The Research Diagnostic Criteria for Temporomandibular Disorders. II: Reliability of Axis I Diagnoses and Selected Clinical Measures

John O. Look, DDS, PhD, MPH
Senior Research Associate
School of Dentistry

Mike T. John, DDS, PhD
Associate Professor
School of Dentistry

Feng Tai, MS, PhD
Research Assistant
Division of Biostatistics

University of Minnesota
Minneapolis, Minnesota

Kimberly H. Huggins, RDH, BS
Affiliate Instructor
Department of Oral Medicine
University of Washington, School of Dentistry
Seattle, Washington

Patricia A. Lenton, RDH, MA
Research Fellow
University of Minnesota, School of Dentistry
Minneapolis, Minnesota

Edmond L. Truelove, DDS
Professor
Department of Oral Medicine
University of Washington, School of Dentistry
Seattle, Washington

Richard Ohrbach, DDS, PhD
Associate Professor
Department of Oral Diagnostic Sciences
University at Buffalo
Buffalo, New York

Gary C. Anderson, DDS, MS
Associate Professor

Eric L. Schiffman, DDS, MS
Associate Professor

School of Dentistry
University of Minnesota
Minneapolis, Minnesota

Correspondence to:
Dr John O. Look
School of Dentistry, University of Minnesota
6-320 Moos Tower, 515 Delaware Street SE
Minneapolis, MN 55455
Fax: 612-626-0138
Email: lookj@umn.edu

Aims: *The primary aim was to determine new estimates for the measurement reliability of the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) Axis I diagnostic algorithms. A second aim was to present data on the reliability of key clinical measures of the diagnostic algorithms.* ***Methods:*** *Kappa (κ), computed by generalized estimate equation procedures, was selected as the primary estimate of interexaminer reliability. Intersite reliability of six examiners from three study sites was assessed annually over the 5-year period of the RDC/TMD Validation Project. Intrasite reliability was monitored throughout the validation study by comparing RDC/TMD data collections performed on the same day by the test examiner and a criterion examiner.* ***Results:*** *Intersite calibrations included a total of 180 subjects. Intersite reliability of RDC/TMD diagnoses was excellent (κ > 0.75) when myofascial pain diagnoses (Ia or Ib) were grouped. Good reliability was observed for discrete myofascial pain diagnoses Ia (κ = 0.62) and Ib (κ = 0.58), for disc displacement with reduction (κ = 0.63), disc displacement without reduction with limited opening (κ = 0.62), arthralgia (κ = 0.55), and when joint pain (IIIa or IIIb) was grouped (κ = 0.59). Reliability of less frequently observed diagnoses such as disc displacements without reduction without limited opening, and osteoarthrosis (IIIb, IIIc), was poor to marginally fair (κ = 0.31–0.43). Intrasite monitoring results (n = 705) approximated intersite reliability estimates. The greatest difference in paired estimates was 0.18 (IIc).* ***Conclusion:*** *Reliability of the RDC/TMD protocol was good to excellent for myofascial pain, arthralgia, disc displacement with reduction, and disc displacement without reduction with limited opening. Reliability was poor to marginally fair for disc displacement without reduction without limited opening and osteoarthrosis.*
J OROFAC PAIN 2010;24:25–34

Key words: diagnostic criteria, reliability, temporomandibular joint disorders, temporomandibular muscle and joint disorders

For the purposes of research addressing temporomandibular disorders (TMD), the most commonly used diagnostic classification protocol is the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD).[1] The full acceptance of the RDC/TMD by dental and medical personnel as a taxonomic system for TMD is necessarily predicated on a rigorous assessment of its reliability and validity, and this was the aim of the multisite RDC/TMD Validation Project.

It has been postulated that the reliability of a diagnostic instrument sets the upper limit for its validity.[2] However, a review of past reliability studies reveals that the entire RDC/TMD protocol has not been adequately tested. Some single-center studies were designed only to evaluate certain clinical signs and a subset of RDC/TMD diagnoses for purposes of examiner calibration prior to initiation of subsequent research.[3,4] Other studies assessed the reliability of modified RDC/TMD-type palpation procedures by employing palpation pressures up to two times greater than the pressures specified for the RDC/TMD.[5–8]

A major advance in reliability studies of the RDC/TMD occurred with the formation of the International Consortium for RDC/TMD-based Research. Reliability testing by international clinical TMD researchers has been performed for all eight RDC/TMD diagnoses. There are now such estimates of reliability from 10 international sites with a total of 30 examiners and 230 participants.[9] Three of the 10 sites have also published individual findings in greater detail.[10–12] This international initiative has provided good heterogeneity with respect to participants and examiners, and offers greater potential for generalizability of the findings than do the previous single-site studies. However, the evidence for establishing the reliability of the RDC/TMD protocol, from both the international and the single-site studies, must still be characterized as weak because the sample sizes have been too low for estimation of 95% confidence intervals. Thus, the role of chance as it might affect the magnitude of the point estimates of reliability has not been assessed.

