Reliability of Clinical Findings in Temporomandibular Disorders

Anton de Wijer, PT, MT
Physiotherapist, Manual Therapist
Department of Oral-Maxillofacial Surgery, Prosthodontics and Special Dental Care

Agnes M. Lobbezoo-Scholte, DDS, PhD
Department of Oral-Maxillofacial Surgery, Prosthodontics and Special Dental Care

Michel H. Steenks, DDS, PhD
Department of Oral-Maxillofacial Surgery, Prosthodontics and Special Dental Care

Frederik Bosman, PhD
Professor
Department of Oral-Maxillofacial Surgery, Prosthodontics and Special Dental Care

Utrecht University
Utrecht, Netherlands

Correspondence to:
Anton de Wijer
Department of Oral-Maxillofacial Surgery, Prosthodontics and Special Dental Care
Utrecht University
Universiteitsweg 100
3584 CG Utrecht
Netherlands

The aim of the present investigation was to study the interexaminer reliability of orthopedic tests and palpation techniques routinely used in the clinical diagnosis of disorders of the masticatory system. The tests were performed by a dentist and a physiotherapist, who both used the tests routinely when examining patients with temporomandibular disorders. Seventy-nine patients participated in this study. In the analysis, percentage agreement, intraclass correlation, and Cohen's kappa were used. The interexaminer reliability of the tests measuring maximal active mouth opening and registration of clicking during active mouth opening was high. The interexaminer reliability was fair for the tests measuring the intensity of pain during active movements and moderate for tests recording joint sounds (κ = 0.47 to 0.59). There was high interobserver agreement on several items of the traction and translation tests, although the kappa values were low. The interexaminer reliability of the multitest scores for compression was substantial for joint sounds (κ = 0.66) and fair for pain (κ = 0.40). The interexaminer reliability of the multitest scores for muscle palpation and joint palpation was moderate (κ = 0.51) and fair (κ = 0.33), respectively. It can be concluded that most variables determined during active movements can be measured with satisfactory reliability, whereas variables for other tests are not measured with the same reliability on the basis of the kappa scores. The main symptoms of temporomandibular disorders can be evaluated reliably with multitest scores. It is recommended that clinicians calibrate their techniques regularly to improve the reliability of results in daily practice.
J OROFACIAL PAIN 1995;9:181–191.

key words: temporomandibular disorders, interexaminer reliability, orthopedic tests

Temporomandibular disorders (TMD) is a collective term embracing a number of clinical problems that involve the masticatory musculature, the temporomandibular joint (TMJ), or both.[1] Many authors agree on the following main symptoms: pain in the area of the TMJ and the masticatory muscles, limited function of the jaw and/or deviations in jaw opening, and joint sounds such as clicking, popping, or crepitation.[1–5] In the past, diagnostic procedures for patients with TMD placed emphasis on active movements and palpation. Hansson et al[6] advocated using an additional diagnostic procedure, based on principles from orthopedic medicine and manual therapy, for the differential diagnosis.[7,8] In this procedure, tension is applied in different ways to individual structures from which symptoms may originate. Besides traditional tests such as active movements and palpation, the status of the joint and the muscle can be also evaluated with additional tests such as passive movements, traction, translation, and dynamic and static muscle resistance tests. Other authors have also sug-

gested the use of tests based on orthopedic principles.[4,9–11] In differential diagnosis, active movements, passive opening, and palpation are used to determine the presence or absence of TMD.[12] If these test results are positive, additional tests are necessary to classify the patient in one of the subgroup categories.[1,12]

A basic concern about these orthopedic tests is the reliability of the tests. The intraexaminer and interexaminer reliability of the functional examination of the masticatory system has been investigated, and many studies show that clinical signs and symptoms important to the differential diagnosis of TMD can be measured with varying reliability.[13–17] The vertical range of motion of both unassisted and assisted mouth opening could be measured reliably in millimeters by both calibrated and noncalibrated clinicians.[16–18] Measurements of horizontal mandibular movements were less reliable.

Most studies demonstrated a higher intraobserver than interobserver consistency, and a conclusion was that longitudinal studies or treatment evaluations should be performed by the same observer.[13,14,18] In a previous study, the reliability of six orthopedic tests applied to diagnostic subgroups of TMD was described.[19] Multitest scores (MTS) were composed for each test or combination of tests for the three main symptoms of TMD. The interexaminer reliability of the determination of the three main symptoms of TMD was satisfactory.

To gain more insight into the effect that the individual tests have on MTS results, the aim of the present study was to assess the interexaminer reliability of individual orthopedic tests used in the routine clinical examination.

Materials and Methods

Subjects and Examiners

The patients included in the study were randomly selected from patients who were referred to the Department of Temporomandibular Disorders and Orofacial Pain of the University Hospital of Utrecht, Utrecht, Netherlands. The study included 60 females (76%) aged 16 to 69 years (mean age 35.9, SD 13.4) and 19 males (24%) aged 18 to 50 years (mean age 33.2, SD 10.4).

