Journal of Physiotherapy and Rehabilitation

Research Article

A SCITECHNOL JOURNAL

Inter-Rater Reliability for a Recently Developed Cluster of Headache Assessment Tests

Luedtke, K[1,4], Starke W[1,2], May A[1], Schoettker-Koeniger T[2] and Schaefer A[3]

Abstract

Objectives

An important component of assessment-based physiotherapy management of patients with headache is a thorough physical examination to identify musculoskeletal dysfunctions. A recently published international consensus study identified 11 clinically useful headache assessment tests (HATs). Test properties are not yet documented for all HATs. Different rating methods include absolute values and a 0-100 visual analog scale (VAS) to indicate the clinical relevance of a test.

Methods

To evaluate the inter-rater reliability of different scales, two manual therapists, blinded towards the diagnosis, examined 25 patients with headache and 25 headache-free controls. Inter-rater reliability was evaluated using the intraclass correlation coefficient and Bland-Altman plots for interval data; Cohen's kappa and Gwet´s AC for categorical data.

Results

Substantial to excellent reliability (Gwet´s AC or ICC > 0.7) was observed for flexion-rotation test, muscle strength, upper cervical quadrant, the examination of active range of motion and reproduction and resolution of symptoms. Forward-head posture, cranio-cervical flexion test, passive accessory intervertebral movements and the observation of latent trigger points, showed moderate levels of interrater agreement (Gwet´s AC or ICC> 0.5), all other tests showed only low levels of agreement.

Discussion

We recommend a 0-100 scale to indicate the clinical relevance of a test result. Substantial to excellent reliability (ICC > 0.7) was observed for active range of motion, trigger point palpation and upper cervical quadrant. Reliability was higher in the headache group. VAS can be handled as interval data; for clinical purposes it can be dichotomised at a cut-off point at 20 mm to indicate a positive or negative test result.

Keywords

Migraine; Headache; Physical examination; Physiotherapy; Manual; Musculoskeletal; Muscle; Joints

Introduction

The lifetime prevalence of any headache is more than 90% in the European population, indicating the relevance and burden of headache [1]. Tension-type headache (TTH) and migraine are the most common recurrent headache types with prevalence's of 20.77% and 14.7%, respectively [2]. Headache compromises a person's quality of life, their functional capability and their work capacity and results in high socioeconomic cost mainly due to days off work [3-9]. While the diagnosis of headache mainly relies on the subjective examination or patient interview, an important component of patient-centered, assessment-based physiotherapy management is a thorough physical examination. Hall et al. propose that only patients with a cervical movement dysfunction are likely to respond to physiotherapy and highlight the importance of physical examination to identify such treatable dysfunctions [10]. In a systematic review, Howard et al. provide an overview of published manual examination techniques in the diagnosis of cervicogenic headache [11]. The tests included are: flexion-rotation test (FRT), active range of cervical motion (AROM), passive accessory intervertebral movements (PAIVMs), passive physiological intervertebral movements (PPIVMs), cervical muscle strength, cross-sectional area measurements of cervical extensors, cranio-cervical flexion test (CCFT), palpation for trigger-points (TPs), pressure pain thresholds, and cervical kinesthetic sense/joint position sense. An additional test that has been recommended for cervicogenic headache, but not for migraine, is the forward head posture (FHP) [12-14]. Other tests that have been suggested for cervicogenic as well as other headache types include the temporomandibular joint assessment and neurodynamic testing [15]. To use all tests during an initial physiotherapy examination is infeasible; hence, the therapist has to choose the appropriate tests based on information from the patient interview and the clinical reasoning process. It is likely to therapists often rely on favored techniques; therefore important information might be missed using this unguided approach. Recently, an international consensus was published that identified the clinically most useful physical examination tests. Eleven headache assessment tests (HATs) were identified, and agreed upon by the experts on as the most useful to detect potentially treatable musculoskeletal dysfunctions in patients with headaches. These HATs were: FHP, cervical AROM, FRT, TP palpation, and muscle strength tests of the shoulder girdle, CCFT, PAIVMs, reproduction and resolution of headaches symptoms, PPIVMs, screening of the thoracic spine and combined cervical movement tests (upper cervical quadrant). However, these tests are only the minimum standard for the physical examination and additional tests might be required in specific clinical situations such as facial pain, dizziness or neuropathic pain [16]. While this expert opinion serves as initial guidance for the physical examination of patients with headache, HATs need to be interpreted with caution, until test properties have been evaluated for all included tests. Recently, Rubio-Ochoa et al. provided an overview of all published clinometric properties of physical examination tests for cervicogenic headache [17]. The inter-rater reliability was evaluated for PAIVMs [kappa 0.68 and PABAK 0.74 (prevalence-adjusted and bias-adjusted)], cervical AROM (kappa 0.08 to 0.89) and for the FRT (PABAK 0.67 to 0.85) in cervicogenic headache populations. For other HATs, inter-rater reliability has only been evaluated in non-headache populations, e.g. TP palpation and FHP were evaluated in

***Corresponding author:** Kerstin Luedtke, Department of Systems Neuroscience, University Hospital Hamburg-Eppendorf, Martinistr 52, 20246 Hamburg, Germany, Tel: 0049 40 741058943; E-mail: k.luedtke@uke.de

Received: February 01, 2018 **Accepted:** February 19, 2018 **Published:** February 24, 2018