The primary aim of this study was to determine new estimates for the measurement reliability of the RDC/TMD Axis I diagnostic algorithms. A second aim was to present data on the reliability of key clinical measures of the diagnostic algorithms.

Materials and Methods

Setting

Data collections were carried out at three sites: the University at Buffalo (UB), the University of Min - nesota (UM), and the University of Washington (UW). A total of nine clinicians served as the examiners for the RDC/TMD Validation Project, including two criterion examiners (CEs) and one test examiner/dental hygienist (TE) for each study site. All six CEs were TMD and orofacial pain experts with between 12 and 38 years of experience in research and treatment of TMD. The three TEs were dental hygienists who were trained and calibrated to perform the RDC/TMD examination protocol. The separate data collections for the reliability and validity assessments in this project were performed concurrently so that the reliability estimates would be temporally relevant to support the credibility of the validity estimates reported in the third article in this series.[13]

Intersite and Intrasite Training, Calibration, and Reliability Assessment

At the beginning of the study, the nine examiners were convened for a 3-day meeting at UM to review the operational definitions of the published RDC/TMD, to formally operationalize the new diagnostic tests authorized by the External Advisory Panel (AP) appointed by the US National Institute of Dental and Craniofacial Research (NIDCR) for the project, and to undergo training for all of the clinical measurements. Following the initial session, ongoing reliability assessment for the clinical measurements and algorithmic diagnoses was based principally on two methods.

The first method for reliability assessment involved formal intersite calibration studies that were conducted annually over the 5-year project period with six examiners, including all three TEs but just one CE from each site. These calibration exercises consisted of measurement of all examination items as specified by the operational definitions of the RDC/TMD. For categorical measures, including the RDC/TMD diagnoses, kappa (κ) = 0.4 was set as the minimum acceptable level for agreement. A kappa of 0.4 represents 70% agreement when the characteristic being measured has 50% prevalence in the participant sample. The minimum goal for intraclass correlation coefficients (ICC) was 0.70. All intersite reliability sessions were carried out at UM. The methodological details for the assessment of intersite reliability are provided in detail below.

The second method for reliability assessment was focused on intrasite reliability, and was accomplished concurrently with the validation study that assessed the eight Axis I diagnoses against the reference standard diagnoses that are described in the third article in this series.[13] During the formal validation study, one of the CEs and the TE from each site each performed examinations of the same participant the same day while blinded to the other's findings. The criterion examination protocol included all the RDC/TMD examination items as well as a much-expanded set of

diagnostic procedures described in the first article in this series.[14] Thus, the reliability of the TE relative to the CE could be monitored with respect to RDC/TMD examination items and the algorithmic diagnoses based on the exam findings. Because the two CEs at each site alternated between successive participants in their role as first or second examiner, the reliability of the TE was compared to both CEs over the course of the validation study.

Participant Recruitment for Intersite Reliability Assessment

One of the design goals for the intersite calibrations was for the participants to be as similar as possible to the participants recruited for the formal validation study. Putative case status, as specified for the validation study, included individuals who reported minimum or mild TMD symptoms. Thus, calibration cases and controls were drawn from lists of available participants, irrespective of the severity of their TMD condition. Putative controls for the validation study required a report of a lifetime absence of jaw pain, temporomandibular joint (TMJ) sounds or jaw locking, and this was also specified for the reliability studies. A complete presentation of the inclusion and exclusion criteria for the validation study is reported in the first article of this series.[14] In brief, exclusions from participation as a calibration participant were based on any of the following findings obtained from history, medical reports, examination, or joint imaging:

- **Medical conditions:** Systemic rheumatic, neurologic/neuropathic, endocrine, collagen vascular or (auto)immune conditions; pregnancy.
- **Prior history:** head/neck radiation; TMJ surgery; internal derangements of the TMJ other than disc displacement or arthrosis; trauma from any cause ″ 2 months.
- **Orofacial conditions:** any non-TMD orofacial pain/ neuralgia/neuropathy; odontogenic pain/ infection.
- **Medications:** unable to discontinue use of muscle relaxants, narcotic and nonsteroidal anti-inflammatory pain relievers ≥ 3 days prior to any study visit; unstable dose for anti-depressant therapy during prior 60 days; current illicit drug use.

Calibration participant recruitment differed from that of the validation study by the fact that no attempt was made to selectively enrich the calibration participant sample for the less common diagnoses (IIb, IIc, IIIb, and IIIc), although this practice was needed for the validation study as it progressed. Also, no calibration participants were respondents to study flyers and advertisements. All were participants of record with their temporomandibular status having been established at the TMD and Orofacial Pain Clinic of UM, or by their participation in either the validation study or another TMD study at UM.

Participant Enrollment and Preparation for Intersite Reliability Assessment

The research manager at UM assessed, by phone or in-person interview, all calibration study participants based on the same inclusion and exclusion criteria as those for the formal validation study described in the first article in this series.[14] The research manager also had the responsibility for ordering or preparing all calibration participant charts, securing all data collection forms, training of recorders, scheduling of participants, and preparation of the examination cubicles. After seating the participant, the research manager reviewed the participant's medical history for change in status, requested the participant read the participant information sheet, answered any questions the participant had, and reminded the participant not to discuss any prior examination with a subsequent examiner. The principal investigator then determined that the participant fully understood the procedure and obtained consent. All procedures were reviewed and approved by the Institutional Review Boards overseeing each study site.