The patients were examined by one of the four dentists and were classified as having TMD when they reported pain in the TMJs and/or in the masticatory muscles and/or limitations or deviations in the mandibular range of motion whether or not combined with joint sounds during mandibular function and related to the chief complaint. The examination consisted of history-taking, extraoral and intraoral inspection, occlusal analysis, orthopedic tests, ie, *(1)* active movements of the mandible, *(2)* passive opening, *(3)* traction and translation, *(4)* compression, *(5)* static pain (isometric resistance), *(6)* palpation of the joint and the masticatory muscles, and *(7)* radiography (panoramic view, supplemented by transcranial radiography if indicated). After patients had given their informed consent, a second examination was performed by a physical therapist who repeated the orthopedic tests of the stomatognathic system. This examination was performed within 2 hours of the first examination. The physiotherapist was not aware of the results of the first session. All patients were seen by the same physiotherapist. Two years before the study the dentists and the physiotherapist had calibrated their examination techniques. The tests had been in routine use for 2 years before the study started, and during this time there was some supervision and recalibration as necessary. No further calibration took place during the study. Since not all tests were performed on every patient, different frequencies were reported in the results.

Clinical Variables

The orthopedic tests used to examine the function of the TMJ and masticatory muscles are described elsewhere.[11] During all mandibular movements, the examiner assessed the presence or absence of signs and symptoms, such as joint sounds, intensity of clicking, and crepitation, on a 0-to-2 rating scale (none, hardly audible, clearly audible). The patient scored pain on a 0-to-3 rating scale (none, mild, moderate, severe pain). The head and neck were in an orthostatic position at the start of the tests.[20,21]

Active Movements. The ranges of movement were measured to the nearest millimeter with a ruler and classified as normal (≥ 40 mm), moderately restricted (≥ 30 mm and < 40 mm), and severely restricted (< 30 mm).[2,9,12] Mouth opening was measured interincisally and was corrected for overbite. Lateral movements were assessed by asking the patient to move the mandible in a lateral direction as far as possible, with a maximum mouth opening of 5 mm; they were classified as normal (≥ 7 mm) and abnormal (< 7 mm). In active protrusion, the distance between the mesioincisal ridge of the right maxillary and mandibular central incisors was measured and added to the overjet. The results were classified as normal (≥ 7 mm) and abnormal (< 7 mm).[2,10]

Passive Opening. The patient was first asked to open his/her mouth as wide as possible, and then gentle pressure was exerted on the maxillary and mandibular incisors until the end position was reached.[10] The end-feel was assessed and the range of motion was measured in the same way as for active movements. End-feel is the resistance perceived by the examiner at the end of the passive range-of-motion test. A normal end-feel is elastic. An abnormal end-feel can be blocking or empty (for example, end-feel could not be established because of pain).[6,9,10,22] The end-feel distance, the difference between the range of passive opening movement and the range of active opening, was calculated.[23]

Traction and Translation Tests. Traction and translation tests evaluate passive accessory motion in the joint (joint play). Joint play is the possibility to separate both joint partners in a passive way, or to be able to slide them (the joint partners) in a parallel direction.[8,10,24] Traction and translation movements were performed with the examiner's thumb placed on the occlusal surfaces of the molars; the examiner moved the condyle in a caudal direction for traction and in a ventrodorsal and mediolateral direction for translation. The palpatory finger on the TMJ registered the moving lateral pole of the joint. The patient had to be relaxed when the joint was examined. Each direction was evaluated separately, and the extent of the joint play (normal, abnormal), the end-feel, and the presence of pain (0-to-3 rating scale) were evaluated.

Compression Tests. The intra-articular joint structures were loaded by moving the mandible in a dorsocranial direction.[10,11] The fixing hand provided for counterforce in the contralateral direction.

Resistance Tests. The mandible was kept in a stationary position, and a gradually increasing force was applied to the mandible on laterotrusion left and right, during opening, and during closing.[9,10,25,26] To standardize the test, each patient was instructed to resist the force for at least 5 seconds to allow the patient to feel maximal resistance.

Palpation. The lateral pole of the TMJ was palpated with the mouth closed and open. The dorsal pole was palpated posteriorly (external auditory meatus) in an anterior direction during opening and closing. The masseter muscles (superficial and deep part), temporal muscles (anterior and posterior part), and the attachment of the medial pterygoid muscles were palpated extraorally, with a slight rolling pressure of the index or middle finger.[9] The ventral part of the superficial masseter muscle was palpated by grasping the muscle between the index finger and the thumb. Tenderness was graded as mild, moderate, or severe.

Statistics

Interexaminer agreement for dichotomous variables was expressed on a scale from 0% to 100%. Cohen's kappa was used to assess agreement beyond chance.[27-29] The strength of agreement for various ranges of kappa was as described by Landis and Koch[30]:

1. Poor, $\kappa < 0.00$
2. Slight, $0.00 \leq \kappa \leq 0.20$
3. Fair, $0.21 \leq \kappa \leq 0.40$
4. Moderate, $0.41 \leq \kappa \leq 0.60$
5. Substantial, $0.61 \leq \kappa \leq 0.80$
6. Almost perfect, $0.81 \leq \kappa < 1.00$

The intraclass correlation coefficient (ICC) was used to reveal differences in vertical mouth opening in millimeters and was derived from analysis of variance of the data.[29,31] The ICC values greater than .75 are considered acceptable.[17]

Because the diagnostic process is based on combinations of tests, and in some tests only a few signs and symptoms occurred, MTS were composed for each individual test.[32] An MTS is a combination of several related variables. Because correlated signs and symptoms can be counted only, this correlation was checked with Pearson's phi.[29] The MTS was considered positive when a sign brought on by one or more parts of an orthopedic test was present in the multitest score of that test. The reliability of the MTS was also determined with Cohen's kappa. McNemar's tests[29] were performed to analyze whether there was a systematic error due to the fixed sequence of the examiners.

