (0.94–0.98)	TN	One study 244 patients		Serious‡	Not serious	Not serious	Serious×	None	359 (352 to 367)	⊕⊕○○ Low
			FP								15 (7 to 22)	
Structured Headache Questionnaire (52)	Migraine	0.86 (0.78–0.97)	TP	One study ? patients	cross-sectional (cohort type accuracy study)	Very serious*	Not serious	Not serious	Serious×	None	126 (115 to 143)	⊕○○○ Very low
			FN								21 (4 to 32)	
		0.94 (0.86–0.98)	TN	One study ? patients		Very serious*	Not serious	Not serious	Serious×	None	802 (734 to 836)	⊕○○○ Very low
			FP								51 (17 to 119)	
	TTH	0.93 (0.79–0.98)	TP	One study ? patients		Very serious*	Not serious	Not serious	Serious×	None	582 (495 to 613)	⊕○○○ Very low
			FN								44 (13 to 131)	
		0.93 (0.86–1.00)	TN	One study ? patients		Very serious*	Not serious	Not serious	Serious×	None	348 (322 to 374)	⊕○○○ Very low
			FP								26 (0 to 52)	

*Prevalence in the general population of 14.7% is used for migraine.

**Prevalence in the general population of 62.6% is used for TTH (65).

CoE: certainty of evidence.

‡"Unclear" or "high" risk of bias on ≥50 < 75% of the domains on QUADAS-2.

* "Unclear" or "high" risk of bias on ≥75% of the domains on QUADAS-2.

×Results based on the outcome of one single study.

‡95% confidence interval (CI) calculated by reviewers.

†Not possible to calculate 95% CI.

Table 6. GRADE recommendations for measurement instruments for target population Cervicogenic Headache.

Measurement instrument	Sensitivity (95% CI)	Specificity (95% CI)	Outcome	Number of studies (number of patients)	Study design	Factors that may decrease certainty of evidence					Effect per 1.000 patients tested*	Test accuracy CoE
						Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Pre-test probability of 4.1%*	
Cervical Flexion Rotation Test (57,58)	0.83‡ (0.72–0.94)	0.82‡ (0.73–0.91)	TP	Two studies 43 patients	Cross-sectional (cohort type accuracy study)	Very serious*	Not serious	Serious	Not serious	None	34 (30 to 39)	⊕○○○ Very low
			FN								7 (2 to 11)	
			TN	Two studies 74 patients		Very serious*	Not serious	Serious	Not serious	None	786 (700 to 873)	⊕○○○ Very low
			FP								173 (86 to 259)	

*Prevalence in the general population of 4.1% is used (76).

CoE: certainty of evidence.

*"Unclear" or "high" risk of bias on ≥75% of the domains on QUADAS-2.

‡95% confidence interval (CI) calculated by reviewers.

of migraine and TTH six (13,14,52–56), and for CGH one tool (57,58). The sensitivity and specificity of the measurement instruments for migraine ranged from 0.38 (38) to 0.99 (48) and 0.27 (10) to 0.99 (37) respectively. The sensitivity and specificity for migraine based on the combined measurement instruments ranged from 0.49 (53) to 1.00 (54) and 0.85 (56) to 0.96 (13) respectively. For TTH, the sensitivity and specificity ranged from 0.36 (14) to 1.00 (54) and 0.59 (53) to 0.98 (13) respectively. For the CFRT, the only measurement instrument for cervicogenic headache, both the sensitivity and specificity ranged from 0.70 (57) to 0.91 (58). All measurement tools for migraine and TTH were questionnaires. The measurement tool for CGH was a physical examination test. Migraine and TTH are solely based on information from the history of the patient (15), allowing the diagnosis to be derived from a questionnaire. However, the choice of gold standard within headache research is inconsistent. Some studies used the International Classification of Headache Disorders (ICHD) first, second or third edition (15,60,61), others used the diagnosis of a neurologist or a headache nurse and for CGH the Sjaastad criteria were used (62). As the ICHD is based on the most recent scientific findings and clinical expertise from experts worldwide, the newest version of the ICHD is recommended as the gold standard (15,63).

The aim of each measurement instrument is described in Table 1. This was unclear for five measurement instruments. Nine measurement instruments are meant to be used as a screening tool in a broader population before seeing a medical specialist for a definitive diagnosis. These screening instruments are recommended for health care providers like PTs, as they are not trained for medical diagnoses but do see these patients often and can refer them to the medical specialist (64). Three measurement instruments studied were meant as a replacement test for the gold standard. This may be efficient for research purposes, as this allows the researchers to diagnose the patients without an extensive visit to a specialist. However, no conclusion was drawn from the included articles as to whether the measurement instruments were better than the gold standard (the medical specialist), therefore the presence of a medical specialist is still recommended in clinical practice.