Examination Sequence for Intersite Reliability Assessment

All calibration participants underwent three examinations in a single visit. The study epidemiologist prepared a randomly ordered examination sequence by first assigning examiners to a participant in groups of three, with each examiner designated a, b, or c. Their within-group order of examination was then randomly selected from among abc, acb, bca, bac, cab, or cba. If, for example, "bca" was selected, then the next participant for the same group of examiners was ordered "cab," and the third "abc." This rotation was designed to control for order effects related to examination-induced sensitization of the participant. The principle was to equalize among examiners, as much as possible, the differential bias that is associated with changes in the participant's temporomandibular status due to repeated examinations being done. After completion of the third examination, the three examiners involved with a given

Table 1 Twelve Pair-wise Comparisons Required for Intersite Calibration Studies

	B1	B2	M3	M4	W5	W6
B1			X	X	X	X
B2			X	X	X	X
M3					X	X
M4					X	X

B1 and B2 represent the two examiners from UB, M3 and M4 are those from UM, and W5 and W6 are those from UW. X: required pairing.

participant regrouped with this participant still present to discuss any differences that they found, but their data collections were not altered.

Each annual intersite calibration exercise had a total of 36 participants (with two additional alternates). Typically, 33 participants were TMD cases, and three were normal participants. All examinations were performed with the examiners blinded to the participants' temporomandibular status. Each examiner examined 18 participants, and examination pairings were structured to allow eight out of the 18 participants to be also examined by each of the two examiners from the other two centers. Thus, there were 12 pair-wise examiner comparisons of interest that matched each examiner from a site with the four examiners from the two other sites (Table 1). Examiners from the same site were not paired with each other; this assessment was left for informal within-site training exercises and intrasite monitoring.

Reliability Estimates Specified for this Study

RDC/TMD Diagnoses. The primary purpose was to evaluate the reliability of the eight RDC/TMD Axis I diagnoses that include:

- **Group I** Muscle Disorders: (Ia) myofascial pain; (Ib) myofascial pain with limited opening.
- **Group II** Disc Displacements: (IIa) disc displacement with reduction; (IIb) disc displacement without reduction with limited opening; (IIc) disc displacement without reduction without limited opening.
- **Group III** Arthralgia, Arthritis, Arthrosis: (IIIa) arthralgia; (IIIb) osteoarthritis; (IIIc) osteo - arthrosis.

Additional collapsed groups of diagnoses specified for testing: (1) Ia or Ib: any Group I diagnosis, *(2)* IIa, IIb, or IIc: any Group II diagnosis, *(3)* IIIa or IIIb: any joint pain diagnosis, and *(4)* IIIb or IIIc: any arthrosis.

For each of these diagnoses or collapsed groups of diagnoses, the study measured study sample prevalence, overall percent agreement, point estimates for kappa and the ICC, and the 95% confidence intervals (CIs) for kappa and the ICC.

Individual Clinical Measures Selected for Their Explanatory Value Relative to the Reliability of RDC/TMD Diagnoses

All of the RDC/TMD Axis I diagnoses are derived from diagnostic algorithms using a classification tree design. The theory and the uses of classification tree methods are reviewed in the Discussion. The Group I algorithm includes 5 nodes, or diagnostic decision branches, some of which are defined by multiple clinical measures. The Group II algorithm has 12 nodes involving many exam and questionnaire variables. The Group III algorithm has 3 nodes, each involving multiple variables as well. The reliability of each of these diagnostic algorithms is a function of the reliability of the clinical measures that make up the trees. For example, the diagnosis of myofascial pain (Group I) is limited by the reliability of the report of pain in response to muscle palpation.

As a secondary analysis, key clinical examination measures with TMD diagnostic importance were selected for reliability assessment. These examination measures were muscle pain, limited opening, disc click, joint pain, and coarse crepitus, and all employed a dichotomous scoring system that facilitated assessment of kappa and percent agreement among examiners. *Muscle pain* is based on the sum of positive pain responses to palpation of 16 extraoral masticatory muscle sites and 4 intraoral muscle sites. In the diagnostic algorithm for myofascial pain (Group I), this variable is dichotomous; 0 to 2 positive pain sites (a negative finding) are differentiated from 3 to 20 positive pain sites (a positive finding). Thus, assessment of agreement was based on this dichotomy. Differentiation of Ib from Ia in Group I is a function of another dichotomous variable, *limited opening*. This variable is positive when the maximum pain-free unassisted jaw opening, corrected for vertical overbite, is less than 40 mm. *Disc click* is fundamental to the diagnosis of a displaced disc (Group II) as defined by the RDC/TMD. For this dichotomous variable to be positive, the disc click must occur during a minimum of 2 out of 3 openings/closings of the jaw, either reciprocally (both opening and closing) or during one such movement plus during an excursive movement. Diagnosis of TMJ pain is fundamental to the diagnosis

