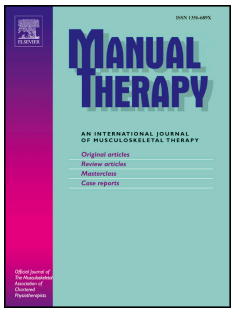
Accepted Manuscript



Physical Examination Tests for Screening and Diagnosis of Cervicogenic Headache: A Systematic Review

J. Rubio-Ochoa, J. Benítez-Martínez, E. Lluch, S. Santacruz-Zaragozá, P. Gómez-Contreras, C.E. Cook, Professor and Vice Chief of Research, Doctor of Physical Therapy Program

PII: S1356-689X(15)00182-4

DOI: 10.1016/j.math.2015.09.008

Reference: YMATH 1765

To appear in: *Manual Therapy*

Received Date: 17 June 2015

Revised Date: 22 August 2015

Accepted Date: 7 September 2015

Please cite this article as: Rubio-Ochoa J, Benítez-Martínez J, Lluch E, Santacruz-Zaragozá S, Gómez-Contreras P, Cook C, Physical Examination Tests for Screening and Diagnosis of Cervicogenic Headache: A Systematic Review, *Manual Therapy* (2015), doi: 10.1016/j.math.2015.09.008.

Physical Examination Tests for Screening and Diagnosis of Cervicogenic Headache: A Systematic Review

Rubio-Ochoa J[1]; Benítez-Martínez J[1]; Lluch E[1, 2, 3, 4]; Santacruz-Zaragozá S[1]; Gómez-Contreras P[1]; Cook CE[5]

[1]Department of Physical Therapy, Faculty of Physiotherapy, University of Valencia, Spain

[2]Department of Human Physiology and Physiotherapy, Faculty of Physical Education & Physiotherapy, Vrije University Brussels, Belgium

[3]Department of Physical Medicine and Physiotherapy, University Hospital Brussels, Belgium

[4]Pain in Motion Research Group, www.paininmotion.be

[5]Professor and Vice Chief of Research, Doctor of Physical Therapy Program, Duke University, 2200 W. Main St. Durham, North Carolina, USA.Chad.cook@duke.edu. 919 684 8905

Address for correspondence: Enrique Lluch Girbés, Department of Physical Therapy, University of Valencia, Gascó Oliag,5, 46010 Valencia, Spain. phone: ++034 3983853 E-mail: enrique.lluch@uv.es.

ABSTRACT

It has been suggested that differential diagnosis of headaches should consist of a robust subjective examination and a detailed physical examination of the cervical spine. Cervicogenic headache (CGH) is a form of headache that involves referred pain from the neck. To our knowledge, no studies have summarized the reliability and diagnostic accuracy of physical examination tests for CGH. The aim of this study was to summarize the reliability and diagnostic accuracy of physical examination tests used to diagnose CGH.A systematic review following PRISMA guidelines was performed in four electronic databases (MEDLINE, Web of Science, Embase and Scopus). Full text reports concerning physical tests for the diagnosis of CGH which reported the clinometric properties for assessment of CGH, were included and screened for methodological quality. *Quality Appraisal for Reliability Studies* (QAREL) and *Quality Assessment of Studies of Diagnostic Accuracy* (QUADAS-2) scores were completed to assess article quality. Eight articles were retrieved for quality assessment and data extraction. Studies investigating diagnostic reliability of physical examination tests for CGH scored poorer on methodological quality (higher risk of bias) than those of diagnostic accuracy. There is sufficient evidence showing high levels of reliability and diagnostic accuracy of the selected physical examination tests for the diagnosis of CGH. The cervical flexion-rotation test (CFRT) exhibited both the highest reliability and the strongest diagnostic accuracy for the diagnosis of CGH.

Keywords: Cervicogenic headache; physical examination; diagnostic accuracy; reliability.

INTRODUCTION

Headache is a common disorder affecting up to 66% of the general population (Stovner et al., 2007). With an estimated lifetime prevalence of 96% (Rasmussen et al., 1991), headaches negatively influence both quality of life and labor productivity (Lipton and Stewart 1994; Diener et al., 2001; van Suijlekom et al., 2003). The individual and socio-economic burden, which consists of direct costs (associated with pursuance of healthcare) and indirect costs [related with sickness leave and reduced productivity (Pradalier et al., 2004)] of headaches around the world is substantial (Rasmussen, 1999).