patients suffering from neck or shoulder pain (ICC 0.62 and 0.83, respectively) or in healthy subjects (FHP: ICC 0.75), CCFT in persons with and without neck pain (ICC 0.69) and PPIVMS in patients with low back pain (weighted kappa 0.11 to 0.32) and neck pain (kappa 0.28-0.43 or kappa 0.09-0.63) [18-24]. The results for inter-rater reliability of combined cervical movement (upper cervical quadrant) varied from kappa 0.15 (movement to the right) to 0.61 (movement to the left) in patients with neck pain [24]. Reliability of thoracic mobility assessment was described as poor in patients with shoulder pain [25]. No reports for the inter-rater reliability of shoulder girdle muscle strength testing and reproduction and resolution could be identified. Furthermore, different rating methods are available for the different tests: e.g. active range of movement is commonly measured using degrees of movement using a cervical range of motion device (CROM) (absolute value in degrees). Manual joint palpation on the other hand is usually reported on a binary scale (joint dysfunction / no joint dysfunction). In a reliability study for the flexion-rotation test, Hall et al. measured degrees of rotation using a goniometer and added a binary test evaluation (test result is positive or negative) [26]. Interestingly the inter-rater agreement for inexperienced examiners was higher for the binary test than for the absolute values of the goniometer measurements. In that study, the decision whether a test result is positive, was based on the estimated range of motion (10° difference from the normal value of 44° rotation). Such numerical cut-off values are not available for all tests due to the nature of some tests (i.e. manual joint palpation, passive physiological intervertebral movements). Hence, the decision whether a test is considered positive or negative for these tests relies on the therapists' perception of altered joint mobility or on the patient's response to a test (i.e. pain, muscle contraction). We therefore propose a method to further refine such clinical test results on a visual analog scale (VAS) from 0 to 100 to indicate how clinically relevant the test result was. For the purpose of HATs evaluation, 0 would indicate a negative test result (no relevant clinical finding) and 100 would indicate an extremely relevant clinical finding. The advantage of this type of measurement is, that all tests can be evaluated on the same scale while offering more response options than a binary assessment method. The rating of the clinical relevance of a test result closely reflects a clinician's decision making during the physical examination procedure. In order to decide on a potentially effective treatment strategy, the evaluation of tests is crucial. Clinically, tests are not always measured in degrees or centimetres, but e.g. range of motion is estimated ("eyeballed") differences in sides and directions are compared and symptom responses are noted. Clinical decisions will therefore often be based on the overall impression of a test and not necessarily on a numerical cut-off value. This procedure is reflected by the 0-100 VAS for clinical relevance. Furthermore, the VAS is considered a linear scale providing all the statistical advantages of interval level scaled data [27,28]. The purpose of this study was therefore, to evaluate the inter-tester reliability of key tests for all components of the HATs identified in the international consensus study and to compare two different rating options regarding their level of inter-rater reliability.

Materials and Methods

The study presented in this article is reported in line with and based on the background of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [29]. Following the informed consent procedure, participants with and without headache were examined by two physiotherapists blinded towards the diagnosis and using clinically feasible key tests to represent the HATs included in the international consensus study. A third researcher (to ensure blinding of the physiotherapists conducting the physical examination) arranged the appointments and instructed all participants prior to the examination not to reveal their status as a patient or control participant. This was important to allow for an unbiased examination of the patient. The second examination was conducted 30 minutes after the first examination. The order of the two examiners was randomised using a computer generated randomisation list. Following this procedure, the examiners were blinded towards the diagnosis and towards the other examiner's test results. The order of the clinical tests was standardized as follows:

FHP **2.** Cervical AROM **3**. Upper cervical quadrant **4.** PPIVM's **5.** FRT **6.** TP palpation **7.** CCFT **8.** Muscle test of the upper trapezius muscle **9.** PAIVM's, **10.** Reproduction and resolution **11.** Screening of the thoracic spine.

Eligibility criteria

Patients were recruited at a specialised university headache clinic and diagnosed by experienced neurologists according to the current IHS criteria [30]. Included were consecutive patients diagnosed with either episodic or chronic migraine (with or without aura) or TTH, with and without symptom contributions from the neck. Patients were not included if they additionally suffered from a diagnosed cervical spine pathology, rheumatoid, neurological or psychiatric disease. Other exclusion criteria were cervical spine surgery or whiplash associated disorder in the past 2 years. Headache-free participants were recruited by online advertising, age and gender-matched, and included if they had less than three headache episodes per year that did not fulfill the criteria for migraine or TTH or any other primary headache type. The same exclusion criteria as for the headache patients, applied to the headache-free participants.

Sample size

For categorical data, many chance-corrected reliability coefficients are based on the percent agreement; therefore we followed the recommendations of Gwet and determined the sample size on the percent agreement. The optimal sample size represents the number of subjects that minimizes the standard error associated with the percent agreement between two raters. Because we decided that it is sufficient that the estimated percent agreement falls within 15% of its true error-free value, we calculated the required sample size to be 44 subjects [31]. For interval scaled test results we used the ICC. For two repeated measurements and an expected ICC of 0.8 with ± 0.1 95% CI DeVet proposed a minimum of 50 subjects [32].

Examiners

Two experienced (> 7 years postgraduate) manual therapists with similar postgraduate education background (Orthopaedic Manual Therapy according to the standard of the International Federation of Orthopaedic Manipulative Physical Therapists) and expertise in the diagnosis and treatment of patients with headache, conducted the physical examinations. Both examiners had been involved in the development of the HATs protocol and the standardisation of procedures, such as the order of tests, the number of repetitions for each test and exact test procedure.

Headache assessment tests

All tests identified in the international consensus study as clinically useful were included [16]. However, while some physical examination

tests have been described in detail and can be applied in a standardised manner based on the available literature (e.g. FRT), other tests are less well described and had to be standardised for the purpose of this study (procedures are detailed below) [26]. Care was taken to achieve a test procedure that can be feasibly conducted during a standard 30-minutes physiotherapy appointment. To ensure a high level of reproducibility of our data, all test procedures are detailed below. References provided indicate publications that describe the same test procedures in more detail. For each test, absolute measurement values were recorded, the clinical decision whether this test was considered positive by the examiner and a 0-100 VAS estimation of how clinically relevant the test result was, were documented. Clinical indicators for the decision of a positive or negative test result are presented in Table 1. This binary decision was used to calculate the best cut-off values for each test. The order of the tests was standardised. Test results were documented in a standardised protocol Figure 1.

Procedures for each HAT

FHP: CROM device with forward head arm (Cervical Range of Motion Instrument, Performance Attainment Associates, 1988, University of Minnesota). Ventral translation was measured in centimetres using C7 as a reference point and maintaining neutral flexion/extension position. The seating posture was standardised by returning to a relaxed position after 2 repetitions of maximum upright and maximum slouched position [19].

AROM (CROM device): Mean value of 3 repetitions: Flexion, extension sidebending right and left, rotation right and left Measurements were recorded in degrees of movement [33].

Combined movements (Upper Cx Quadrant): Upper cervical extension + rotation and ipsilateral lateral flexion [34]. This test was chosen as the key test to represent the category combined cervical movements. Test results were documented as hypomobility (yes=1; no=0) and / or pain provocation (yes=1; no=0).

PPIVMs: Flexion, extension, lateral flexion right and left, rotation right and left [34]. Test results were documented as hypomobility or hypermobility or normal (yes=1; no=0).

FRT: Degrees of painfree cervical rotation in maximum cervical flexion, additionally the provocation of symptoms was recorded as yes=1 and no=0 [35].

Trigger point palpation: Four TPs sternocleidomastoid muscle sternal head, 4 TPs masseter muscle, 3 TPs temporal muscle, 2 TPs sub occipital muscles, 4 TPs trapezius muscle [36]. The number of active and latent trigger points was recorded.