Table 2 Intersite Reliability (n = 180, Three Examinations per Participant)

Diagnosis	Prevalence based on detection rates*	GEE kappa for diagnostic agreement	95% CI for GEE kappa	Percent agreement†	ICC for diagnostic agreement	95% CI for ICC (by bootstrap)
Any Group I	0.63	0.84	0.74 – 0.94	93	0.84	0.77 – 0.91
Ia	0.41	0.62	0.53 – 0.72	81	0.62	0.53 – 0.71
Ib	0.22	0.58	0.41 – 0.74	85	0.58	0.46 – 0.68
Any Group II	0.36	0.60	0.48 – 0.72	82	0.60	0.54 – 0.67
IIa	0.34	0.63	0.51 – 0.75	84	0.64	0.57 – 0.70
IIb	0.01	0.62	0.00 – 1.00	99	0.60	0.00 – 0.87
IIc	0.02	0.31	0.00 – 0.86	97	0.30	0.00 – 0.59
Any joint pain (IIIa or IIIb)	0.23	0.59	0.45 – 0.73	85	0.59	0.51 – 0.67
Any arthrosis (IIIb or IIIc)	0.07	0.39	0.16 – 0.63	91	0.38	0.25 – 0.51
IIIa	0.19	0.55	0.39 – 0.71	86	0.55	0.45 – 0.64
IIIb	0.04	0.40	0.06 – 0.74	95	0.38	0.22 – 0.56
IIIc	0.03	0.43	0.09 – 0.78	96	0.42	0.22 – 0.60

*Prevalence based on the probability of any examiner making a positive diagnosis within this study sample.
†Percent agreement is computed as an overall agreement based on three pairings for Group I diagnoses, and six pairings for the side-specific diagnoses (Groups II and III).

Table 3 Intrasite Reliability (n = 705, Two Examinations per Participant)

Diagnosis	Prevalence based on detection rates*	Prevalence of reference standard diagnoses†	GEE kappa for diagnostic agreement between TE & CE	95% CI for GEE kappa	Percent agreement‡
Any Group I	0.61	495/705 = 0.70	0.82	0.77 – 0.87	91
Ia	0.25	210/705 = 0.30	0.60	0.51 – 0.69	85
Ib	0.36	285/705 = 0.40	0.70	0.64 – 0.76	86
Any Group II	0.24	898/1410 = 0.64	0.58	0.49 – 0.66	84
IIa	0.21	532/1410 = 0.38	0.60	0.52 – 0.69	87
IIb	0.02	91/1410 = 0.06	0.51	0.20 – 0.82	98
IIc	0.01	275/1410 = 0.20	0.13	-0.10 – 0.36	98
Any joint pain (IIIa or IIIb)	0.27	689/1410 = 0.49	0.55	0.47 – 0.62	82
Any arthrosis (IIIb or IIIc)	0.05	343/1410 = 0.24	0.33	0.18 – 0.48	94
IIIa	0.24	466/1410 = 0.33	0.52	0.44 – 0.60	82
IIIb	0.03	223/1410 = 0.16	0.36	0.17 – 0.56	97
IIIc	0.02	120/1410 = 0.09	0.28	0.07 – 0.49	97

*Prevalence based on the probability of either the CE or the TE making a positive diagnosis within this study sample.
†Prevalence based on the reference-standard criterion diagnoses (see the first article in this series[14]). For Group I, there is only one diagnosis per participant and the rate denominator is 705. For Groups II and III, there are two joints per participant and the diagnosis rate denominator is 1,410.
‡Percent agreement is computed as an overall agreement based on one pairing for Group I diagnoses, and two pairings for the side-specific diagnoses (Groups II and III).

arthralgia (Group IIIa). This dichotomous item is positive if there is pain to palpation of either the lateral pole of the joint or the posterior attachment of the joint. Joint palpation was performed with 1 pound of pressure as specified by the RDC/TMD. The absence or presence of *coarse crepitus* is used diagnostically to differentiate IIIb/IIIc from IIIa.

Statistical Procedures

Prevalence of diagnoses in the study sample and overall percent agreement for diagnoses were computed using the Proc Freq procedure (SAS Institute).

A generalized estimate equations (GEE) procedure has been described by Williamson et al[15] for kappa estimates. Kappa point estimates were computed with this procedure as well as variance estimates for calculating 95% CIs. This mathematical model yields estimates adjusted not only for chance but also for correlated data between the right and left joints within the same participant. In addition, this GEE kappa procedure allows for assessment of agreement between multiple examiners. Based on this model, estimates of agreement over the six examiners were computed for the intersite calibration study (n = 180) in Table 2. The GEE kappa procedure was also used for estimating agreement between two examiners at each site (one CE and the TE) for the intrasite reliability assessment (n = 705) presented in Table 3, and for

Table 4 Intersite Data Collection: Reliability of Selected Exam Items with Corresponding Diagnostic Algorithm Reliability