Results

The interexaminer reliability for active movements is described in Table 1. The measurements for active maximal mouth opening (ICC = .8) and registration of clicking during mouth opening (κ = 0.7) were highly reliable. There was moderate agreement between the examiners for the classification of maximum mouth opening and joint sounds on active mouth opening (κ = 0.56 and κ = 0.59, respectively). The kappa value for joint sounds on active movements ranged from 0.47 to 0.59. The MTS showed fair agreement for pain (κ = 0.32) and crepitation (κ = 0.29), moderate agreement for restriction of movement (κ = 0.52) and for clicking (κ = 0.70), and substantial agreement for total of joint sounds (κ = 0.61). On passive opening there was fair and slight interexaminer agreement for pain (κ = 0.34), joint sounds (κ = 0.24), and end-feel (κ = 0.01) (Table 2).

Table 1 Interexaminer Reliability During Active Movements (n = 79)

	% A	κ	D	P
Maximal mouth opening classified < 30, 30 to 40, ≥ 40	95	0.56	8.9*	10.2*
Pain on opening	68	0.28	38.0	24.1
Joint sounds on opening	79	0.59	51.9	36.4
Pain on laterotrusion right	71	0.28	33.8	18.2
Joint sounds on laterotrusion right	83	0.57	27.3	26.0
Pain on laterotrusion left	78	0.28	22.1	15.6
Joint sounds on laterotrusion left	83	0.50	24.7	18.2
Pain on protrusion	80	0.36	18.7	20.0
Joint noises on protrusion	80	0.47	21.3	28.0
Multitest scores				
Pain	65	0.32	58.4	39.0
Clicking	86	0.70	43.4	39.5
Crepitation	77	0.29	25.7	13.5
Joint sounds (total)	80	0.61	61.8	47.4
Restriction of movement (opening < 40 and/or laterotrusion < 7 mm)	84	0.52	8.9	10.2

* Value is less than 40 mm.
% A = percentage agreement; D = frequency (%) of the signs registered by the dentist; P = frequency (%) of the signs registered by the physiotherapist.

Table 2 Interexaminer Reliability of MTS on Passive Opening

	n	% A	κ	D	P
Pain on opening	73	67	0.34	47.9	39.7
Joint sounds	73	83	0.24	15.3	9.7
Abnormal end-feel	70	59	0.01	40.0	10.0

% A = percentage agreement; D = frequency (%) of the signs registered by the dentist; P = frequency (%) of the signs registered by the physiotherapist.

For traction and translation (Table 3), more than 85% agreement was reached for several signs and symptoms. However, kappa ranged from poor to moderate for pain ($-0.08 \leq \kappa \leq 0.50$), from poor to substantial for joint sounds ($-0.02 \leq \kappa \leq 0.66$), and from poor to slight for end-feel ($-0.13 \leq \kappa \leq 0.20$). The MTS ranged from poor to moderate for signs and symptoms (interexaminer reliability score for pain, $\kappa = 0.46$).

The kappa values for compression ranged from 0.19 to 1.00 (Table 4), with a moderate score for pain on compression on the left side ($\kappa = 0.47$) and a fair score for the MTS pain ($\kappa = 0.40$). In the resistance tests, the kappa values for pain varied from 0.15 to 0.30 (Table 5). The kappa values for pain on palpation of the muscles and the joint ranged from 0.16 to 0.45 (Table 6). The MTS for pain on muscle palpation was moderate ($\kappa = 0.51$), and for the joint it was fair ($\kappa = 0.33$).

The MTS for the combination of tests for the three main symptoms of TMD, namely pain, joint sounds, and restriction of movement, are shown in Table 7. The reliability of pain assessment was best when palpation was added to the five other tests. The reliability of assessing joint sounds and restriction of movement was best when active movements were used alone.

Discussion

Methods

A basic requirement for a proper diagnosis is the reliability of the diagnostic procedures. An earlier study established that in patients with TMD, the diagnosis should not be based on the results of one test, but on a combination of findings.[19] Although

Table 3 Interexaminer Reliability for the Traction and Translation Tests (n = 67)

	% A	κ	D	P
Traction right				
Joint play*	87	–0.03	11.8	1.5
Pain	85	–0.08	8.8	5.9
Joint noises	97	–0.02	1.5	1.5
End-feel*	88	–0.05	8.7	2.9
Translation right				
Joint play	84	0.08	13.4	6.0
Pain	85	0.50	14.9	20.9
Joint noises	76	0.07	19.4	10.4
End-feel	85	0.20	—	—
Traction left				
Joint play	80	–0.05	16.7	3.0
Pain	93	0.25	3.0	7.5
Joint noises	99	0.66	1.5	3.0
End-feel	90	–0.05	6.0	4.5
Translation left				
Joint play	77	–0.10	10.7	6.1
Pain	74	0.28	22.7	24.2
Joint noises	73	–0.02	22.7	7.6
End-feel	76	–0.13	14.9	9.0
MTS Traction and Translation				
Joint play	75	0.08	22.2	9.5
Pain	75	0.46	35.4	35.4
Joint noises	64	–0.01	29.6	19.7
End-feel	70	0.07	25.0	14.1

*Abnormal.
% A = percentage agreement; D = frequency (%) of the signs registered by the dentist; P = frequency (%) of the signs registered by the physiotherapist.