The International Headache Society categorizes headaches into *primary* and *secondary* classifications (IHS, 2004). Primary headaches are the most common and are often defined as idiopathic, suggesting that these often occur without an underlying disease or process. Secondary headaches may be a consequence of a serious underlying disease such as a brain tumor, aneurysm, infection, substance abuse or withdrawal, or inflammatory disease; but may present as referred pain from other regional structures such as the teeth, nose, ears, or neck. One type of secondary headache is cervicogenic headache (CGH), which refers to a headache resulting from musculoskeletal dysfunction of the cervical spine, particularly the upper three cervical segments (Bogduk, 1994; Jull 2002a; Zito et al., 2006). CGH constitutes about 15-20% of all chronic and recurrent headaches (Nilsson, 1995).

The complex neurophysiological interactions within the cervical-trigeminal nucleus are the cause of the referral of pain to regions of the head (Bogduk et al., 1997). The interface between the trigeminal afferent and efferent processes from the three upper cervical nerves is bidirectional (Bartsch and Goadsby, 2002, 2003), which also explains why cervical pain is not an exclusive feature of CGH. This bidirectional mechanism

creates similar referred pain from the cervical spine in other forms of headaches such as migraine or tension-type headache (Hagen et al., 2002). The overlap in signs and symptomatology of CGH with other forms of headaches greatly complicates an appropriate diagnosis, leading to incorrect diagnoses in approximately 50% of cases of CGH (Pfaffenrath and Kaube, 1990). Thus, correct headache diagnosis is mandatory in order to establish an appropriated treatment, especially considering that CGH is the headache classification that most commonly responds positively to long-term physiotherapy treatment (Jull et al., 2002b; Bronfort al., 2004).

Because of the overlap between signs and symptoms of the different types of headache (D'Amico et al., 1994; Nicholson and Gaston, 2001) it has been suggested that differential diagnosis should consist of a robust subjective examination (Sjaastad et al, 1998; IHS 2004, 2013) as well as a detailed physical examination of the cervical spine (Hall et al, 2008a). In this regard, some studies have documented the presence of specific cervical spine musculoskeletal dysfunction in patients with CGH (Hall and Robinson, 2004; Zito et al., 2006). To our knowledge, no studies have synthesized the utility of physical examination testing of the cervical spine and its influence on CGH. Consequently, the objective of this study was to review the available evidence regarding the physical examination tests used for diagnosis of CGH. In particular, we were interested in evaluating the utility of the physical examination by scrutinizing the reliability and diagnostic accuracy of the selected tests.

METHODS

Search strategy

This systematic review was written in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement for reporting systematic reviews and meta-analyses of studies (Liberati et al., 2009) and the Cochrane Diagnostic Accuracy Group recommendations. To identify relevant articles concerning the study objective, a systematic search was performed in four electronic databases [MEDLINE, Web of Science, EMBASE and Scopus] from each databases' inception until June 2015. A search strategy was built using the following keywords: *"cervicogenic headache" AND (diagnosis (MeSH) OR "diagnostic accuracy") AND ("physical assessment" OR "physical examination")*. Relevant hand searched articles were also included to obtain as complete information as possible.

Study selection

Articles were eligible for this systematic review if each fulfilled the following inclusion criteria: (I) the authors studied at least one physical test for the diagnosis of CGH in humans; (II) the clinometric properties (e.g., reliability, sensitivity or specificity) of the test used to assess CGH were reported or data were provided to allow for individual calculation; (III) articles included full text reports of original studies; and (IV) studies were published in English or Spanish. Physical examination tests were operationally defined as clinician performed tests or measures that were designed to be a proxy for a diagnosis or impairment.