	Intersite data collection (n = 180)*				Reliability of algorithmic RDC/TMD diagnoses	
Dichotomous examination items	Examiner-based prevalence rate	GEE kappa for examination item agreement	95% CI for GEE kappa	Percent agreement	Diagnosis	GEE kappa
Muscle pain	0.76	0.75	0.58 – 0.92	91	Group I myofascial pain	0.84
Limited opening	0.29	0.58	0.45 – 0.71	82	Ib	0.58
Disc click	0.34	0.63	0.51 – 0.75	84	IIa	0.63
Joint pain	0.32	0.42	0.29 – 0.55	74	IIIa	0.55
Coarse crepitus	0.12	0.53	0.31 – 0.75	90	Any arthrosis (IIIb or IIIc)	0.39

*Three examinations per participant.

Table 5 Intrasite Data Collection: Reliability of Selected Exam Items with Corresponding Diagnostic Algorithm Reliability

	Intrasite data collection (n = 705)*				Reliability of algorithmic RDC/TMD diagnoses	
Dichotomous examination items	Examiner-based prevalence rate	GEE kappa for examination item agreement	95% CI for GEE kappa	Percent agreement	Diagnosis	GEE kappa
Muscle pain	0.66	0.77	0.71 – 0.83	90	Group I myofascial pain	0.82
Limited opening	0.45	0.76	0.71 – 0.81	88	Ib	0.70
Disc click	0.20	0.61	0.51 – 0.70	87	IIa	0.60
Joint pain	0.30	0.41	0.33 – 0.48	75	IIIa	0.52
Coarse crepitus	0.07	0.53	0.37 – 0.69	94	Any arthrosis (IIIb or IIIc)	0.33

*Two examinations per participant.

the selected dichotomous clinical measures in Tables 4 and 5.

Reliability can also be estimated by the ICC using a random-effects analysis of variance (ANOVA) model described by Shrout and Fleiss.[16] To compare GEE kappa results to ICC results, the ANOVA procedure was applied to the intersite data from the six participating examiners. The bootstrap method was employed to take into account the correlated data between sides within the same participant and compute the 95% CIs for the ICCs.

For the formal intersite calibration studies, the consensus (reference standard) methodology was not applied when establishing the prevalence of diagnoses. In this context, diagnostic prevalence can then be understood as the probability of any examiner making a positive diagnosis. Percent agreement as reported in the tables was computed as follows: when, for example, three examiners saw a given participant for the intersite calibrations, this allowed for three diagnostic pairings (1 versus 2, 1 versus 3, 2 versus 3) to be assessed for Group I diagnoses, and six pairings to be assessed for Group II and Group III diagnoses, because the latter two diagnostic groups are side-specific. Overall percent agreement was the average over all participants examined for a given diagnosis or for individual examination items.

The study used the following guidelines[17] for interpreting either the kappa statistic or the ICC: > 0.75 denotes excellent reproducibility; 0.4 to 0.75 demonstrates fair to good reproducibility; and < 0.4 expresses poor reproducibility.

Results

Reliability of Diagnostic Algorithms

Using the guidelines of Fleiss et al,[17] the reliability of the RDC/TMD Axis I diagnoses was excellent ($\kappa > 0.75$) only for "any Group I" (Ia or Ib) in the intersite and intrasite assessments. Intersite reliability of Ia, Ib, IIa, IIIa, and "any joint pain" (IIIa or IIIb) was consistently good ($\kappa = 0.55$ to 0.63; Table 2). All but one of the lower confidence limits for these estimates also showed values considered to be fair to good (≥ 0.41). The intrasite reliability estimates for these same diagnoses were similar, with $\kappa = 0.52$ to 0.70 (Table 3).

For the less common Axis I diagnoses (ie, IIb, IIc, IIIb, IIIc, and any arthrosis [IIIb or IIIc]), intersite reliability was either poor or at a low level of acceptability (κ = 0.31 to 0.43), with only IIb found to have good reliability (κ = 0.62) (Table 2). In addition, the CIs for the less common diagnoses suggest, in most cases, considerable uncertainty. Intrasite reliability for the less common diagnoses was again comparable to intersite reliability for four of the five diagnostic categories, with κ = 0.28 to 0.51. The exception in the intrasite data was IIc, which was associated with a very low detection rate and with κ = 0.13 (Table 3).

The GEE kappa method for computing reliability estimates was equivalent to the ICC from the random-effects ANOVA method, differing by no more than 0.02 for any pairing of estimates in Table 2. Percent agreement was no lower than 81% in both the intersite and intrasite assessments (Tables 2 and 3). Percent agreement for four dichotomous key clinical measures in both assessments was no lower than 82% (Tables 4 and 5). However, agreement on joint pain was 74% and 75% in the intersite and intrasite assessments, respectively (Tables 4 and 5).