Muscle tests shoulder girdle: The upper trapezius muscle was chosen as the key muscle to represent the upper crossed syndrome [37]. Active hold against maximum resistance without substitution of the rhomboids or latissimus dorsi was considered normal and given a value of 2, active hold was documented as a value of 1, no active hold was given a value of 0. The patient was positioned in prone, the scapula positioned in neutral (without excessive elevation and rotation) and the arm placed in 180° elevation [38].

Table 1: Clinical indicators for a positive test results.

HAT	Test regarded positive if one of the listed findings was present:
FHP	• Obvious ventral translation of the head • Symptom change on posture correction [19]
AROM	• Marked movement restriction • Marked side difference • Symptom provocation [39]
Upper Cx Quadrant	• Marked movement restriction • Marked side difference • Symptom provocation [34]
PPIVMs	• Marked movement restriction • Marked side difference [34]
FRT	• Marked movement restriction (>10°) • Marked side difference • Symptom provocation [35]
Trigger point palpation	• Palpable taught band • Referred pain • Muscle twitch • Active top: symptom provocation [36]
Muscle tests shoulder girdle	• Marked side difference • Cannot hold against maximum resistance [38]

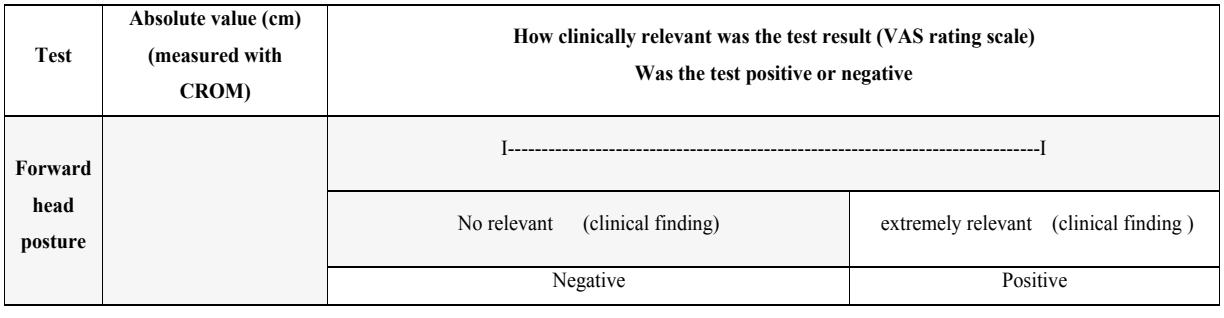

Figure 1: Documentation protocol (example: forward head posture).

PAIVMs: Central and unilateral posterior-anterior movement at C0-3 [34]. The number of hypomobile and/or painful joints and movement directions was recorded.

Reproduction and resolution: Sustained unilateral posterior-anterior movement at C0-3 [40]. If reproduction and resolution of headache symptoms was achieved this was given a value of 1.

Thoracic mobility: Active rotation right and left, extension; central and unilateral posterior-anterior manual joint palpation [34]. Hypomobility or pain provocation was recorded as a clinical sign. The number of clinical signs (max. 6: extension, rotation right/left, central PA, unilateral PA right/left) was recorded.

CCFT: Pressure Biofeedback Unit 20-30 mmHG; 10 second hold without substitution of superficial muscles [41]. The mm HG value that was held for 10 seconds without substitution was recorded (Table 1 and Figure 1).

Statistical analyses: Reliability levels in the presence or absence of a diagnosis might differ, therefore reliability levels were calculated for each group (headache and control) separately. Inter-rater reliability for the two examiners was evaluated using the intraclass correlation coefficient ($ICC_{agreement,\ single\ measure}$ 2.1) and Bland-Altman plots for interval data. For interval data we also calculated the standard error of measurement ($SEM_{agreement} = \sqrt{(\sigma^2 error)}$ and the smallest detectable change ($SDC = 1.96 \times \sqrt{6} \times SEM_{agreement}$). The interpretation of the ICC according to the Scientific Advisory Committee of the Medical Outcome Trust, was that an ICC>0,7 was considered sufficient inter-rater agreement [42]. For all nominal variables the inter-rater agreement is presented in terms of percent agreement, Cohen's kappa and Gwet's AC_1 with 95 % confidence intervals (CI). For ordinal variables we used the quadratic weighted kappa and Gwet´s AC_2. Cases with missing values were excluded.

The kappa statistic is influenced by prevalence and rater bias [43]. In situations where a large proportion of the ratings are either positive or negative, the unbalanced prevalence of the trait will lead to a reduced kappa coefficient. In situations where there is a systematic difference between the two rater's tendencies to make particular ratings, the kappa coefficient may be inflated. Therefore, beside Cohen's kappa, we also reported Gwet´s AC and percent agreement, as well as the prevalence of a positive test result. Gwet's AC_1 and AC_2 (a quadratic weighted version of Gwet´s AC), is not affected by trait prevalence or rater bias. Prevalence was calculated based on the number of positive cases, as judged by both raters, then calculated as a percentage of the total number of cases, and inter-rater reliability. Descriptive data were analysed using the paired t-test for interval data and cross-tabulations for ordinal and binary data. Furthermore, we estimated the correlation between the absolute measurement and the clinical decision on the VAS scale, using Pearson´s r, point biserial correlation or sommer´s d depending on the scale of the absolute data. To get an impression how good the clinical VAS rating represents a positive/negative test result, we performed a ROC analyses (nonparametric receiver operating characteristic analyses) using the binary decision as the reference variable and the VAS rating as the classifier to calculate the AUC (area under the curve) with its 95% CI. Within the ROC analyses "cut-of points" were estimated using the smallest sum of 1-sensitivity and 1-specificity which is the "best threshold" to dichotomize the interval scaled VAS test result [44]. All data were analysed with STATA® version 13.1.

Results

Thirty-six (72%) of the 50 examined participants were female. The average age of the study population was 41.1 years (SD=14.3), ranging from 21 to 79 years. The description of the study population is detailed in Table 2. In general, kappa values in the headache group where higher than in the asymptomatic group Figure 3. This phenomenon is caused by the low prevalence rates of a positive test result in the asymptomatic group. In Figure 3 the effect of the prevalence on percent agreement, kappa and Gwet´s AC is visualized. Based on the 95%CI of the reliability coefficients the statistical significance on their difference can be estimated [45]. For the absolute measures, the reliability values for the asymptomatic group, the headache group and both groups pooled are given in Tables 3a and 3b. The Bland-Altman plots for the pooled groups are presented in Figure 4.