Selection process

After performing the literature search, duplicate articles were removed. Eligibility assessment was performed based on title and abstract. The full-text article was searched and analyzed when the article seemed to fulfill the inclusion criteria. When there was uncertainty regarding the content of the paper based on title and abstract, the full text was read and evaluated against the inclusion criteria. Screening was performed by two researchers independently (JR and SS). A consensus meeting was organized to discuss potential disagreements. When consensus could not be reached, a third opinion was provided by a trained experienced researcher (CC). The full text versions of all articles that met the inclusion criteria were retrieved for methodological quality assessment and data extraction.

Quality assessment

Two independent researchers evaluated the quality of two forms of studies; reliability and diagnostic accuracy. The *Quality Appraisal for Reliability Studies* (QAREL) checklist is an 11 item appraisal tool recently developed to assess the quality of studies of diagnostic reliability (Lucas et al., 2010).

Quality of the diagnostic accuracy studies was evaluated using the *Quality Assessment of Diagnostic Accuracy Studies* (QUADAS-2) scale (Whiting et al., 2011). QUADAS-2 provides assessment opportunities in four key areas: patient selection, clinical trial studied, standard reference and flow and timing. In addition, clinical applicability of a study is evaluated based on selection of patients, test analyzed and reference standard. For both categories (i.e. risk of bias and applicability analysis), each study criteria was classified as "low risk", "high risk" or "unclear". In both assessments,

the reviewers reached a definitive score during a consensus meeting, resulting in a final quality score.

Risk of bias assessment

Risk of bias was defined as the risk of a systematic error or deviation from the truth, in the results or inferences in each study. In particular, we qualitatively evaluated the internal validity of each study for a believability assessment of the results. Risk of bias assessment differs from quality assessment, as it represents "the extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error" (Viswanathan et al., 2012; Higgins et al., 2011).

Tabulation of the diagnostic clinometrics

All included studies needed to incorporate the same diagnostic criteria (IHS, 2004) and present data for analysis of reliability and/or diagnostic accuracy of clinical tests. For our study, the diagnostic reliability of a clinical test was determined by the Kappa coefficient indicating consistency between different evaluators to identify cervical dysfunction (Hall et al., 2010a) or Intraclass correlation coefficient (ICC), which involves the reliability of multiple measurements or ratings. Cohen suggested the following Kappa value interpretations: values ≤ 0 equals no agreement, whereas 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement (Cohen, 1960). ICC values were interpreted as follows: >0.75 was excellent, 0.40–0.75 was fair to good and <0.40 was poor (Fleiss 1986)

The diagnostic accuracy of clinical tests was determined based on the sensitivity, specificity, positive likelihood ratio (LR+) and/or negative likelihood ratio (LR-). Sensitivity is defined as the percentage of subjects who test positive for a specific disease among a group of individuals who have the disease, whereas the specificity is the percentage of subjects with a negative result for a specific disease among a group of individuals who don´t have that disease (Cook and Hegedus, 2011). A higher value for LR+ indicates that a test is able to confirm the presence of a finding when the result is positive. A lower LR- suggests a test is useful in ruling out a diagnosis when the test is negative. For clinical practice, we used values that have been advocated previously: sensitivity >90% with LR- <0.2 for a test to be useful for ruling out disorders and specificity>90% with LR+>5 for a test to confirm a specific diagnosis (Cook and Hegedus, 2008).

RESULTS

Study selection

The selection process of the articles is presented in Figure 1. The initial search resulted in 220 hits (26 in MEDLINE, 33 in Web of Science, 107 in EMBASE and 36 in Scopus) and, after removing duplicates, 118 studies remained. From these, 113 studies were excluded after screening based inclusion and exclusion criteria. References from our reference lists and independent hand search revealed an additional 4 articles, thus 9 articles were finally retrieved for quality assessment and data extraction.

In this review, three studies included manual examination tests from the upper cervical spine (Jull et al., 1997; van Suijlekom et al., 2000; Hall et al., 2010a), four studies were based on the cervical flexion-rotation test (CFRT) (Ogince et al., 2007; Hall et al., 2008; Hall et al., 2010b; Hall et al., 2010c) and one study was based on a combination of tests to diagnose CGH (Jull et al., 2007b).