Table 4 Interexaminer Reliability of the Compression Tests (n = 57)

	% A	κ	D	P
Compression right				
Pain	83	0.19	10.3	13.8
Joint sounds	98	NA	0	1.7
Compression left				
Pain	93	0.47	3.5	10.5
Joint sounds	100	1.00	1.8	1.8
MTS compression				
Pain	83	0.40	14.0	21.1
Joint sounds	98	0.66	1.8	3.6

% A = percentage agreement; D = frequency (%) of the signs registered by the dentist; P = frequency (%) of the signs registered by the physiotherapist; NA = not applicable.

the different tests in our study showed different reliabilities, the three main symptoms of TMD could be determined reliably (Table 7).

Haas[32] and Haas et al[33] noted that kappa values become unstable when there is high agreement for only one of the possible rating choices. For example, observed agreement (%) can be high in a healthy group of subjects as a result of the total absence or the low frequency of positive findings in this group. According to Haas et al,[33] the slightest disagreement can be strongly penalized by the kappa statistic when a large proportion of agreement (> 85%) is for only one of the possible alternative categories. This is termed *limited variation*.[32-34] As a result of the high percentage of agreement and the lack of positive findings in some tests, these problems were also encountered in the present study (see Tables 2 to 6). For example, lim-

Table 5 Interexaminer Reliability of the Static Pain Tests (Isometric Resistance) (n = 45)

	% A	κ	D	P
Open				
Pain	75	0.24	19.1	23.4
Close				
Pain	77	0.30	23.3	18.6
Lateral right				
Pain	76	0.28	26.7	15.6
Lateral left				
Pain	69	0.26	35.6	22.2
MTS static pain tests				
Pain	57	0.15	52.3	36.41

% A = percentage agreement; D = frequency (%) of the signs registered by the dentist; P = frequency (%) of the signs registered by the physiotherapist.

Table 6 Interexaminer Reliability of the Palpation Tests (n = 79)

	% A	κ	D	P
Masseter, superficial				
Pain	62	0.16	60.8	73.4
Muscle tone (abnormal)	70	0.07	15.2	25.3
Structural changes	68	0.03	12.7	26.6
Masseter, deep				
Pain	58	0.19	51.9	17.7
Muscle tone	70	0.03	30.4	2.5
Structural changes	76	−0.05	21.5	2.5
Temporalis, anterior				
Pain	72	0.45	45.6	53.2
Muscle tone	84	−0.08	5.1	11.4
Structural changes	85	0.06	7.6	10.1
Temporalis, posterior				
Pain	84	0.23	13.9	10.1
Muscle tone	99	NA	1.3	0.0
Structural changes	99	NA	1.3	0.0
Pterygoid, medial				
Pain	60	0.23	31.6	57.0
TMJ				
Lateral pole, open mouth	56	0.16	58.4	35.1
Lateral pole, mouth closed	66	0.34	57.1	44.2
Dorsal via external meatus	65	0.13	33.8	19.5
MTS pain on palpation				
Masseter	75	0.33	75.9	73.4
Temporalis	71	0.42	48.1	54.4
Pterygoid medial attachments	60	0.23	31.6	57.0
Total masticatory muscles	87	0.51	83.5	86.1
TMJs	68	0.33	70.1	53.2

% A = percentage agreement; D = frequency (%) of the signs registered by the dentist; P = frequency (%) of the signs registered by the physiotherapist; NA = not applicable.

ited variation existed in the traction test because of the lack of positive findings (see Table 3). Since diagnostic decisions are based mainly on multiple test findings, MTS were formed for each individual test and for the main signs and symptoms of TMD.[32] However, we intentionally did not combine the heterogeneous findings, pain, joint noises, and restriction of movement, as described in the literature,[2,18] because this would unnecessarily mask the reliability of the individual diagnostic components.

Table 7 Interexaminer Reliability of the MTS of Combinations of Tests Determined for the Main Signs and Symptoms of TMD

	% A	κ
MTS pain		
during active movements	65	0.32
during function (active movements and/or additional tests)	89	0.73
during function and/or palpation	96	0.78
MTS joints sounds		
during active movements	80	0.61
MTS restriction of movement		
during active movements	92	0.56

% A = percentage agreement.

The physiotherapist detected more signs in the separate compression test and the muscle palpation tests than did the dentists, but fewer signs in the tests involving active movements, passive opening, traction and translation, and TMJ palpation (see Tables 3 to 7). These differences were compensated in the MTS, and McNemar's test showed no statistically significant differences between the dentists and the physiotherapist.