Tables 3a, 3b and Figures 3 and 4. For the pooled groups, inter-rater reliability for movement tests utilizing mechanical devices (FHP, AROM) ranged from ICC 0.63 to 0.84 (Table 3a), while reliability values for manual movement testing with PPIVMs, PAIVMs and thoracic mobility ranged from ICC 0.16 – 0.37 (Table 3a). Testing for upper cervical dysfunction with the upper quadrant test showed kappa values of 0.55 (Gwet´s AC 0.73). Manual muscle testing showed kappa values of 0.48 (Gwet´s AC 0.88). Reliability of the test for deep neck flexor muscle recruitment (CCFT) showed an ICC of 0.51 (Table 3a). The reliability for tests with symptom reproduction (PAIVM (pain), upper Cx pain and reproduction and resolution) was higher than for those that observed movement limitations. Also, the correlation between the absolute measurement values and the clinically based VAS rating was better for symptoms, e.g. pain during FRT than for biomechanical aspects e.g. ROM during FRT. The inter-rater agreement for the clinical judgement on a 0-100 scale is sufficient (pooled groups; ICC>0.7) for active ROM, the upper cervical quadrant and the trigger point palpation. For most tests, the inter-rater reliability is higher in the headache group compared to the non-headache group. The mathematically best cut-off point was always in the region of about 20 mm on the VAS scale with values for sensitivity and specifity near 1. This is visualised in Figure 2. The near to perfect AUC of the ROC analyses further indicates that the 0-100 scale almost perfectly reflected the clinical judgement of a positive/negative test result.

Table 2: Descriptive statistics Notes: N/A= not applicable; SD=standard deviation; TTH=tension-type headache.

Variable		Control N=25 (%)	Migraine N=19 (%)	TTH N=6 (%)
Gender	Male	9 (36)	1 (5.3)	4 (66.7)
	Female	16 (64)	18 (94.7)	2 (33.3)
Age	Mean (sd)	40 (13.5)	39 (11.9)	54 (20.3)
Headache history	Years mean (sd)	N/a	14 (12.6)	19 (10.6)
Headache frequency	Days per month (median)	N/a	10	28

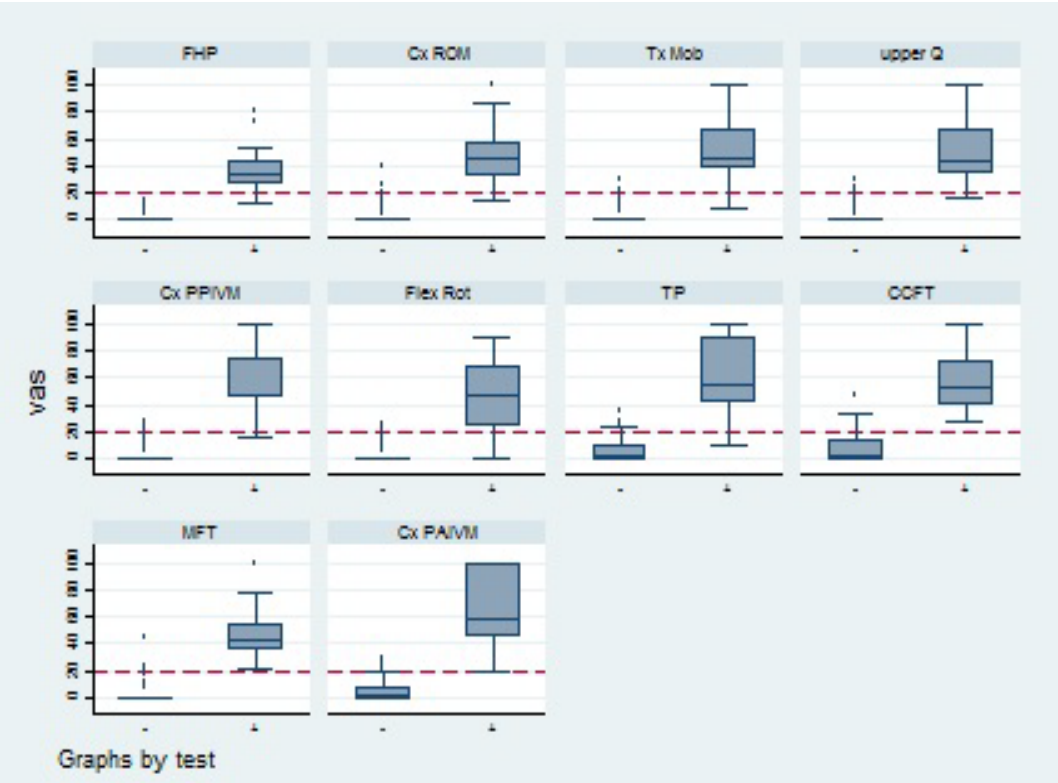

Figure 2: Box plots of the congruence of binary rating (positive/negative test result) and its relationship to the 0-100 VAS rating scale. The cut-off value to dichotomise the VAS scale (dotted line) was around 20mm on the VAS scale for all tests. x-axis = positive / negative rating of the test result; y-axis= VAS rating (0 to 100) of the clinically relevance of the test result. Values are presented in Table 5.

Discussion

The highest inter-rater reliability for absolute measurement values was shown for AROM and FHP. Both tests used measurement devices designed for the purpose of reliability. The inter-rater agreement for such measurements reflects previously published data for AROM testing of the cervical spine in a population with neck disorders and published inter-rater reliability results for the FHP in healthy participants and patients with both studies using the CROM device [19,46,47]. On the other hand, PAIVMs and PPIVMs showed low levels of inter-rater reliability of their absolute values. A systematic review on the reliability of spinal motion palpation indicated that out of 44 studies investigating either excursion or end-feel, eight showed high levels of reliability (four if only studies of high methodological quality were considered), hence our results for joint palpation are in line with most of the published data [48]. The entire protocol was conducted within a 30 minute period, indicating its feasibility for a regular physiotherapy setting. In the current study, the 0-100 rating for clinical relevance included the patient's verbal and non-verbal response and therefore reflected a clinical construct, while the absolute measurements were more biased towards biomechanical function. Jull et al. published a study in 1994 evaluating the inter-rater reliability for manual joint examination and reported that although examiners did not have to rely on the patient's response to identify a dysfunctional joint, results are best if all information, manual palpation as well as the patient's pain response, are included [49]. A patient is more likely to have clearly identifiable musculoskeletal dysfunctions than a "control" participant. Especially if symptoms are reproduced, the decision that a test is rated as more relevant is comparably easier than in situations of e.g. some stiffness in the absence of symptoms. When looking at interrater reliability in the subgroup of participants with headache, results were acceptable for 7 out of 11 tests for the 0-100 VAS scale for clinical relevance.