Relationship Between the Reliability of Diagnostic Clinical Measures and the Reliability of Diagnoses

As noted above, the reliability of algorithmic diagnostic trees such as those of the RDC/TMD is expected to be similar to the reliability of the clinical measures that make up the trees. The reliability of *muscle pain* was estimated at κ = 0.75 (intersite data) to κ = 0.77 (intrasite data), and Group I diagnostic reliability was seen to be similar at κ = 0.82–0.84 (Tables 4 and 5). For *limited opening*, which is used to differentiate a diagnosis of Ib from Ia, reliability was estimated at κ = 0.58 (intersite data) and κ = 0.76 (intrasite data). The diagnostic reliability for Ib was κ = 0.58 (intersite data) and κ = 0.70 (intrasite data). The diagnostic *disc click* as specified by the RDC/TMD had a reliability of κ = 0.63 (intersite data) and κ = 0.61 (intrasite data), and the diagnostic reliability for Group II disc displacement with reduction (IIa) was κ = 0.60-0.63. Diagnosis of temporomandibulal *joint pain* had a reliability of κ = 0.42 (intersite data) and κ = 0.41 (intrasite data). The corresponding estimate of reliability for the diagnosis of arthralgia (IIIa) was κ = 0.52–0.55. In both the intersite and intrasite assessments, the finding of *coarse crepitus* used for differentiating IIIb (osteoarthritis) or IIIc (osteoarthrosis) from IIIa did not vary (κ = 0.53), and the reliability for the diagnosis of any arthrosis (IIIb or IIIc) varied little, from κ = 0.33–0.39.

Low Examiner Detection Rates for the Less Common Diagnoses

The problem of low examiner detection rates was experienced when the published RDC/TMD Axis I protocol was used for detection of the less common diagnoses (IIb, IIc, IIIb, IIIc). Detection rates averaged around 2% of the study sample (see Table 3), corresponding to approximately four participants in the intersite sample, and 15 participants in the intrasite sample, both of which are too few for stable reliability estimates.

Discussion

This study evaluated comprehensively the interexaminer reliability of RDC/TMD Axis I diagnoses. It provided reliability estimates for less common TMD disc displacement or osteoarthrosis disorders, which had previously not been available. When published guidelines[17] were used for the interpretation of the magnitude of reliability coefficients, the reliability of RDC/TMD Axis I diagnoses was excellent (κ > 0.75) only for one diagnosis, the combination Group I (Ia or Ib). Reliability estimates for Ia, Ib, IIa, IIb, IIIa, and "any joint pain" (IIIa or IIIb) were generally good, whereas the reliability of IIc, IIIb, IIIc, and any arthrosis (IIIb or IIIc) was poor (κ < 0.4) to marginally fair (κ ≤ 0.43). The reliability of common diagnoses such as myofascial pain with limited opening (Ib), disc displacement with reduction (IIa), and arthralgia (IIIa) varied little from the reliability of single key examination measures: interincisal opening, disc click, and joint pain to palpation, respectively. The difference in kappa ranged from 0 to 0.13. This demonstrates the statistical influence of single variables for their capacity to predict and limit the reliability of the diagnostic algorithms.

Comparisons of Reliability Estimates of This Study with Past Studies

Although a number of studies have been published that assessed TMD signs and symptoms operationalized according to the RDC/TMD specifications,[10,18–20] only a few studies have reported reliability findings for the actual diagnoses. Lausten et al[21] reported kappa values for myofascial pain with (0.496) and without limited opening (0.681), disc

displacement with reduction (0.621), and arthralgia (0.405). These reliability coefficients are slightly above the present study's estimates for myofascial pain without limited opening, similar to the results for disc displacement with reduction, and lower than the findings for myofascial pain with limited opening and arthralgia. John and Zwijnenburg have reported kappa values between 0.32 and 0.51 for disc displacements with reduction.[7] This is lower than the findings observed in the present study. In youths aged 12 to 18 years, Wahlund et al found good to excellent reliability for each of the RDC/TMD major groupings with all reliability coefficients > 0.78 for RDC/TMD Groups I, II, and III.[4] The present study showed this level of reliability only for Group I (muscle pain). Only three single-site studies have been published that were designed to examine the entire set of RDC/TMD Axis I diagnoses. All were included among the 10 international consortium sites reported on by John et al.[9] In one single-site study, Lobbezoo et al found fair to good reliability coefficients, ie, ICC values between 0.4 and 0.75 for myofascial pain diagnoses, disc displacement with reduction as well as without reduction and limited opening, and arthralgia. The diagnosis of osteoarthrosis reached excellent reliability with ICC = 0.79. This latter estimate of reliability, approximately twice as high as that observed in the present study, may be related to their low sample size (which would increase susceptibility to chance), and/or by differences in case severity.[11] Similar results were reported from another single-site study by Schmitter et al.[12] These authors found fair to good reliability for common RDC/TMD diagnoses such as myofascial pain, disc displacement with reduction, and arthralgia. Only disc displacement without reduction without limited opening showed poor reliability. The third single-site study to publish did not have the possibility of assessing four of the RDC/TMD diagnoses due to the low prevalence of these diagnoses in their study sample, but their results for Ia and Ib demonstrated higher reliability by 0.18 than in the present study.[10]

When the present study is compared to the international multicenter study,[9] which was the most comprehensive reliability study to date, the present data are equal or better. For half of the diagnostic categories, reliability coefficients observed in the present study were similar to the multicenter study, ie, within a 0.10 range. For the other half of the diagnoses, reliability coefficients in the present study were higher than in the international study. The international study was not able to determine the reliability of diagnoses such as disc displacement without reduction and without limited opening, osteoarthritis, and osteoarthrosis.