Quality assessment of individual studies

Five studies met the inclusion criteria for quality assessment of reliability (Table 1). One study exhibited poor reliability results, yielding a score of 4/11 (van Suijlekon et al., 2000). Four of five studies scored unclear scores for Item 4 “Were raters blinded to their own prior findings of the test under evaluation?” and Item 7 “Were raters blinded to additional cues that were not part of the test?”. Item one resulted in the greatest number of “no” scores, suggesting that the samples used did not reflect subjects typically seen in clinical practice.

All four of the diagnostic accuracy studies exhibited low risk of bias in the majority of QUADAS-2 categories (Table 2). Three of the four studies exhibited poor

quality for risk of bias, patient selection. One study (Hall et al., 2010b) exhibited no risks of bias for any of the QUADAS 2 categories.

Clinometric results

All studies included clinical criteria provided by the IHS (2004) as a requirement for the diagnosis of CGH. Five articles studied the reliability of the physical tests for the diagnosis of CGH (Table 3). Passive accessory intervertebral movements (PAIVMs) tests C0-C3 were used in two studies with Kappa values ranging from 0.53 to 0.72 (Jull et al., 1997) and 0.64 to 0.7 (Hall et al., 2010a). In another study (van Suijlekom et al., 2000), Kappa values ranged from 0.08 to 0.89 for manual examination of the cervical spine and range of motion assessment. Hall et al (2008b) studied the reliability of the CFRT obtaining values ranged from 0.67 to 0.85 for a prevalence adjusted kappa cross-sectionally among raters. Hall and colleagues (2010c) measured the longitudinal reliability of the CFRT, finding excellent reliability when testing both left (ICC=0.97; 95%CI=0.94, 0.99) and right (ICC=0.95; 95%CI=0.90, 0.98) movements.

Five studies studied diagnostic accuracy of physical tests for the diagnosis of CGH (Table 4). Zito et al. (2006) studied the values of diagnostic accuracy of the PAIVMs tests C0-C3 obtaining sensitivity values between 59 and 65%, specificity between 78 and 87%, LR + from 2.9 to 4.9 and LR - from 0.43 to 0.49. Another study (Jull et al., 2007b) showed a sensitivity of 100% and specificity of 94.4% by clustering cervical range of motion, manual examination C0-C3 and the cranio-cervical flexion test. Three studies examined the CFRT (Ogince et al., 2007; Hall et al., 2008b; Hall et al., 2010b). The sensitivity for this test ranged from 70 to 91.3% and specificity from

70 to 92%. CFRT exhibited a LR+ higher than 5 and a LR – lesser than 0.2, indicating the ability to alter significantly the post-test probability (Table 4).

Risk of bias

All studies presented with small sample sizes and/or asymptomatic individuals as control subjects. Asymptomatic controls have the risk of inflating diagnostic accuracy and increasing the level of reliability found in a study. Symptomatic subjects were appropriate for all studies included.

DISCUSSION

The purpose of this review was to investigate the reliability and diagnostic accuracy of the selected clinical tests for the diagnosis of CGH. In our review, the tests that exhibited the highest reliability included the PAIVMs tests C1-C2 (Jull et al., 1997; Hall et al., 2010a) and the CFRT (Hall et al., 2008). The most commonly investigated tests for diagnostic accuracy included the CFRT (Ogince et al., 2007; Hall et al., 2008; Hall et al., 2010b), differentiation tests of PAIVMs (Zito et al., 2006) and the cluster cervical range of motion, manual examination C0-C3,and the cranio-cervical flexion test (Jull et al., 2007b). Whereas all of the tests exhibited good diagnostic accuracy utility, the strongest diagnostic accuracy metrics were associated with the CFRT.

The agreement among clinicians who used the CGH physical examination tests presents values that were better than those identified by random chance expected (Landis and Koch, 1977; Sim and Wright, 2005). In our synthesis, the studies that demonstrated lower levels of reliability and validity for manual screening of the cervical spine (Seffinger et al., 2004; van Trijffel et al., 2005) scored poorer on methodological quality as well. Indeed, the lower reliability metrics may be associated with design quality, a finding that has been identified by others regarding manual examination of the spine (Stochkendahl et al., 2006).