For each measurement tool, the cut-off criteria to recognize headache should be described to allow for comparison of outcomes between studies. In reality, cut-off criteria differed between studies, which resulted in highly variable sensitivity and specificity. The lack of established cut-off points was taken into account within the 'Index Test' domain when assessing both methodological qualities and risk of bias.

Migraine measurement instruments

From the 11 measurement instruments found for migraine, only three were supported by evidence of two or more articles: The 3-question screen (10,41,59), the ID-migraine (12,34,40,44–47,49,50) and the Migraine Screen Questionnaire (11,51). Several studies introduced serious patient selection bias by only recruiting patients with the headache they were interested in studying (10). By doing so, there were no false positives or true negatives present, which resulted in more favourable diagnostic accuracy outcome measures. Other studies excluded participants who had a secondary headache (45), or who did not screen positive for a preliminary screening for migraine (45,46,49). One study selected their participants so 50% had a confirmed migraine diagnosis prior to the index test and 50% did not have migraine (11). This also introduced selection bias in favour of the outcomes, as the prevalence of the studied disorder (50% in the tested group versus 14.7% in the general population) determines the pre-test probability and thus the chance of correct diagnosis (65,66).

Furthermore, serious bias was introduced in the "flow and timing" section of the articles, as some articles did not properly describe the order of receiving the index test and the reference standard diagnosis. Other studies did not include all participants in the analysis (11,12,34,37,38,40,42,43,48,49,59). The introduced biases on both domains resulted in a downgrade of the certainty of evidence on all measurement instruments except for the Migraine Assessment Tool (35). However, as this tool is only studied in one article, the level of evidence was also downgraded for imprecision. Therefore, there are no measurement instruments for migraine with a high level of evidence.

Combined migraine and TTH measurement instruments

Out of the six measurement instruments that looked at both migraine and TTH, only the German language questionnaire is supported by two articles (13,57). However, due to a serious risk of bias and indirectness, there is only a low level of evidence for this questionnaire. In both studies, only patients with headaches that were also studied in the questionnaire were included, which introduced a serious selection bias (13,57). Similarly, the Computerized Headache Assessment Tool (CHAT) presented a sensitivity of 1.00 for both migraine and TTH, but no true negatives or false positives were available, and no specificity was presented (54). In this study, the gold standard was the diagnosis established by a headache nurse (54). As stated before, this is an unreliable gold standard for a headache diagnosis (63).

The seven articles differed in population. Some study samples were retrieved from the general population (53,55,56), others from urgent care or family practice (54), and others from a headache clinic (13,14). In one study, the sample origin was unclear (52). The prevalence used in the GRADE recommendations was for the general population, but in health care settings the prevalence is higher. This increases the pre-test probability of a positive headache diagnosis. This must be taken into consideration when interpreting the results of those studies (14,54,56).

Regarding the flow and timing of these studies, not all participants received both the index test and reference standard (52–54,56). Other studies did not include all participants in the final analyses (13,14,53,55). By excluding participants in these ways, the generalization of results is compromised. All these components resulted in very low to moderate level of evidence for the six combined migraine and TTH measurement instruments.

Cervicogenic headache measurement instruments

Both articles studying the diagnostic accuracy of the cervical flexion rotation test (CFRT) for CGH showed selection bias, as participants were selected based on headache type (57,58). In one study, the sensitivity and specificity were both 0.70 (57), whereas in the other study the sensitivity was 0.91 and the specificity 0.90 (58). In the study with lower diagnostic accuracy, the control group consisted of other headache forms (migraine or multiple headache forms) (57). This makes differentiating between headache types more difficult as other headaches are related to neck problems (5,67,68). The study with higher diagnostic accuracy compared patients with CGH with asymptomatic participants and several patients with migraine (58), which made it easier to recognize the CGH. When this test is applied in the clinic, patients will have a headache complaint and will not be asymptomatic, so the sensitivity and specificity of 0.70 will likely be more accurate.