The rating method that best reflected the therapists' clinical decision, the 0-100 VAS scale for clinical relevance, was closely related to the decision of a positive or negtive test as demonstrated in the ROC analysis. A test result can be judged as negative on a binary scale but still receive a (low) rating on the 0-100 VAS for clinical relevance. An example to clarify this seemingly contradictory situation would be some slight stiffness of the thoracic spine without any symptom provocation. Most therapists would rate this as a negative test result. However, thoracic stiffness might still be relevant in the clinical context, e.g. thoracic mobilisation might be used to improve posture or to influence the sympathetic nervous system. Interestingly, the cut-off value to dichotomise the VAS rating for all tests was around 20 on the 0-100 VAS scale, indicating that if a binary decision was required (positive / negative test), a value above or below 20 on the VAS would represent a positive or a negative test result, respectively. We strongly recommend the use of the 0-100 scale in addition to the absolute measurement values as it provides more information than a binary scale (positive/negative test result) and closely reflects the clinical reasoning process supporting the decision which treatment approach to use. It is particularly useful in tests where the nature of the test (i.e. PPIVMs) does not allow a clear definition of cut-off values to distinguish between positive and negative test results. Furthermore, the 0-100 VAS is suitable for all statistical analyses that require interval data. Generally, the physical examination tests are

Table 3a: Inter-rater agreement and correlation to the clinical (VAS) rating for test results with interval scaled data.

		Asymp group	Headache group	Groups pooled					
		(N=25)	(n=25)	(n=50)					
Test		Inter rater agreement	Inter rater agreement	Descriptive statistics		Inter rater agreement			Corr. To VAS rating
				Rater 1	Rater 2				
	Measure	$ICC_{2,1}$ (95% CI)	$ICC_{2,1}$ (95% CI)	mean (sd)	mean (sd)	$ICC_{2,1}$ (95%CI)	SEM[1]	SDC[2]	Pearson r
FHP	Ventral translation (cm)	0.67*** (0.37-0.84)	0.6** (0.06-0.86)	19.04 1.7	19.35 2	0.63*** (0.40- 0.79)	1.1 cm	3.05 cm	0.06 †
AROM	Range (°) Flexion	0.72*** (0.46-0.86)	0.82*** (0.64-0.92)	56.7 (11.8)	57.4 (14.5)	0.79*** (0.66-0.88)	6°	16.6°	-0.39***
	Extension	0.72*** (0.42-0.87)	0.83*** (0.65-0.92)	70.7 (16.5)	67.8 (14.1)	0.80*** (0.67-0.88	7°	19.4°	-0.48***
	Rotation right	0.46** (0.09-.071)	0.75*** (0.51-0.88)	72.3 (15)	68.8 (13.3)	0.67*** (0.47-0.80)	8°	22.2°	-0.39***
	Rotation left	0.54** (0.19-0.77)	0.78*** 0.56-0.89)	68.9 (14.2)	67.4 (13.5)	0.71*** (0.55-0.83)	7°	19.4°	-0.43***
	Latflex right	0.75*** (0.52-0.88)	0.81*** (0.6-0.9)	43.6 (12.1)	41.8 (11.1)	0.80*** (0.67-0.88)	5°	13.9°	-0.29**
	Latflex left	0.82*** (0.64-0.92)	0.81*** (0.61-0.91)	45.1 (12.4)	45.1 (12.1)	0.84*** (0.73-0.90)	5°	13.9°	-0.29**
FRT (ROM)	Rotation (°)	0.35** (0.1-0.57)	0.47*** (0.24-0.66)	56.8 (10.3)	53.5 (10)	0.47*** (0.29-0.61)	7.5°	20.7°	0.394***
TP palpation	Number of active	0.14† (-0.2-0.48)	0.46** (0.08-0.71)	2.8 (5.7)	1 (2.6)	0.49*** (0.23- 0.68)	3	8.3	0.63***
	latent	0.67*** (0.37-0.86)	0.59** (0.28-0.79)	8.4 (8.1)	6.1 (7)	0.65*** (0.44,-0.79)	5	13.9	0.66***
CCFT	Pressure reached (mmHG)	0.47** (0.09-0.72)	0.49** (0.14-0.74)	25.6 (2.5)	26.5 (2.2)	0.51*** (0.26-0.69)	1.7	4.7	0.834***
PPIVMs Co-C2	Number of findings	0.63*** (0.2-0.83)	0.22† (-0.16-0.55)	1.5 (2.1)	1.4 (2.8)	0.37*** (0.1 -0.59)	1.4	3.9	0.69***
PAIVMs Co-C3	Number of dysfunctions (max 6)	0.12† (-0.23-0.46)	0.03† (-0.35-0.4)	1.6 (1.2)	1.6 (1.2)	0.16† (0 - 0.43)	1.4	3.9	0.54***
	Number of pain responses (max 6)	0.31* (-0.04-0.61)	0.44** (0.08-0.7)	1.4 (1.7)	2.1 (1.9)	0.58*** (0.34-0.74)	1.2	3.3	0.81***
Thoracic mobility	Number of + tests (max 6)	0.012† (-0.35-0.38)	0.34* (-0.04-0.64)	1.8 (1.6)	1.2 (1.7)	0.29* (0.03- 0.52)	1.4	3.9	0.43***

Note: Descriptive statistics: means standard deviations appear in parenthesis underneath means.