The patients included in this study were examined by one of the four dentists and by the physiotherapist. Because differences between the dentists and the physiotherapist did not depend on the dentist who performed the tests, the results of all patients are presented in this study without further specification of which dentist performed the examination.

Restricted Movement

The interexaminer reliability for assessment of continuous variables (mouth opening measured in millimeters) was good, in agreement with other studies.[13-15,17,18,35,36] Kopp and Wenneberg[14] reported a large fluctuation of clinical signs with time but little variability when vertical jaw opening was assessed. Dworkin et al[17] found that noncalibrated examiners had lower reliability scores than did calibrated examiners, but both had a high interexaminer reliability score for active and passive opening. We investigated the interexaminer reliability of routinely used orthopedic tests without further calibration of the examiners during data collection. This may explain our lower ICC for active opening compared with that reported by Dworkin et al.[17] Moreover, we examined TMD patients, which may have resulted in more differences in possible rating choices because of the variability of most signs and symptoms of TMD; other investigators studied asymptomatic or healthy subjects.[13,14,36,37]

If a patient is unable to complete an active range of motion or if test results seem inconclusive, the tests of the passive range of motion are indicated.[38] According to Friedman et al[24] and Hansson et al,[6] passive-movement and joint-play tests aim at loading the TMJ while the muscles are relaxed, thus providing information about the function of the TMJ only. An increased end-feel distance (> 2 mm) indicates a myogenous involvement. The criteria used to define a symptomatic joint are an abnormal end-feel; abnormal quality of resistance to motion; and pain, both local or referred. Each joint motion may have some natural variation of normal end-feel because of the individual differences. In our study, joint play could not be assessed with adequate reliability. Other authors have also reported poor results for interexaminer reliability when testing joint play or end-feel of other joints. This might be due to the lack of standards.[39-42] The variability of test results may be caused by the differences in the force and test directions used in these tests. Changes in the passive range of movements are also possible due to the tolerance of the patient, the patient being able to relax during these tests, and the intensity of the signs and symptoms. Another possible explanation for the variation of clinical signs and symptoms, for example, in the passive joint-play tests, is that the patient is more familiar with the test during the second session and can be influenced by experience or by a different attitude of the examiner, or by socially desirable behavior.[37]

Joint Sounds

The interexaminer reliability for joint sounds in a clinical setting was acceptable ($0.47 \leq \kappa \leq 0.59$) for the individual active movements and high for the MTS clicking on active movements. This is in line with the results of Dworkin et al.[17] These authors found a marked difference between calibrated and noncalibrated examiners for both palpation (κ = 0.62 and 0.30) and stethoscopy (κ = 0.61 and 0.35) in the detection of joint sounds. We used results obtained with the palpation technique in the analysis. These results were consistent with Dworkin's data[17] for the detection of joint sounds on vertical opening, whereas our results for the detection of joint sounds on lateral excursions were better than the stethoscopy results of Dworkin's study. These differences may be caused

by the different populations examined. We recruited patients with TMD, but Dworkin selected asymptomatic healthy volunteers as well. A high agreement between both examiners would be expected in patients with internal derangement with reduction, as a result of the intra-articular morphologic changes that can be the reason for the reproducibility of the joint sounds.[5,43] Recently Wabeke et al[44] compared the palpation method and stethoscopy in a nonpatient population. Sound was assessed electronically. Stethoscopy was found to be more sensitive, particularly with regard to crepitation. It was concluded that both palpation and stethoscopy can be used in clinical settings but that both methods have their limitations.

The detection of joint sounds was far from reliable in the passive opening and TMJ traction and translation tests. This may be because of the extent of mouth opening that is needed to manipulate the jaw and alterations in joint position caused by these manipulations. Limited variation may also have influenced the results.

Pain

Pain during palpation is a relevant sign for diagnosis and for treatment evaluation. Some authors claim that a pressure algometer provides a more reliable means of evaluating this sign.[45,46] This may be true for pain intensity, but it is also important to get an impression of structural changes in the muscles, of the tissue reaction to palpation, such as twitches in combination with trigger points,[47] and of changes in muscle tonus and pain during palpation. Unlike Bergholz[15] and Duinkerke et al,[37] we did not have even moderate reliability scores for pain on palpation of the muscles, with exception of the anterior temporal muscles, and the joints. Dworkin et al[16] found that retraining improved reliability more for the muscles than for the TMJ. The difference between these studies may be influenced by the fact that some investigators used healthy people,[14,37] and the high percentage of no-pain results for the asymptomatic subjects can influence the agreement between examiners. Another reason may be the examination procedure itself. In our study, the masticatory system was palpated after the functional tests, at the end of the first and second examinations. Thus, muscles could have been influenced by the foregoing tests, and this may have had consequences on the pain response to palpation.[48] Dworkin et al[17] stated that clinical signs and symptoms, like pain, can be more variable than what would be suggested by the results of some authors, and they may change spontaneously over time, making it difficult to achieve the same score. The clinical signs and symptoms of mandibular dysfunction themselves can therefore be variable and influence reliability results between examiners. Because pain is a subjective experience, its sensory and affective components can also give rise to variability. Other sources of variability have been described, such as the amount of pressure applied, different palpation techniques, differences in size of the distal phalanx of the examiners, different specific anatomic areas palpated, and different interpretations by examiners or patients.[4,49] Dworkin et al[17] made a summary score for each muscle group (extraoral, intraoral), which improved the agreement between examiners. In our study, the MTS were also better than the individual scores, which is also consistent with the findings of Fricton and Schiffman.[18] Thus, the assessment of pain on palpation seems to be highly variable, which may reflect both the low intraindividual constancy and the high observer variability. The MTS on palpation and/or a pressure algometer seems to provide a more reliable means for evaluating the intensity of pain on palpation.