Just as in the current review, another recent systematic review describing physical examination tests for screening and diagnosis of CGH, the CFRT was determined to be the most useful test with the highest reliability and strongest diagnostic accuracy (69). There is, however, a debate in the literature on the reliability of manual ROM tests of the spine (70). Inter-examiner reliability for the cervical spine passive ROM ranged from poor to substantial. The manual tests of the

upper cervical spine (C1/2, C2/3) have a fair to substantial level of reliability (70). The reliability of the CFRT has been established to be good to excellent (71). However, CFRT reliability was established by comparing a manual diagnosis of C1/2 dysfunction with the outcome of the CFRT (71). If the reliability of the manual diagnosis of dysfunction is only fair, then the reliability of the CFRT is questionable. However, in another study where the cervical ROM was measured with a device (CROM), a significant difference was found between the ROM in patients with CGH compared to patients with migraine and healthy subjects, which confirms the findings of the included papers of this review (57,58,72). In conclusion, the CFRT is a valid and reliable measure to recognize CGH, though the reliability is higher when using a CROM device rather than assessing the ROM manually.

Strengths and limitations of the study

The current review is, to the authors' knowledge, the first review establishing an overview of the diagnostic accuracy of measurement instruments for headaches associated with musculoskeletal symptoms. By using the QUADAS-2 and COSMIN tool, the methodological quality was assessed in a well-known and internationally accepted manner (24,25). By using the GRADE recommendations, the findings of this review are transparent and easy to translate to the clinical practice (27).

There are, however, also a few limitations of this study. Comparison between index and reference test was not easy, as the validation of the index test was performed in a different population compared to the population in which the reference standard was developed. It is important to keep in mind that the diagnostic accuracy is dependent on the prevalence of the target condition in the population; the study sample needs to be taken into consideration when interpreting the results. The prevalence of the target condition is the pre-test probability of a person having that condition, and a good measurement instrument will increase the chance of recognizing the target condition correctly. However, if the study sample is biased by having a very high prevalence in the target condition whereas the measurement instrument would normally be used in a setting with a low prevalence of the target condition, the diagnostic accuracy is not valid for that specific population. Validation studies of measurement instruments should therefore always test the measurement instrument in the population and setting for which it is being validated.

Also, some measurement tools were used in different languages and cultures, which must also be considered

when interpreting these results. In this review, great variability was found between the different studies, as illustrated in the S-ROC curves in Figure 3(a) and (c). These S-ROC curves show the uncertainty of the findings compared to reality, so the pooled data should be used with caution. The clear gap between the diagnostic accuracy of some measurement instruments between studies showed the necessity of conformation by multiple studies within the same population and against the same reference standard.

Implications for practice

The findings of the current review support the use of the ID-Migraine questionnaire to diagnose migraine with a moderate level of certainty (Table 4). However, patients with headaches often experience multiple headache forms (7,13,74). This warrants a measurement instrument that can diagnose more than one headache. From the questionnaires that looked at both migraine and TTH, the HSQ has the highest level of evidence within this review (Table 5). To establish if there is a migraine and/or a TTH present, this questionnaire is therefore recommended. As CGH needs to be confirmed by physical examination (15), the CFRT is recommended (Table 6). No other measurement instruments for secondary headache related to musculoskeletal complaints were found. Therefore, for these headache types, such as secondary headache attributed to temporomandibular disorders or headache attributed to whiplash injury, no recommendations can be made.

Implications for future research

Currently, there are many questionnaires for migraine and TTH, most of them validated by one study. Future research should use the recommended measurement instruments and validate them in different samples of the same population to increase the level of certainty that the diagnostic accuracy is realistic. The QUADAS-2 and COSMIN tools should be used when designing their studies to enhance their methodological quality.

Furthermore, additional clinimetric properties of measurement instruments for headache should be examined. Clinimetric properties such as reliability and responsiveness are important to enhance the care of headache complaints and monitor the course of these complaints. For that reason, the authors are conducting a complementary review to establish the clinimetric properties of measurement instruments for these symptoms and factors (Figure 2).

In conclusion, only a few measurement instruments reached a moderate level of evidence for the diagnostic

accuracy. For migraine, the ID-Migraine is recommended. For migraine and TTH, the HSQ is recommended, and the CFRT is advised to be used for

CGH. However, more studies are needed to validate these instruments further to enhance the level of evidence.

Article highlights

- ID-migraine is the most studied diagnostic accuracy measurement instrument for migraine and has a moderate level of certainty.
- Six measurement instruments are examined that establish the diagnostic accuracy for both migraine and tension-type headache.
- The Headache Screening Questionnaire has the highest level of evidence to screen for both migraine and tension-type headache.
- Only the Cervical Flexion Rotation Test studies the diagnostic accuracy for cervicogenic headache, but the level of evidence is very low.