In our review, two studies (Jull et al., 1997; Hall et al., 2010a) demonstrated high reliability (Kappa 0.68 and 0.74 PABAK) for the manual examination of PAIVMs, and identified the segment C1-C2 as the most common symptomatic segment (63% of positive cases in subjects with CGH). This finding is consistent with previous studies (Hall et al., 2004; Zito et al., 2006; Hall et al., 2010a), where C1-C2 was the most prevalent symptomatic segment with up to 72% of cases showing positive results. Both studies exhibited high QAREL scores of 7/11 (Jull et al., 1997) and 9/11 (Hall et al.,

2010a) and were considered to be low risk of bias. Furthermore, Hall et al (2008b) investigated the CFRT showing high levels of reliability among experienced examiners (Kappa 0.85) and inexperienced examiners (Kappa 0.67), a finding in line with previously published studies (Hall et al., 2004; Ogince et al., 2007). We feel these findings support that the CFRT should be considered a useful clinical test to evaluate movement dysfunction at the C1-C2 segment and can assist in the differential diagnosis of CGH (Hall et al., 2008b).

According to our study, the test that has shown greater diagnostic accuracy for CGH was the CFRT (Ogince et al., 2007; Hall et al., 2008b; Hall et al., 2010b). Hall et al (2008b) reported a sensitivity of 90% and specificity of 88% for the CFRT. This is a similar finding to other studies (Hall et al., 2004; Ogince et al., 2007), where sensitivities and specificities of 91% (Ogince et al., 2007) and 86% and 100% (Hall et al., 2004) were shown. Both studies indicated an average value for a positive test in 33º rotation of the C1-C2 segment for CFRT. It is possible that the higher values from Ogince et al (2007) were associated with increased risk of bias and the use of asymptomatic subjects. Asymptomatic controls may overvalue the accuracy of the results in the absence of a comparison group with involvement of the cervical spine in a headache. For example, Hall et al (2010b) compared subjects with CGH to individuals with multiple forms of migraine headache (MFMH) and, although the CFRT still demonstrated good diagnostic utility, the diagnostic accuracy findings were not nearly as robust at studies in which asymptomatic controls were used.

All data investigated in this review have shown that physical examination of the upper cervical spine has a good utility for differential diagnosis in headache. We advocate for the use of physical examination testing in a stepwise fashion in clinical practice. Use of the IHS criteria (2004, 2013) can help us formulate our first hypotheses

during the subjective examination. Later, at the beginning of the physical examination using a test with good reliability, high sensitivity and a low negative likelihood ratio such as PAIVMs C0-C3 testing is recommended (Zito et al., 2006). When confirming a finding, the use of a reliable test with high specificity and high positive likelihood ratio such as the CFRT (Ogince et al., 2007; Hall et al., 2008b; Hall et al., 2010b) can be used near the end of the examination. Using the appropriate tests in the appropriate order can bring us a reliable differential diagnosis of CGH in a non-invasive way.

Possible limitations in our study are the small number of studies of reliability and diagnostic accuracy of clinical tests for CGH that have been included in this review. In addition, some studies have shown a low score on the QAREL scale (van Suijlekom et al., 2000). This fact may have overvalued some reliability results. The same occurred with diagnostic accuracy studies, where three of the four studies evaluated showed a high risk of bias QUADAS-2 in the selection of the sample. All studies included in this review have had a design based case-control study cases except one (Ogince et al., 2007). Case-control designs may overvalue reliability and diagnostic accuracy data (Lijmer et al., 1999; Kelly et al., 2001), and thus may have biased the results of this review. Lastly, all studies in our review used the IHS criteria as inclusion criteria for their own enrollments. Since the IHS criteria requires that the headache resolve within 3 months of treatment of the causative factor or lesion there is a risk that some of the patients in the articles included may have been unintentionally misdiagnosed; since none of the articles actually looked at resolution of symptoms.

Conclusion

There is sufficient evidence showing high levels of reliability and diagnostic accuracy of the selected physical examination tests for the diagnosis of CGH. The

CFRT has better level of evidence and highest values of validity, reliability and diagnostic accuracy for use in the differential diagnosis of CGH. Therefore, the clinical tests selected for evaluation of the upper cervical spine can be used by therapists in a reliable and accurate way for the diagnosis of CGH. More high quality case-based, case control studies in relation to the prevalence of CGH in different groups of population are necessary.