Because the current estimates of reliability are based on a large multicenter sample, the authors hold that these estimates of reliability are both credible and generalizable to other populations in which the prevalence of diagnoses is comparable. This is the first reliability study for which the diagnostic prevalence of each diagnosis was established by reference standard methodology (intrasite data, Table 3). While such criterion procedures may not be feasible for most future reliability studies, investigators should clearly delineate their recruitment goals and techniques as well as all available information on their study participant diagnoses. This requirement is already set forth in the Consolidated Standards for Reporting Trials (CONSORT) recommendations for clinical trials,[22,23] and is recommended by the Standards for Reporting of Diagnostic Accuracy (STARD) initiative for diagnostic accuracy studies.[24]

Additional Interpretations of the Results

Classification tree methods. Classification trees are widely used in applied fields such as medicine because they lend themselves to graphical displays that are relatively easy to interpret. Such diagnostic trees are composed of nodes that categorize or divide participants based on a "split" condition. By observing the conditions that are satisfied between the initial node and the terminal node, one can understand diagnostic characteristics that predict membership in categorical classes such as case or noncase.

The influence of examiner-based detection rates in a participant sample. The split conditions chosen for the RDC/TMD Axis I diagnostic system were designed to have an expected 70% sensitivity and 95% specificity.[1] The intent was to minimize false-positive diagnoses while accepting higher levels of false-negative diagnoses. Diagnostic criteria with moderate sensitivity and high specificity generally make it more difficult to make a positive diagnosis, particularly as the participants' signs and symptoms become milder and less numerous. However, as it becomes more difficult for an examiner to make a positive diagnosis, the examiner-based prevalence of the diagnosis (ie, examiner detection rate) in the study sample decreases and, as a result, both kappa and ICC estimates may fall significantly. It has previously been shown that the magnitude of the reliability coefficients depends on the prevalence of the disorder.[25,26] The prevalence of consensus diagnoses for IIb, IIc, IIIb,

and IIIc in the intrasite sample was 0.06, 0.20, 0.16, and 0.09, respectively. In comparison, the examiner detection rates for these diagnoses were 0.02, 0.01, 0.03, and 0.02, respectively (Table 3). These detection rate estimates pertain to *any* examiner making a positive diagnosis, and are not based on multiple raters having to agree that the diagnosis is present for it to be counted. These diagnostic rates also include both true and false-positive diagnostic renderings. Thus, it is predictable that examiner reliability would be lower for the less common Axis I diagnoses (IIb, IIc, IIIb, and IIIc).

A potential ceiling for some examination items. In 1992, Dworkin and LeResche reported that agreement on joint sounds with auscultation of the joint showed kappa values in the 0.60s whereas joint pain to palpation was somewhat lower, with kappa values in the 0.40s.[1] The present study found almost identical reliability estimates with disc click at κ = 0.61–0.63, and joint pain at κ = 0.41–0.42. These findings suggest that there may be predictable levels of reliability associated with certain signs, and that these observations could be relatively consistent over time, participants, and examiners. The implication of this less-than-excellent clinical measure reliability for the diagnostic algorithms' reliability is reflected in the data in Tables 4 and 5. Kappas for disc click and kappas for agreement on IIa differ by no more than 0.01. Kappas for joint pain predict within a range of 0.11 to 0.13 the kappas for diagnostic agreement on arthralgia.

Limitations of this Study

The validation study that assessed the eight Axis I diagnoses against the reference standard diagnoses, as reported in the third article in this series,[13] was designed to have sufficient statistical power to yield CIs no greater than 0.10 on either side of point estimates measured on a 0.0 to 1.0 scale. However, the intersite reliability assessments were not designed to have such statistical power, given that the total sample size was 180 as compared to 705 participants for the validation study. Thus, the present study is able to report new information as to point estimates for the reliability of the less common Axis I diagnoses (IIb, IIc, IIIb, and IIIc), but, due to low detection rates, the associated CIs indicate considerable uncertainty as to the true magnitude of these estimates.