Registration

This review is registered on PROSPERO (CRD42017062472).

ORCID iD

Hedwig A van der Meer (iD) http://orcid.org/0000-0002-6848-9629
Maria WG Nijhuis van der Sanden (iD) http://orcid.org/0000-0003-2637-6877

References

1. Hagen K, Einarsen C, Zwart J-A, et al. The co-occurrence of headache and musculoskeletal symptoms amongst 51 050 adults in Norway. *Eur J Neurol* 2002; 9: 527–533.
2. Bendtsen L, Ashina S, Moore A, et al. Muscles and their role in episodic tension-type headache: Implications for treatment. *Eur J Pain* 2016; 20: 166–175.
3. Fernandez-de-las-Peñas C, Pérez-de-Heredia M, Molero-Sánchez A, et al. Performance of the craniocervical flexion test, forward head posture, and headache clinical parameters in patients with chronic tension-type headache: A pilot study. *J Orthop Sports Phys Ther* 2007; 37: 33–39.
4. Fernandez-de-las-Peñas C, Cuadrado ML, Arendt-Nielsen L, et al. Myofascial trigger points and sensitization: An updated pain model for tension-type headache. *Cephalalgia* 2007; 27: 383–393.
5. Fernández-De-Las-Peñas C, Cuadrado ML and Pareja JA. Myofascial trigger points, neck mobility and forward head posture in unilateral migraine. *Cephalalgia* 2006; 26: 1061–1070.
6. Fernandes G, Franco AL, Goncalves DAG, et al. Temporomandibular disorders, sleep bruxism, and primary headaches are mutually associated. *J Orofac Pain* 2013; 27: 14–20.
7. van der Meer HA, Speksnijder CM, Engelbert RHH, et al. The association between headaches and temporomandibular disorders is confounded by bruxism and somatic complaints. *Clin J Pain* 2017; 33: 835–843.
8. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders, 3rd edition. *Cephalalgia* 2018; 38: 1–211.
9. Gaul C, Visscher CM, Bhola R, et al. Team players against headache: Multidisciplinary treatment of primary headaches and medication overuse headache. *J Headache Pain* 2011; 12: 511–519.
10. Cady RK, Borchert LD, Spalding W, et al. Simple and efficient recognition of migraine with 3-Question Headache Screen. *Headache* 2004; 44: 323–327.
11. Láinez MJA, Domínguez M, Rejas J, et al. Development and validation of the Migraine Screen Questionnaire (MS-Q). *Headache* 2005; 45: 1328–1338.
12. Lipton RB, Dodick D, Sadovsky R, et al. A self-administered screener for migraine in primary care. *Neurology* 2003; 61: 375–382.
13. Fritsche G, Hueppe M, Kukava M, et al. Validation of a German language questionnaire for screening for migraine, tension-type headache, and trigeminal autonomic cephalgias. *Headache* 2007; 47: 546–551.
14. van der Meer HA, Visscher CM, Engelbert RHH, et al. Development and psychometric validation of the

headache screening questionnaire – Dutch Version. *Musculoskelet Sci Pract* 2017; 31: 52–61.

15. Šimundić A-M. Measures of diagnostic accuracy: Basic definitions. *EJIFCC* 2009; 19: 203–211.

16. World Health Organization. International classification of functioning, disability and health, http://www.who.int/classifications/icf/en/ (2015, accessed 31 March 2018).

17. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med* 2009; 6: e1000097.

18. Olesen J, Burstein R, Ashina M, et al. Origin of pain in migraine: Evidence for peripheral sensitisation. *Lancet Neurol* 2009; 8: 679–690.

19. Bendtsen L and Fernández-De-La-Peñas C. The role of muscles in tension-type headache. *Curr Pain Headache Rep* 2011; 15: 451–458.

20. European Region of the World Confederation for Physical Therapy. *European physiotherapy benchmark statement*. Brussels, Belgium: ER-WCPT, 2003, pp.1–47.

21. Terwee CB, Jansma EP, Riphagen II, et al. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009; 18: 1115–1123.

22. Veritas Health Innovation. Covidence systematic review software, https://www.covidence.org (n.d., accessed 17 March 2017).

23. Whiting P, Rutjes AWS, Reitsma JB, et al. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BioMed Cent Med Res Methodol* 2003; 13: 1–13.

24. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155: 529–536.

25. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Qual Life Res* 2010; 19: 539–549.

26. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63: 737–745.

27. Schünemann HJ, Schünemann AHJ, Oxman AD, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; 336: 1106–1110.

28. Lumley T. rmeta: Meta-analysis, https://cran.r-project.org/package=rmeta (2012, accessed 10 January 2018).

29. Doebler P, Münster W and Holling H. Meta-analysis of diagnostic accuracy with mada. http://nbcgib.uesc.br/mirrors/cran/web/packages/mada/vignettes/mada.pdf (2015, accessed 10 January 2018).

30. Leeflang MMG. Systematic reviews and meta-analyses of diagnostic test accuracy. *Clin Microbiol Infect* 2014; 20: 105–113.

31. Reitsma JB, Glas AS, Rutjes AWS, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; 58: 982–990.

32. Doebler P. mada: Meta-analysis of diagnostic accuracy, https://cran.r-project.org/package=mada (2017, accessed 10 January 2018).

33. Deeks JJ, Macaskill P and Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005; 58: 882–893.

34. Lipton RB, Serrano D, Buse DC, et al. Improving the detection of chronic migraine: Development and validation of Identify Chronic Migraine (ID-CM). *Cephalalgia* 2016; 36: 203–215.

35. Marcus DA, Kapelewski C, Jacob RG, et al. Validation of a brief nurse-administered migraine assessment tool. *Headache* 2004; 44: 328–332.

36. Michel P, Dartigues J, Henry P, et al. Validity of the International Headache Society Criteria for Migraine. *Neuroepidemiology* 1993; 12: 51–57.

37. Michel P, Henry P, Letenneur L, et al. Diagnostic screen for assessment of the IHS criteria for migraine by general practitioners. *Cephalalgia* 1993; 13: 54–59.

38. Rueda-Sánchez M and Díaz-Martínez L. Validation of a migraine screening questionnaire in a Colombian university population. *Cephalalgia* 2004; 24: 894–899.

39. Shaik MM, Hassan NB, Tan HL, et al. Validity and reliability of the Malay version of the Structured Migraine Interview (SMI) Questionnaire. *J Headache Pain* 2015; 16: 1–9.

40. Siva A, Zarifoglu M, Ertas M, et al. Validity of the ID-Migraine screener in the workplace. *Neurology* 2008; 70: 1337–1345.

41. Wahab K, Ugheoke A, Okokhere P, et al. Validation of the 3-Question Headache Screen in the diagnosis of migraine in Nigeria. *Ethiop J Health Sci* 2016; 26: 5–8.

42. Walters A and Smitherman TA. Development and validation of a four-item migraine screening algorithm among a nonclinical sample: The Migraine-4. *Headache* 2016; 56: 86–94.

43. Wang S-J, Fuh J-L, Huang S-Y, et al. Diagnosis and development of screening items for migraine in neurological practice in Taiwan. *J Formos Med Assoc* 2008; 107: 485–494.

44. Brighina F, Salemi G, Fierro B, et al. A validation study of an Italian version of the "ID Migraine". *Headache* 2007; 47: 905–908.

45. de Mattos ACMT, de Souza JA, Moreira Filho PF, et al. ID-Migraine™ questionnaire and accurate diagnosis of migraine. *Arq Neuropsiquiatr* 2017; 75: 446–450.

46. Ertaş M, Baykan B, Tuncel D, et al. A comparative ID migraine screener study in ophthalmology, ENT and neurology out-patient clinics. *Cephalalgia* 2009; 29: 68–75.

47. Gil-Gouveia R and Martins I. Validation of the Portuguese version of ID-Migraine. *Headache* 2010; 50: 396–402.

48. Kallela M, Wessman M and Färkkilä M. Validation of a migraine-specific questionnaire for use in family studies. *Eur J Neurol* 2001; 8: 61–66.

49. Karli N, Ertas M, Baykan B, et al. The validation of ID migraine[TM] screener in neurology outpatient clinics in Turkey. *J Headache Pain* 2007; 8: 217–223.

50. Kim ST and Kim C-Y. Use of the ID Migraine questionnaire for migraine in TMJ and Orofacial Pain clinic. *Headache* 2006; 46: 253–258.

51. Láinez MJ, Castillo J, Domínguez M, et al. New uses of the Migraine Screen Questionnaire (MS-Q): Validation in the primary care setting and ability to detect hidden migraine. MS-Q in primary care. *BMC Neurol* 2010; 10: 39.