REFERENCES

Bartsch T, Goadsby P. Stimulation of the greater occipital nerve induces increased central excitability of dural afferent input. Brain 2002;125:1496-509.

Bartsch T, Goadsby P. Increased responses in trigemino cervical nociceptive neurons to cervical input after stimulation of the duramater. Brain 2003;126:1801-3.

Bogduk N. Cervical causes of headache and dizziness. In: Boyling J, Palastanga N, editors. *Grieves Modern Manual Therapy*. 2nd ed. Edindurgh: Churchill Livingstone; 1994.p.317-32.

Bogduk N. Headache and the neck. In: Goadsby P & Silberstein S, eds. Headache. Melbourne, Australia: Butterworth-Heinemann, 1997.

Bronfort G, Nilsson N, Haas M, Evans R, Goldsmith CH, Assendelft WJ, et al. Non-invasive physical treatments for chronic/recurrent headache. Cochrane Database Syst Rev 2004:CD001878.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.

Cook C, Hegedus EJ. Orthopedic physical examination tests: an evidence based approach. Upper Saddle River, NJ: Prentice Hall; 2008.

Cook C, Hegedus EJ. Diagnostic utility of clinical tests for spinal dysfunction. Man Ther 2011;16: 21-25.

D'Amico D, Leone M, Bussone G. Side-locked unilaterality and pain localization in long-lasting headaches-migraine, tension-type headache, and cervicogenic headache. Headache 1994;34(9):526-30.

Diener I. The impact of cervicogenic headache on patients attending a private physiotherapy practice in Cape Town. S Afr J Physiother 2001;57:35-39.

Fleiss J. The design and analysis of clinical experiments. New York: John Wiley & Sons, 1986.

Hagen K, Einarsen C, Zwart J, Svebak S, Bovim G. The co-occurrence of headache and musculoskeletal symptoms amongst 51 050 adults in Norway. Eur J Neurol2002; 9:527-33.

Hall T, Robinson K. The flexion-rotation test and active cervical mobility: a comparative measurement study in cervicogenic headache. Manual Therapy 2004;9(4):197-202.

Hall T, Briffa K, Hopper D. Clinical evaluation of cervicogenic headache. J Man Manip Ther 2008a;16:73–80.

Hall TM, Robinson KW, Fujinawa O, Akasaka K, Pyne EA. Intertester Reliability and Diagnostic Validity of the Cervical Flexion-Rotation Test. J Manipulative Physiol Ther 2008b;31(4):293-300.

Hall T, Briffa K, Hopper D, Robinson K. Reliability of manual examination and frequency of symptomatic cervical motion segment dysfunction in cervicogenic headache. Man Ther 2010a;15(6):542-546.

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

Hall T, Briffa K, Hopper D, Robinson K. Long-term stability and minimal detectable change of the cervical flexion-rotation test. J Orthop Sports Phys Ther 2010c;40:225-229.

Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of interventions version 5.1.0. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

International Headache Society. The International Classification of Headache Disorders.2nd ed. Cephalalgia2004;24Suppl 1:9-160.

International Headache Society. The International Classification of Headache Disorders.3rd ed. Cephalalgia 2013;33(9):629-808.

Jull G, Zito G, Trott P, Potter H, Shirley D, Richardson C. Inter-examiner reliability to detect painful upper cervical joint dysfunction. Aust J Physiother 1997;43:125–9.

Jull G. Management of cervicogenic headache. In: Grant R, editor. Physical Therapy of the cervical and thoracic spine.3rd edition. New York: Churchill Livingstone;2002a.p.239-272.

Jull G, Trott P, Potter H. A randomized controlled trial of exercise and manipulative therapy for cervicogenic headache.Spine.2002b; 27:1835-1843; discussion 1843.

Jull G, Amiri M, Bullock-Saxton J, Darnell R, Lander C. Cervical musculoskeletal impairment in frequent intermittent headache, part 1: subjects with single headaches. Cephalalgia 2007;27(7):793-802.

Kelly S, Harris KM, Berry E, Hutton J, Roderick P, Cullingworth J, et al. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. Gut 2001;49:534e9.

Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74

Liberati A, Altman DG, Tetzlaff J,Mulrow C, Gotzsche PC, Ioannidis LPA. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6(7):e1000100.

Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. Journal of the American Medical Association1999;282:1061e6.

Lipton RB, Stewart WF. The epidemiology of migraine. EurNeurol1994; 34Suppl 2:6-11.

Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL).J ClinEpidemiol.2010;63(8):854-61.

Nicholson GG, Gaston J. Cervical headache. J Orthop Sports PhysTher2001;31:184-193.

Nilsson N. The prevalence of cervicogenic headache in a random population sample of 20–59 year olds. Spine 1995;20:1884-8.

Ogince M, Hall T, Robinson K, Blackmore AM. The diagnostic validity of the cervical flexion-rotation test in C1/2-related cervicogenic headache. Man Ther 2007;12(3):256-262.

Pfaffenrath V, Kaube H. Diagnostics of cervicogenic-headache. FunctNeurol1990;5:159-164.

Pradalier A, Auray JP, El Hasnaoui A, et al. Economic impact of migraine and other episodic headaches in France: data from the GRIM 2000 study. Pharmacoeconomics 2004; 22(15):985-99.

Rasmussen BK, Jensen R, Schroll M, Olesen J. Epidemiology of headache in a general population—a prevalence study. Journal of Clinical Epidemiology 1991;44:1147-57.

Rasmussen BK. Epidemiology and socio-economic impact of headache. Cephalalgia.1999;25:20-3.

Seffinger MA, Najm WI, Mishra SI, Adams A, Dickerson VM, Murphy LS, Reinsch S. Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. Spine (PhilaPa 1976) 2004;29(19):E413e25.

Sjaastad O, Fredriksen TA, Pfaffenrath V. Cervicogenic headache: diagnostic criteria. Headache 1998;38:442-5.

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85(3):257-68.

Stochkendahl M, Christensen H, Hartvigsen J, Vach W, Haas M, Hestbaek L, Adams A, Bronfort G. Manual examination of the spine: a systematic review of reproducibility. J ManipulativePhysiolTher2006;29:475-85.

Stovner L, Hagen K, Jensen R, et al. The global burden of headache: A documentation of headache prevalence and disability worldwide. Cephalalgia2007; 27:193-210.

Van Suijlekom HA, De Vet HC, Van Den Berg SG, Weber WE. Interobserver reliability in physical examination of the cervical spine in patients with headache. Headache 2000;40(7):581-586.

Van Suijlekom HA, Lame I, Stomp-van den Berg SG, Kessels AG, Weber WE. Quality of life of patients with cervicogenic headache: a comparison with control subjects and patients with migraine or tension-type headache. Headache 2003;43:1034-1041.

vanTrijffel E, Anderegg Q, Bossuyt PM, Lucas C. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. Man Ther 2005;10(4):256e69.

Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews.

March 2012. AHRQ Publication No. 12-EHC047-EF. Available at: www.effectivehealthcare.ahrq.gov/

Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155:529–36.

Zito G, Jull G, Story I. Clinical tests of musculoskeletal dysfunction in the diagnosis of cervicogenic headache. Man Ther 2006; 11(2):118-129.

1 **FIGURE LEGENDS**

2

3 **Figure 1**: Flow chart study selection.

Table 1. Quality Appraisal of Diagnostic Reliability (QAREL) Checklist

Study	Type	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Total "yes" scores
Jull et al., 1997	Inter	N	Y	U	U	Y	Y	U	Y	Y	Y	Y	7/11
van Suijlekom et al., 2000	Inter	Y	Y	U	U	U	N	U	U	Y	U	Y	4/11
Hall et al., 2008	Inter	N	Y	Y	U	Y	Y	U	Y	Y	Y	Y	8/11
Hall et al., 2010a	Inter	N	Y	Y	U	Y	Y	Y	Y	Y	Y	Y	9/11
Hall et al., 2010c	Intra	N	Y	N	Y	Y	Y	U	U	Y	Y	Y	7/11

Scoring: Y=Yes, N=No, U=Unclear, and N/A=Not applicable

Item 1: Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied?
Item 2: Was the test performed by raters who were representative of those to whom the authors intended the results to be applied?
Item 3: Were raters blinded to the findings of other raters during the study?
Item 4: Were raters blinded to their own prior findings of the test under evaluation?
Item 5: Were raters blinded to the results of the reference standard for the target disorder (or variable) being evaluated?
Item 6: Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design?
Item 7: Were raters blinded to additional cues that were not part of the test?
Item 8: Was the order of examination varied?
Item 9: Was the time interval between repeated measurements compatible with the stability (or theoretical stability) of the variable being measured?
Item 10: Was the test applied correctly and interpreted appropriately?
Item 11: Were appropriate statistical measures of agreement used?