Inter rater agreement: ICC Typ $2,1_{agreement}$, with their 95% confidence intervals; 1. $SEM_{agreement}$ =Standard error of the Measurement=$\sqrt{(\sigma^2 error)}$ which includes the systematic error [1]. 2. Smallest detectable change =1.96 x $\sqrt{2}$ x $SEM_{agreement}$. Correlation to VAS rating: All values are pearson r. positive or negative values represents the direction of correlation **Stat. significance**: *=p≤0.05; **= p≤0.01; ***= p≤0.001; †=p not sig. N for the correlations=100, FRT (rom) left and right side pooled

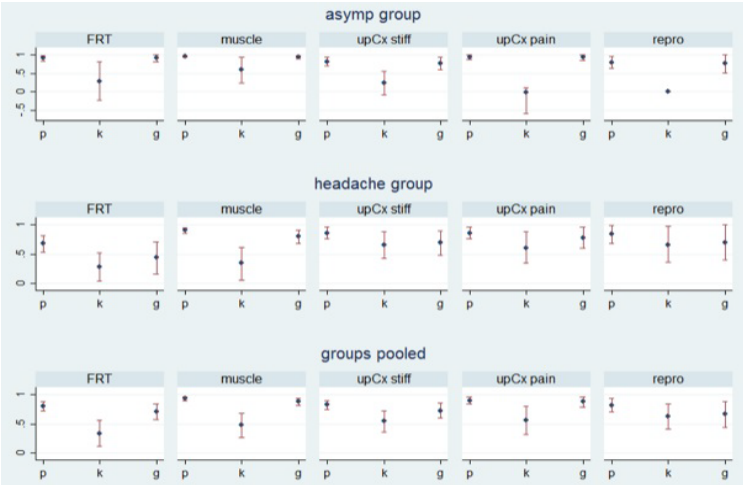

Figure 3: Reliability coefficients for categorical data for different groups: p=percent agreement, k=Cohen´s kappa, g=Gwet´s (AC). The values are presented in Table 4b.

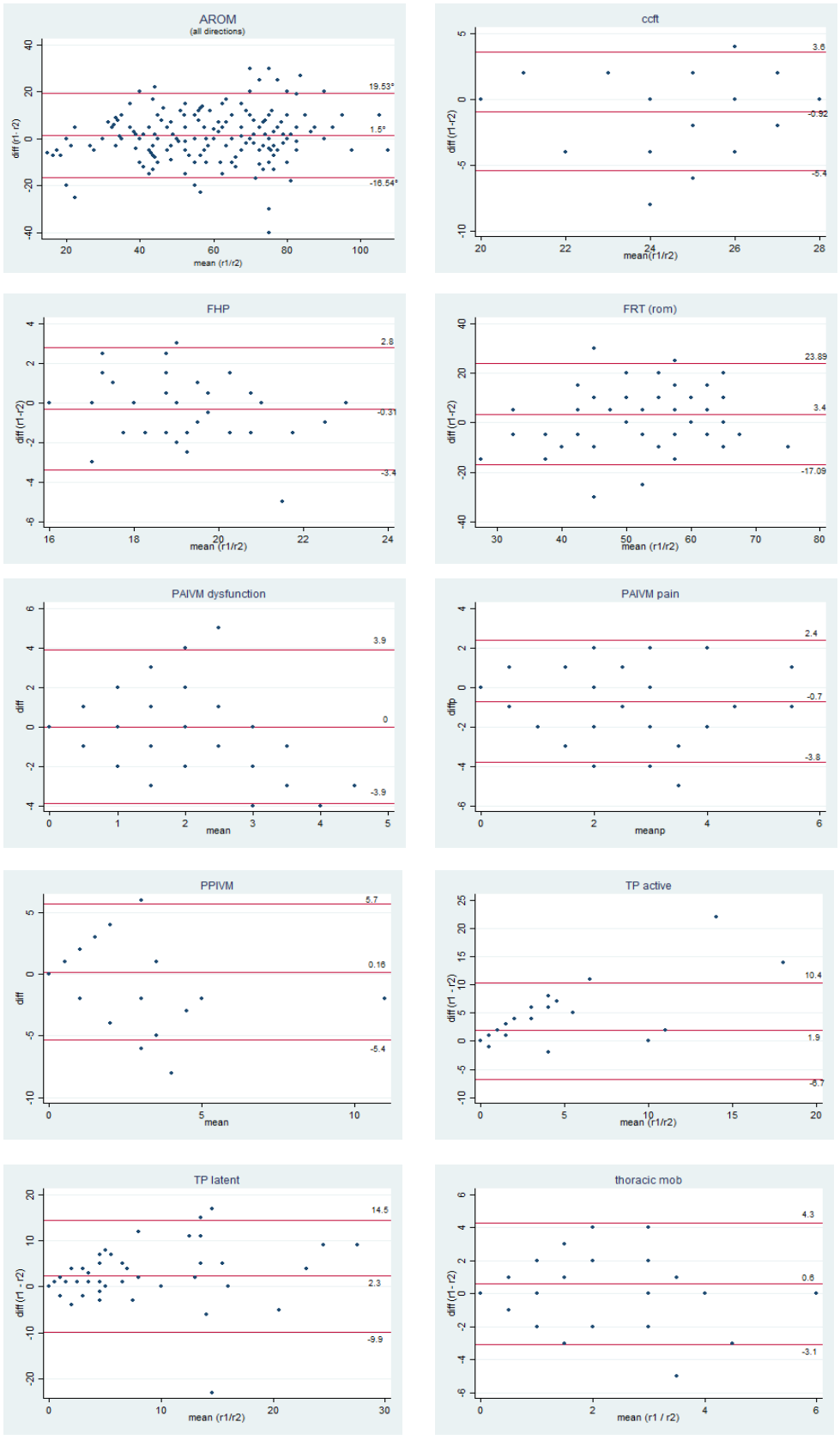

Figure 4: Bland Altmann plots for each test; x-axis= mean rater1/rater2; y-axis: rater1 – rater2; reference lines = bias and 95% limits of agreement.

Table 3b: Inter-rater agreement and correlation to the clinical (VAS) rating for test results with interval scaled data.