Several authors have concluded that it is not possible to register pain reliably in a functional examination of the masticatory system.[13,14] However, Dworkin et al[17] found good results ($\kappa > 0.70$) for the interexaminer assessment of pain during active movements. We did not find similar results for the individual tests. The reliability score for pain on active laterotrusion and protrusion of our examiners was equal to the fair score of the noncalibrated examiners in Dworkin's study.[17] Pain during function was reliable when assessed with a multitest score ($\kappa = 0.78$).

The static pain-resistance test was used to evaluate pain on muscle contraction. We did not use the dynamic pain test because this test loads both the masticatory muscles and the TMJ, and as such gives no additional information to that provided by the other tests. Another reason was that in clinical practice, it is impossible not to elicit reactions in the neck during execution of the dynamic pain test.

Variation in Clinical Signs and Symptoms

The clinical signs and symptoms of TMD can themselves be variable.[14,17] However, this variability of signs and symptoms may also be the result of a lack of standards and may be influenced by several therapist- and patient-related aspects.[16] A large observer variation is not restricted to the clinical registration of signs and symptoms of the locomo-

tor system. Other authors[50–52] have described large variations in the recording of cardiovascular signs, gastrointestinal signs, and respiratory signs. In a clinical setting, we must accept a certain amount of variation in signs and symptoms, and for this reason, it is important to not base a diagnosis on the results of a single test. The test results, history, and clinical examination must point in the same direction to arrive at a proper working diagnosis.

Conclusion

Most variables of the examination could be determined with satisfactory reliability during active movements. However, although additional individual tests (traction, translation, resistance) were less reliable in evaluating the main signs and symptoms of TMD, the three main symptoms of TMD could be determined reliably.

Computing multitest scores improves the reliability of tests used to assess patients with TMD. It is advisable that clinicians calibrate their techniques regularly to improve the reliability of results in daily practice.

Acknowledgments

The authors gratefully acknowledge the clinical contributions to this study made by Dr R. Buchner, Dr F. Lobbezoo, and Dr L. W. Olthoff. They also wish to thank Dr J. A. J. Faber for providing valuable statistical advice.

References

1. American Academy of Orofacial Pain. McNeill C (ed). Temporomandibular Disorders. Guidelines for Classification, Assessment, and Management. Chicago: Quintessence, 1993.
2. Helkimo M. Studies on function and dysfunction of the masticatory system. Index for anamnestic and clinical dysfunction and occlusal state. Swed Dent J 1974;67: 101–121.
3. McNeill C, Danzig WM, Farrar WB, Gelb HY, Lerman MD, Mossett BC, et al. Craniomandibular (TMJ) disorders—The state of the art. Position paper of the American Academy of Craniomandibular Disorders. J Prosthet Dent 1980;44:434–437.
4. Fricton JR, Kroening RJ, Hathaway KM. TMJ and Craniofacial Pain: Diagnosis and Management. St Louis, MO: Ishiyaka EuroAmerica, 1988.
5. De Boever JA, Steenks MH. Epidemiologie, Symptomatologie und Ätiologie der Kraniomandibulären Dysfunktionen. In: Steenks MH, de Wijer A (herausgeber). Kiefergelenksfehlfunktionen aus Physiotherapeutischer und Zahnmedizinischer Sicht, Diagnose und Therapie. Berlin: Quintessenz, 1991:35–45.
6. Hansson T, Wessman C, Öberg T. Säkrare diagnoser med ny teknik, Förslag till funktionsbedömning av käkleder och tuggmuskler. Tandläkartidningen 1980;72(24): 1372–1374.
7. Janda V, Lewit K. Krankengymnastik und die muskuläre Fehlsteuerung der Wirbelsäule. In: Lewit K (ed). Manuelle Therapie, im Rahmen der ärztlichen Rehabilitation. Leipzig: Johann Ambrosius Barth, 1973:288–359.
8. Kaltenborn FM. Manuelle therapie der Extremitätengelenke. Oslo: Olaf Norlis Bokhandel, 1976.
9. Solberg WK. Temporomandibular disorders: Physical tests in diagnosis. Br Dent J 1986;19:273–278.
10. Palla S. Neue Erkenntnisse und Methoden in der Diagnostik der Funktionsstörungen des Kausystems. Schweiz Monatsschr Zahnmed 1986;96:1329–1351.
11. Steenks MH, de Wijer A (ed). Kiefergelenksfehlfunktionen aus physiotherapeutischer und zahnmedizinischer Sicht, Diagnose und Therapie. Berlin: Quintessenz, 1991.
12. Lobbezoo-Scholte AM. Diagnostic Subgroups of Craniomandibular Disorders [PhD dissertation]. Utrecht, Netherlands, Utrecht Univ, 1993.
13. Carlsson GE, Egermark-Eriksson I, Magnusson T. Intra- and inter-observer variation in functional examination of the masticatory system. Swed Dent J 1980;4:187–194.
14. Kopp S, Wenneberg B. Intra- and interobserver variability in the assessment of signs of disorder in the stomatognathic system. Swed Dent J 1983;7:239–246.
15. Bergholz P. Zur Untersucherubereinstimmung bei der klinischen Funktionsanalyse nach Krogh-Poulsen. Dtsch Zahnärtztl Z 1985;40:182–185.
16. Dworkin SF, LeResche L, DeRouen T. Reliability of clinical measurement in temporomandibular disorders. Clin Pain 1988;4(2):89–99.
17. Dworkin SF, LeResche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. J Prosthet Dent 1990;63:574–579.
18. Fricton JR, Schiffman EL. Reliability of a craniomandibular index. J Dent Res 1986;65:1359–1364.
19. Lobbezoo-Scholte AM, de Wijer A, Steenks MH, Bosman F. Interexaminer reliability of six orthopaedic tests in diagnostic subgroups of craniomandibular disorders. J Oral Rehabil 1994;21:273–285.
20. Rocabado M. Biomechanical relationship of the cranial, cervical and hyoid regions. J Craniomand Pract 1983; 3:62–66.
21. Braun BL. Postural differences between asymptomatic men and women and craniofacial pain patients. Arch Phys Med Rehabil 1991;72:653–657.
22. Keating JC, Bergmann TF, Jacobs GE, Finer BA, Larson K. Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. J Manipulative Phys Ther 1990;13:463–470.
23. Hansson T, Honée W, Hesse J. Craniomandibulaire Dysfunctie. Alphen aan de Rijn/Brussel: Samson Stafleu, 1985.
24. Friedman MH, Weisberg J. Joint play movements of the temporomandibular joint: Clinical considerations. Arch Phys Med Rehabil 1984;65:413–417.
25. Friedman MH, Weisberg J. Application of orthopedic principles in evaluation of the temporomandibular joint. Phys Ther 1982;62: 597–603.
26. Friedman MH, Weisberg J. Temporomandibular Joint Disorders, Diagnosis and Treatment. Chicago: Quintessence, 1985.

27. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measurement 1960;20:37–46.
28. Cohen J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. Psychol Bull 1968;70:213–220.
29. Fleiss JL. Statistical methods for rates and proportions, ed 2. New York: John Wiley & Sons, 1981.
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
31. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420–423
32. Haas M. Interexaminer reliability for multiple diagnostic test regimens. J Manipulative Physiol Ther 1991;14: 95–103.
33. Haas M, Nyiendo J, Peterson C, Thiel H, Sellers T, Cassidy D, Yong-Hing K. Interrater reliability of roentgenological evaluation of the lumbar spine in lateral bending. J Manipulative Physiol Ther 1990;13:179–189.
34. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the Kappa Statistic. Arch Gen Psychiatry 1985;42:725–728.
35. Mezitis M, Rallis G, Zachariades N. The normal range of mouth opening. J Oral Maxillofac Surg 1989;47: 1028–1029.
36. Westling L, Helkimo E, Mattiasson A. Observer variation in functional examination of the temporomandibular joint. J Craniomandib Disord Facial Oral Pain 1992; 6:202–207.
37. Duinkerke ASH, Luteijn F, Bouman TK, de Jong HP. Reproducibility of a palpation test for the stomatognathic system. Community Dent Oral Epidemiol 1986;14: 80–85.
38. Hoppenfeld S. Physical examination of the spine and the extremities. New York: Appleton-Century-Crofts, 1976:128–133.
39. Cooperman JM, Riddle DL, Rothstein JM. Reliability and validity of judgments of the integrity of the anterior cruciate ligament of the knee using the Lachman's test. Phys Ther 1990;70:225–233.
40. Gonella C, Paris SV. Reliability in evaluating passive intervertebral motion. Phys Ther 1982;62:436–444.
41. McClure PW, Rothstein JM, Riddle DL. Intertester reliability of clinical judgments of medial knee ligament integrity. Phys Ther 1989;69:268–275.
42. Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. Phys Ther 1985;65: 1671–1675.
43. Stegenga B. Temporomandibular joint osteoarthrosis and internal derangement [PhD dissertation]. Netherlands, Univ Groningen, 1991.
44. Wabeke KB, Spruijt RJ, van der Zaag J. The reliability of clinical methods for recording temporomandibular joint sounds. J Dent Res 1994;73:1157–1163.
45. Sung-Chang C, Bo-Yong U, Hyung-Suk K. Evaluation of pressure pain threshold in head and neck muscles by electronic algometer: Intrarater and interrater reliability. J Craniomand Pract 1992;10:28–35.
46. Murphy GJ, McKinney MW, Gross WG. Temporomandibular-related pressure thresholds: A model for establishing baselines. J Craniomand Pract 1992;10: 118–123.
47. Travell JG, Simons DG. Myofascial pain and dysfunction. Triggerpoint manual, vol 1. Baltimore: Williams and Wilkins, 1983.
48. Jaeger B, Reeves JL. Quantification of changes in myofascial trigger point sensitivity with the pressure algometer following passive stretch. Pain 1986;27:203–210.
49. Smith JP. Observer variation in the clinical diagnosis of mandibular pain dysfunction syndrome. Community Dent Oral Epidemiol 1977;5:91–93.
50. Brooks D, Wilson L, Kelsey C. Accuracy and reliability of 'specialized' physical therapists in auscultating tape-recorded lung sounds. Physiother Can 1993;45:21–24.
51. Godfrey S, Edwards RTH, Campbell EJM, Armitrage P, Oppenheimer EA. Repeatability of the physical signs in airway obstruction. Thorax 1969;24:4–9.
52. Koran LM. The reliability of clinical methods, data and judgments. N Engl J Med 1975;293:642–646.