Table 2. Tabular Presentation for QUADAS-2 Results

Study	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Jull et al., 2007b	☹	☺	☹	☺	☺	☹	☺
Ogince et al., 2007	☹	☺	☺	☺	☺	☺	☺
Zito et al., 2006	☹	☺	☺	☺	☺	☺	☺
Hall et al., 2010c	☺	☺	☺	☺	☺	☺	☺

☺ = Low Risk; ☹ = High Risk' ? = Unclear

Table 3. Reliability findings of clinical tests for cervicogenic headache. CGH, cervicogenic headache; PAIM, passive accessory intervertebral movement; CFRT, Cervical flexion-rotation test.

Study	Study design	Subjects	Clinical test assessment	Results
Jull et al., 1997	Cross-sectional	20 with CGH and 20 without CGH	PAIM test CO-C1 PAIM test Cl-C2 PAIM test C2-C3	0.72 Kappa* 0.68 Kappa* 0.53 Kappa*
van Suijlekom et al., 2000	Cross-sectional	24 subjects Group A: CGH subjects Group B: migraine subjects Group C: tension-type headache subjects	Range of movement Head pain provocation Pressure pain zygapophyseal joint Pressure pain occiput Pressure pain mastoid process	0.44 and 0.46 Kappa** 0.53-0.67 Kappa 0.27 Kappa*** 0.08 Kappa*** 0.89 Kappa***
Hall et al., 2008b	2 single blind comparative measurement	**Study 1** Group A:20 subjects CGH with C1-C2 dysfunction group B:10 subjects CGH with different dysfunctional levels than C1/2 Group C: 10 asymptomatic controls **Study 2** Group A:12 subjects CGH Group B:12 asymptomatic controls	CFRT 2 experienced examiners CFRT 2 inexperienced examiners	0.85 Kappa 0.67 Kappa
Hall et al., 2010a	Cross-sectional	Group A: 60 subjects with CGH Group B: 20 asymptomatic controls	PAIM test C1-C2 PAIM test C2-C3	0.64 Kappa/ 0.74 PABAK**** 0.7 Kappa/ 0.7 PABAK****
Hall et al., 2010c	Longitudinal (tested 4 times over a 14 day period)	Group A: 15 subjects with CGH Group B: 15 asymptomatic controls	CFRT	ICC (95%CI) 0.95 (0.90, 0.98) Right ICC (95%CI) 0.97 (0.94, 0.99) Left

*Average value between examiners
**Rotation right and left
***Average value
****Adjusted Kappa coefficient

Table 4. Diagnostic accuracy of clinical tests for cervicogenic headache. PAIM, passive accessory intervertebral movement; CFRT, Cervical flexion-rotation test.

Study	Clinical test assessment	Sensitivity / specificity	LR + / LR -
Hall et al., 2008b	CFRT	90 / 85-90 (study 1)* 83 / 83-92 (study 2)**	6-9 / 0.11-0.12 (study 1)* 5-10/ 0.18-0.2 (study 2)**
Hall et al., 2010b	CFRT	70 / 70	2.33 / 0.43
Ogince et al., 2007	CFRT	91.3 / 91.4	10.65 / 0.095
Zito et al., 2006	PAIM test C0-C1 PAIM test C1-C2 PAIM test C2-C3	59 / 82 62 / 87 65 / 78	3.3 / 0.49 4.9 / 0.43 2.9 / 0.44
Jull et al., 2007b	Cervical range of motion, manual examination C0-C3 and cranio-cervical flexion test	100 / 94.4	- / -

*experienced examiners
**inexperienced examiners