			Asymp group (N=25)	Headache group (N=25)	Groups pooled (N=50)	Correlation to VAS rating
Test	**Measure**	**Reliability coef**	Coef (95% CI)	Coef (95% CI)	Coef (95% CI)	
Flexion-rotation test	Pain (no/yes)	Percent agreement	0.92 (0.84-0.99)	0.68 (0.54-0.81)	0.8 (0.72-0.88)	0.63***
		Cohen‘ s Kappa	0.29 (-0.23-0.81)	0.28 (0.04-0.53)	0.34 (0.12-0.56)	
		Gwet´s AC1	0.91 (0.81-1)	0.44 (0.17-0.71)	0.71 (0.58-0.84)	
		Prevalence rate (%)	2%	14%	8%	
Muscle strength	Weak/Moderate/Strong	Percent agreement	0.96 (0.93-0.98)	0.91 (0.86-0.95)	0.93 (0.90-0.96)	[a]0.84***
		Cohen's Kappa (w2)	0.59 (0.23-0.95)	0.35 (0.06-0.62)	0.48 (0.27-0.69)	
		Gwet´s AC2	0.94 (0.90-0.98)	0.8 (0.69-0.91)	0.88 (0.82-0.94)	
		Prevalence rate (%)	8%	28%	18%	
Upper Cx quadrant	Stiffness (no/yes)	Percent agreement	0.82 (0.71-0.93)	0.84 (0.74-0.94)	0.83 (0.75-0.90)	0.61***
		Cohen’ s Kappa	0.24 (-0.09-0.56)	0.66 (0.43-0.88)	0.55 (0.37-0.73)	
		Gwet´s AC1	0.77 (0.60-0.93)	0.7 (0.49-0.90)	0.73 (0.60-0.86)	
		Prevalence rate (%)	4%	28%	16%	
	Pain (no/yes)	Percent agreement	0.94 (0.87-1)	0.86 (0.76-0.96)	0.9 (0.84-0.96)	0.71***
		Cohen´s Kappa	-0.027 (-0.06-0.012)	0.61 (0.35-0.88)	0.56 (0.32-0.80)	
		Gwet´s AC1	0.94 (0.86-1)	0.78 (0.61-0.96)	0.88 (0.79-0.96)	
		Prevalence rate (%)	0%	16%	8%	
Reproduction & resolution	Referral to head (no/yes)	Percent agreement	0.8 (0.64-0.97)	0.84 (0.69-0.99)	0.82 (0.71-0.93)	0.73**
		Cohen’s Kappa	0 (0-0)	0.66 (0.36-0.97)	0.63 (0.42-0.84)	
		Gwet´sAC1	0.76 (0.52-1)	0.7 (0.4-1)	0.67 (0.45-0.88)	
		Prevalence rate (%)	0%	56%	28%	

Note: *Prevalence was calculated based on the number of positive cases, as judged by both raters, then calculated as a percentage of the total number of cases, and inter-rater reliability. For example, when calculating the prevalence of FRT in the headache group, the number of cases in which raters agreed with each other was 7, which was calculated as a percentage of the total number of ratings (50 (N= 25 to each side)), leading to a prevalence rate of 14%.

Correlation: corr between the groups pooled to the VAS rating point biserial correlation; a = somer´s d (hazard ratio), **Stat. significance**: *=p≤0.05; **= p≤0.01; ***= p≤0.001; †=p not sig. N for the correlations=100, for all other tests=50

not conducted to discriminate between patients with headache and patients without headache because the diagnosis is largely made by patient interview. Physical examination tests are rather used to identify musculoskeletal dysfunctions potentially treatable by physiotherapy, to decide on a potentially relieving treatment technique, and to evaluate progress over time. The proposed cluster of tests is a feasible tool for this purpose with an acceptable level of inter-rater reliability, especially when patient responses are considered in the evaluation process.

Limitations

For the purpose of presenting a selection of tests that is “clinician friendly”, we only included one test for each component of the international consensus on HATs. (e.g. Only the trapezius pars ascendens was tested to represent the component “shoulder girdle strength”). If this test or any signs and symptoms from the clinical picture point towards altered shoulder girdle muscle strength it is recommended to focus on this component and additionally test other relevant muscles. The same applies to any of the other HATs components. It cannot be excluded that the order of the tests and the number of tests conducted on one day influenced the results. We therefore standardised the order, so that this effect would be the same in each participant. Tests were conducted on the same day in order not to induce more variability due to two test days.

Conclusion

Overall, most HATs showed acceptable levels of interrater agreement. Interrater reliability improved when using a clinically based rating scale that included the patient’s response. The 0-100 VAS measuring clinical relevance of a test result closely reflected the decision whether a test was positive at a cut-off value of 20 on the 0-100 VAS.

Funding

This work was supported by the Deutsche Migräne- und Kopfschmerz-Gesellschaft (DMKG) and the International Maitland Teachers Association (IMTA).