Conclusions

Diagnoses are, both from conceptual and clinical practice points of view, the preferred way to characterize TMD, and the RDC/TMD protocol is currently the most widely used diagnostic and classification system for Axis I diagnoses. The reliability of RDC/TMD Axis I diagnoses was found to be excellent (κ > 0.75) only for "any Group I" (Ia or Ib). The reliability of Ia, Ib, IIa, IIIa, and "any joint pain" (IIIa or IIIb) was generally good, as demonstrated by kappa values in the range of 0.52 to 0.70, even after taking into consideration the lower confidence limits ($\geq$ 0.39). As for the less common Axis I diagnoses (ie, IIb, IIc, IIIb, IIIc, and any arthrosis [IIIb or IIIc]), reliability point estimates were poor to marginally fair ($\kappa \leq$ 0.43), except for IIb (κ = 0.51–0.62), and all were associated with CIs that suggest considerable uncertainty due to the low detection rate for these diagnoses. Axis I diagnostic classification trees are based on multiple measures, but the reliability of common diagnoses such as myofascial pain with limited opening, disc displacement with reduction, and arthralgia varies little (κ = 0.0–0.13) from the reliability of single key examination measures (interincisal opening, disc click, and joint pain to palpation, respectively).

Acknowledgments

The authors thank the following personnel of the RDC/TMD Validation Project: at the University of Minnesota, Mansur Ahmad, Quentin Anderson, Mary Haugan, Amanda Jackson, Wenjun Kang, and Wei Pan; at the University at Buffalo, Leslie Garfinkel, Yoly Gonzalez, Patricia Jahn, Krishnan Kartha, Sharon Michalovic, and Theresa Speers; and, at the University of Washington, Lars Hollender, Lloyd Mancl, Julie Sage, Kathy Scott, Jeff Sherman, and Earl Sommers. Research supported by NIH/NIDCR U01-DE013331 and N01-DE-22635.

References

1. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord Facial Oral Pain 1992;6:301–355.
2. Smith TW. Measurement in health psychology research. In: Friedman HS, Silver RC. Foundations of Health Psychology. New York: Oxford University, 2007:19–51.
3. 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.
4. Wahlund K, List T, Dworkin SF. Temporomandibular disorders in children and adolescents: Reliability of a questionnaire, clinical examination, and diagnosis. J Orofac Pain 1998;12:42–51.

5. Goulet JP, Clark GT, Flack VF. Reproducibility of examiner performance for muscle and joint palpation in the temporomandibular system following training and calibration. Community Dent Oral Epidemiol 1993;21:72–77.
6. Goulet JP, Clark GT, Flack VF, Liu C. The reproducibility of muscle and joint tenderness detection methods and maximum mandibular movement measurement for the temporomandibular system. J Orofac Pain 1998;12:17–26.
7. John MT, Zwijnenburg AJ. Interobserver variability in assessment of signs of TMD. Int J Prosthodont 2001;14: 265–270.
8. Conti PC, dos Santos CN, Lauris JR. Interexaminer agreement for muscle palpation procedures: The efficacy of a calibration program. Cranio 2002;20:289–294.
9. John MT, Dworkin SF, Mancl LA. Reliability of clinical temporomandibular disorder diagnoses. Pain 2005;118: 61–69.
10. List T, John MT, Dworkin SF, Svensson P. Recalibration improves inter-examiner reliability of TMD examination. Acta Odontol Scand 2006;64:146–152.
11. Lobbezoo F, van Selms MK, John MT, et al. Use of the research diagnostic criteria for temporomandibular disorders for multinational research: Translation efforts and reliability assessments in The Netherlands. J Orofac Pain 2005;19:301–308.
12. Schmitter M, Ohlmann B, John MT, Hirsch C, Rammelsberg P. Research diagnostic criteria for temporomandibular disorders: A calibration and reliability study. Cranio 2005;23:212–218.
13. Truelove EL, Pan W, Look JO, et al. The research diagnostic criteria for temporomandibular disorders. III. Validity of axis I diagnoses. J Orofac Pain 2010;24:35–47.
14. Schiffman EL, Truelove EL, Ohrbach R, et al. The research diagnostic criteria for temporomandibular disorders. I. Overview and methodology for assessment of validity. J Orofac Pain 2010;24:7–24.
15. Williamson JM, Lipsitz SR, Manatunga AK. Modeling kappa for measuring dependent categorical agreement data. Biostatistics 2000;1:191–202.
16. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420–428.
17. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. Hoboken, NJ: Wiley-Interscience, 2003.
18. John MT, Hirsch C, Reiber T, Dworkin SF. Translating the research diagnostic criteria for temporomandibular disorders into German: Evaluation of content and process. J Orofac Pain 2006;20:43–52.
19. Khoo S, Yap AJ, Chan YH, Bulgiba AM. Translating the research diagnostic criteria for temporomandibular disorders into Malay: Evaluation of content and process. J Orofac Pain 2008;22:131–138.
20. Pehling J, Schiffman E, Look J, Shaefer J, Lenton P, Fricton J. Interexaminer reliability and clinical validity of the temporomandibular index: A new outcome measure for temporomandibular disorders. J Orofac Pain 2002;16: 296–304.
21. Lausten LL, Glaros AG, Williams K. Inter-examiner reliability of physical assessment methods for assessing temporomandibular disorders. Gen Dent 2004;52:509–513.
22. Begg C, Cho M, Eastwood S,et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637–639.
23. Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357:1191–1194.
24. Bossuyt PM, Reitsma JB, Bruns DE, et al. Standards for reporting of diagnostic accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 2003;49:1–6.
25. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551–558.
26. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–549.