Resumen

La Fiabilidad de los Hallazgos Clínicos en los Desórdenes Temporomandibulares

El propósito de esta investigación fue el de estudiar la fiabilidad entre los examinadores en relación a los exámenes ortopédicos y técnicas de palpación utilizadas rutinariamente en el diagnóstico clínico de los desórdenes del sistema masticatorio. Los exámenes fueron realizados por un dentista y un fisioterapeuta quienes utilizaron los exámenes rutinariamente cuando examinaron los pacientes con desórdenes temporomandibulares (DTM). Este estudio tuvo 79 participantes. En el análisis se utilizaron los porcentajes que reflejaban los acuerdos, las correlaciones entre las clases, y el kappa de Cohen. La fiabilidad de los exámenes entre los examinadores que median la apertura bucal activa máxima y los registros de los sonidos de "clic" durante la apertura bucal activa fue alta. La fiabilidad entre los examinadores fue regular en relación a los exámenes que midieron la intensidad del dolor durante los movimientos activos, y moderada en el caso de los exámenes que registraban los sonidos de la articulación (κ = 0,47 a 0,59). Se detectó un acuerdo alto entre los observadores en varios detalles de los exámenes de tracción y traslación, aunque los valores kappa fueron bajos. La fiabilidad entre los examinadores en relación a los valores de exámenes múltiples de compresión, fue considerable en los sonidos de la articulación (κ = 0,66) y regular en el dolor (κ = 0,40). La fiabilidad entre los examindaores en los valores de los exámenes múltiples relacionados a la palpación de los músculos y de la articulación fue moderada (κ = 0,51) y regular (κ = 0.33), respectivamente. Se puede concluir que la mayoria de las variables determinadas durante los movimientos activos pueden ser medidas con una fiabilidad satisfactoria, mientras que, las variables para los otros exámenes no son medidas con la misma fiabilidad en base a las valores kappa. Los síntomas principales de los DTM pueden ser evaluados con seguridad con los valores de los exámenes múltiples. Se recomienda que los clínicos deben calibrar sus técnicas regularmente para mejorar la fiabilidad de los resultados en la práctica diaria.

Zusammenfassung

Verlässlichkeit von klinischen Befunden bei Myoarthropathien des Kausystems

Das Ziel der vorliegenden untersuchung war, die Verlässlichkeit von orthopädischen Tests und Palpationstechniken zwischen verschiedenen Untersuchern zu studieren, Techniken, die routinemässig für die klinische Diagnostik von Myoarthropatien des Kausystems (MAP) zur Anwendung gelangen. Die Tests wurden von einem Zahnarzt und von einem Physiotherapeuten durchgeführt, beide brauchten diese Tests routinemässig zur Untersuchung von Patienten mit MAP. 79 Patienten nahmen an der Studie teil. Zur Analyse wurden die prozentuale Übereinstimmung, dei Korrelation innerhalb der Klasse und Cohens Kappa verwendet. Es gab eine hohe Übereinstimmung zwischen den Untersuchern beim Messen der maximalen aktiven Mundöffnung und bei der Registrierung des Knackens während der aktiven Mundöffnung. Dagegen war die Zuverlässigkeit gut für die Tests, die die Intensität des Schmerzes während der aktiven Bewegung bestimmen und mässig für die Tests, die Gelenkgeräusche aufzeichnen (κ = 0,47 bis 0,59). Es gab eine hohe Übereinstimmung zwischen den Untersuchern bei einzelnen Einheiten des Traktions- und Translationstests, auch die Kappawerte waren tief. Die Übereinstimmung des multitest scores war ansehnlich für Gelenkgeräusche (κ = 0,66) und gut für Schmerz (κ = 0,40). Die Übereinstimmung des multitest scores für Muskel- resp. Gelenkpalpation war mässig (κ = 0,51) resp. gut (κ = 0,33). Man kommt zum Schluss, dass die meisten der während der aktiven Bewegungen determinierten Variabeln mit genügender Verlässlichkeit gemessen werden können, während Variabeln anderer Tests nicht mit derselben Verlässlichkeit gemessen werden können (kappa). Die Hauptsymptome der MAP können mit multitest scores zuverlässig erhoben werden. Klinikern wird empfohlen, ihre Technik regelmässig zu kalibrieren, um die Zuverlässigkeit der Resultate in der täglichen Praxis zu verbessern.