References

1. Steiner TJ, Stovner LJ, Katsarava Z, Lainezet JM, Lampl C, et al (2014) The impact of headache in Europe: principal results of the Eurolight project. J Headache Pain 15: 31.
2. Vos T, Flaxman AD, Naghavi M, Lozano R , Michaud C , et al (2012) Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380: 2163-2196.
3. D'Amico D, Grazzi L, Usai S, M. Leonardi, A. Raggi (2013) Disability and quality of life in headache: where we are now and where we are heading. Neurol Sci 34:1-5.
4. Raggi A, Leonardi M, Bussone G, D'Amico D (2013) A 3-month analysis of disability, quality of life, and disease course in patients with migraine. Headache 53: 297-309.
5. Berg J, Stovner LJ (2005) Cost of migraine and other headaches in Europe. Eur J Neurol 12: 59-62.
6. Bloudek LM, Stokes M, Buse DC, Wilcoxet TK, Lipton RB et al (2012) Cost of healthcare for patients with migraine in five European countries: results from the International Burden of Migraine Study (IBMS). J Headache Pain 13: 361-378.
7. Burton WN, Conti DJ, Chen CY, Schultz AB, Edington DW, et al (2002) The economic burden of lost productivity due to migraine headache: a specific worksite analysis. J Occup Environ Med 44: 523-529.
8. Lanteri-Minet M (2014) Economic burden and costs of chronic migraine. Curr Pain Headache Rep 18: 385.
9. Pradalier A, Auray JP, El Hasnaoui EA, Alzahouri K, Dartigueset JF et al (2004) Economic impact of migraine and other episodic headaches in France. Pharmacoeconomics 22: 985-999.
10. Hall T, Briffa K, Hopper D (2008) Clinical Evaluation of Cervicogenic Headache: A Clinical Perspective. J Man Manip Ther 16: 73-80.
11. Howard PD, Behrns W, Martino MD, Mambro AD, McIntyre K, et al (2015) Manual examination in the diagnosis of cervicogenic headache: a systematic literature review. J Man Manip Ther 23: 210-218.
12. Fernández-de-las-Peñas C, Alonso-Blanco C, Cuadrado ML, Pareja JA (2006) Forward head posture and neck mobility in chronic tension-type headache: a blinded, controlled study. Cephalalgia Int J Headache 26: 314-319.
13. Fernández-de-Las-Peñas C, Cuadrado ML, Pareja JA (2006) Myofascial trigger points, neck mobility and forward head posture in unilateral migraine. Cephalalgia Int J Headache 26:1061-1070.
14. Fernández-de-Las-Peñas C, Cuadrado ML, Pareja JA (2007) Myofascial trigger points, neck mobility, and forward head posture in episodic tension-type headache. Headache 47: 662-672.
15. von Piekartz HJM, Schouten S, Aufdemkampe G (2007) Neurodynamic responses in children with migraine or cervicogenic headache versus a control group. A comparative study. Man Ther 12: 153-160.
16. Luedtke K, Boissonnault W, Caspersen N, Castien R, Chaibi A, et al (2016) International consensus on the most useful physical examination tests used by physiotherapists for patients with headache: A Delphi study. Man Ther 23:17-24.
17. Rubio-Ochoa J, Benítez-Martínez J, Lluch E, Santacruz-Zaragozá S, Gómez-Contreras P, et al (2016) Physical examination tests for screening and diagnosis of cervicogenic headache: A systematic review. Man Ther 21: 35-40.
18. Barbero M, Bertoli P, Cescon C, Macmillan F, Coutts F et al (2012) Intra-rater reliability of an experienced physiotherapist in locating myofascial trigger points in upper trapezius muscle. J Man Manip Ther 20:171-177.
19. Garrett TR, Youdas JW, Madson TJ (1993) Reliability of measuring forward head posture in a clinical setting. J Orthop Sports Phys Ther 17:155-160.
20. Nam SH, Son SM, Kwon JW, Lee NK, et al (2013) The Intra- and Inter-rater Reliabilities of the Forward Head Posture Assessment of Normal Healthy Subjects. J Phys Ther Sci 25:737-739.
21. Juul T, Langberg H, Enoch F, Søgaard K (2013) The intra- and inter-rater reliability of five clinical muscle performance tests in patients with and without neck pain. BMC Musculoskelet Disord 14: 339.
22. Phillips DR, Twomey LT (1996) A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. Man Ther 1: 82-87.
23. Smedmark V, Wallin M, Arvidsson I (2000) Inter-examiner reliability in assessing passive intervertebral motion of the cervical spine. Man Ther 5:97–101.
24. Pool JJ, Hoving JL, de Vet HC, van Mameren H, Bouter LM (2004) The interexaminer reproducibility of physical examination of the cervical spine. J Manipulative Physiol Ther 27: 84-90.
25. Burns SA, Cleland JA, Carpenter K, Mintken PE (2016) Interrater reliability of the cervicothoracic and shoulder physical examination in patients with a primary complaint of shoulder pain. Phys Ther Sport 18:46-55.
26. Hall TM, Robinson KW, Fujinawa O Akasaka K , Pyne EA (2008) Intertester reliability and diagnostic validity of the cervical flexion-rotation test. J Manipulative Physiol Ther 31: 293-300.
27. Hofmans J, Theuns P (2008) On the linearity of predefined and self-anchoring Visual Analogue Scales. Br J Math Stat Psychol 61: 401-413.
28. Myles PS, Troedel S, Boquest M, Reeves M (1999) The pain visual analog scale: is it linear or nonlinear? Anesth Analg 89:1517-1520.
29. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud 48: 661-671.
30. Headache Classification Committee of the International Headache Society (IHS) (2013) The International Classification of Headache Disorders, 3rd edition (beta version). Cephalalgia Int J Headache 33: 629-808.
31. Gwet KL (2010) Estimating the number of subjects and number raters when designing an inter-rater reliability study.
32. de Vet HCW, Terwee CB, Mokkink LB, Knol DL (2011) Measurement in Medicine: A practical guide. (1st ed) Cambridge University Press, New York, United State of America.
33. Fletcher JP, Bandy WD (2008) Intrarater reliability of CROM measurement of cervical spine active range of motion in persons with and without neck pain. J Orthop Sports Phys Ther 38:640-645.
34. Maitland GD (2013) Vertebral Manipulation, (4th ed) Butterworth-Heinemann, United kingdom.
35. Hall T, Robinson K (2004) The flexion-rotation test and active cervical mobility--a comparative measurement study in cervicogenic headache. Man Ther 9:197-202.
36. Travell JG, Simons DG (1998) Myofascial Pain and Dysfunction The Trigger Point Manual, (2nd ed) Lippincott Williams&Wilki, Baltimore, United States of America.
37. Moore MK (2004) Upper crossed syndrome and its relationship to cervicogenic headache. J Manipulative Physiol Ther 27: 414-420.
38. Kendall FP, McCreary EK, Provance PG, Rodgers MM, Romani WA (2014) Muscles: Testing and Function, with Posture and Pain: Testing and Function with Posture and Pain, (5th ed) Wolters Kluwer Health, United Kingdom.
39. Hole DE, Cook JM, Bolton JE (1995) Reliability and concurrent validity of two instruments for measuring cervical range of motion: effects of age and gender. Man Ther 1: 36-42.
40. Watson DH, Drummond PD (2012) Head pain referral during examination of the neck in migraine and tension-type headache. Headache 52: 1226-1235.
41. Jull GA, O'Leary SP, Falla DL (2008) Clinical assessment of the deep cervical flexor muscles: The craniocervical flexion test. J Manipulative Physiol Ther 31: 525-533.
42. Lohr KN (2002) Assessing health status and quality-of-life instruments: Attributes and review criteria. Qual Life Res 11:193-205.
43. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL (2013) A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 13: 61.
44. Fraud R, Abel G (2014) Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of Pythagoras. theoretical considerations and an example application of change in health status. PloS One 9: e114468.

45. Cumming G, Finch S (2005) Inference by eye: confidence intervals and how to read pictures of data. Am Psychol 60:170-180.

46. Youdas JW, Carey JR, Garrett TR (1991) Reliability of measurements of cervical spine range of motion-comparison of three methods. Phys Ther 71: 98-104.

47. Dunleavy K, Neil J, Tallon A, Adamo DE (2015) Reliability and validity of cervical position measurements in individuals with and without chronic neck pain. J Man Manip Ther 23: 188-196.

48. Haneline MT, Cooperstein R, Young M, Birkel K (2008) Spinal motion palpation: a comparison of studies that assessed intersegmental end feel vs excursion. J Manipulative Physiol Ther 31: 616-626.

49. Jull G, Treleaven J, Versace G (1994) Manual examination: is pain provocation a major cue for spinal dysfunction? Aust J Physiother 40: 159-165.

Author Affiliations

Top

[1]*University Hospital Hamburg-Eppendorf, Department of Systems Neuroscience, Martinistr. 52, 20246, Hamburg, Germany*

[2]*HAWK University of Applied Sciences and Arts Hildesheim / Holzminden / Göttingen, Faculty of Social Work and Health, Faculty of Occupational Therapy, Speech Therapy, Physiotherapy, Goschentor, Hildesheim, Germany*

[3]*City University of Applied Sciences Bremen, Faculty of Social Sciences, Applied Sciences Speech and Language Therapy and Physiotherapy, Neustadtswall 30, 28199, Bremen, Germany*

[4]*University of Lübeck, Department of Physiotherapy, Osterweide 10, 23562, Lübeck, Germany*