52. El-Sherbiny NA, Shehata HS, Amer H, et al. Development and validation of an Arabic-language headache questionnaire for population-based surveys. *J Pain Res* 2017; 10: 1289–1295.

53. Hagen K, Zwart JA, Aamodt AH, et al. The validity of questionnaire-based diagnoses: The third Nord-Trøndelag Health Study 2006–2008. *J Headache Pain* 2010; 11: 67–73.

54. Maizels M and Wolfe WJ. An expert system for headache diagnosis: The Computerized Headache Assessment tool (CHAT). *Headache* 2008; 48: 72–78.

55. Rasmussen BK, Jensen R and Olesen J. Questionnaire versus clinical interview in the diagnosis of headache. *Headache* 1991; 31: 290–295.

56. Yoon M-S, Obermann M, Fritsche G, et al. Population-based validation of a German-language self-administered headache questionnaire. *Cephalalgia* 2008; 28: 605–608.

57. Hall T, Briffa K, Hopper D, et al. Comparative analysis and diagnostic accuracy of the cervical flexion-rotation test. *J Headache Pain* 2010; 11: 391–397.

58. Ogince M, Hall T, Robinson K, et al. The diagnostic validity of the cervical flexion-rotation test in C1/2-related cervicogenic headache. *Man Ther* 2007; 12: 256–262.

59. Pryse-Phillips W, Aubé M, Gawel M, et al. A headache diagnosis project. *Headache* 2002; 42: 728–737.

60. Headache Classification Subcommittee of the International Headache Society. Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. *Cephalalgia* 1988; 8: 1–96.

61. Headache Classification Subcommittee of the International Headache Society. The International Classification Of Headache Disorders, 2nd edition. *Cephalalgia* 2004; 24: 1–160.

62. Sjaastad O, Fredriksen TA and Pfaffenrath V. Cervicogenic headache: Diagnostic criteria. The Cervicogenic Headache International Study Group. *Headache* 1998; 38: 442–445.

63. Beithon J, Gallenberg M, Johnson K, et al. *Diagnosis and treatment of headache 2013*. 11th ed. Bloomington, MN: Institute for Clinical Systems Improvement, 2013, p.90.

64. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: Assessing new tests against existing diagnostic pathways. *BMJ* 2006; 332: 1089–1092.

65. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: Communicating the performance of diagnostic tests. *Clin Biochem Rev* 2008; 29: S83–S87.

66. Stovner LJ and Andree C. Prevalence of headache in Europe: A review for the Eurolight project. *J Headache Pain* 2010; 11: 289–299.

67. Ashina S, Bendtsen L, Lyngberg AC, et al. Prevalence of neck pain in migraine and tension-type headache: A population study. *Cephalalgia* 2015; 35: 211–219.

68. Bevilaqua-Grossi D, Pegoretti KS, Goncalves MC, et al. Cervical mobility in women with migraine. *Headache* 2009; 49: 726–731.

69. Rubio-Ochoa J, Benítez-Martínez J, Lluch E, et al. Physical examination tests for screening and diagnosis of cervicogenic headache: A systematic review. *Man Ther* 2016; 21: 35–40.

70. van Trijffel E, Anderegg Q, Bossuyt P, et al. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: A systematic review. *Man Ther* 2005; 10: 256–269.

71. Hall T, Robinson K, Fuijnawa O, et al. Intertester reliability and diagnostic validity of the cervical flexion-rotation test. *J Manipulative Physiol Ther* 2008; 31: 293–300.

72. Zito G, Jull G and Story I. Clinical tests of musculoskeletal dysfunction in the diagnosis of cervicogenic headache. *Man Ther* 2006; 11: 118–129.

73. Valentine JC, Pigott TD and Rothstein HR. How many studies do you need? A primer on statistical power for meta-analysis. *J Educ Behav Stat* 2010; 35: 215–247.

74. World Health Organization. *Atlas of headache disorders and resources in the world*. Geneva, Switzerland: WHO, 2011, pp.1–72.

75. Schiffman E, Ohrbach R, List T, et al. Diagnostic criteria for headache attributed to temporomandibular disorders. *Cephalalgia* 2012; 32: 683–692.

76. Ravishankar K. The art of history-taking in a headache patient. *Ann Indian Acad Neurol* 2012; 15: S7–S14.

77. Sjaastad O and Bakketeig LS. Prevalence of cervicogenic headache?: Va study of headache epidemiology. *Acta Neurol Scand* 2008; 117: 173–180.