Published in final edited form as:

J Oral Rehabil. 2010 October ; 37(10): 744–759. doi:10.1111/j.1365-2842.2010.02121.x.

Reliability and Validity of Axis I of the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) with Proposed Revisions

John O. Look, Eric L. Schiffman, Edmond L. Truelove, and Mansur Ahmad

Keywords

Temporomandibular muscle and joint disorders; TMD; validity; reliability; research diagnostic criteria

Introduction

The most successful diagnostic protocol for temporomandibular muscle and joint disorders (TMD) is the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD).¹ This protocol is used internationally having been translated into more than 20 languages (International RDC/TMD Consortium Network <www.rdc-tmdinternational.org>). The RDC/TMD incorporates a dual system for assessment of TMD with regard to Axis I physical diagnoses and Axis II psychological status and pain-related disability. Because both the content validity and the construct validity of the RDC/TMD are generally accepted, much research on TMD pain and dysfunction has been performed using this diagnostic protocol. Although the original form of the RDC/TMD published in 1992 has met with broad acceptance by the TMD research community, it was never intended to be an end product but rather a work-in-progress that would be tested and modified as found to be necessary.¹

Given that a comprehensive characterization of the reliability and criterion validity of the RDC/TMD had never before been accomplished, the National Institute of Dental and Craniofacial Research (NIDCR) funded in 2001 the most definitive research to date on the RDC/TMD as a U01 project entitled, “Research Diagnostic Criteria: Reliability and Validity” (referred to hereafter in this paper as the Validation Project). With the U01 research designation, NIDCR was directly involved in the conduct of the study by establishing an Advisory Panel to oversee the project. This panel consisted of 12 experts who represented each of the pertinent clinical and basic science areas. The results of the Validation Project were first presented at a pre-session workshop of the Toronto general session of the International Association of Dental Research (IADR) on July 2, 2008. This meeting entitled, “Validation Studies of the RDC/TMD: Progress toward Version 2,” was sponsored by the International RDC/TMD Consortium Network. Building on positive feedback from this workshop, this paper is intended to complement all that has taken place in terms of discussion and international consensus. Presented here are summaries of the Validation Project Axis I presentations at the IADR Toronto meeting as well as solicited critiques of these presentations.

Reliability of RDC/TMD Axis I diagnoses based on clinical signs and symptoms

The RDC/TMD Axis I protocol is a standardized series of diagnostic tests based on clinical signs and symptoms. Diagnostic algorithms using different combinations of clinical and questionnaire measures are used to differentiate 8 RDC/TMD-defined Axis I diagnoses for TMD. These diagnoses include myofascial pain (Ia), myofascial pain with limited opening (Ib), disc displacement with reduction (IIa), disc displacement without reduction with limited opening (IIb), disc displacement without reduction without limited opening (IIc), arthralgia (IIIa), osteoarthritis (IIIb), and osteoarthritis (IIIc). The reliability of a clinical assessment is the measure of its consistency when it is performed on the same subject by multiple examiners (inter-rater reliability), or when a single examiner performs the diagnostic protocol repeatedly on the same subject (intra-rater). Although reliability (reproducibility) is conceptually different from validity (accuracy), these two characteristics may be in one sense connected; it has been suggested that the reliability of a diagnostic instrument sets the upper limit for its validity.²

Following the formation of the International RCD/TMD Consortium Network, reliability testing on the RDC/TMD Axis I diagnoses was conducted at 10 sites internationally with an overall total of 30 examiners and 230 participants.³ This initiative provided good heterogeneity with respect to examiners and subjects, but the individual studies were too small to allow for assessment of the influence of chance on the estimates of reliability. To that extent, these point estimates of reliability remained in question.

Methods for testing the reliability of RDC/TMD Axis I diagnoses based on the published RDC/TMD diagnostic protocol

The reliability assessment of the RDC/TMD Axis I diagnoses that was conducted as a part of the Validation Project has been described in detail elsewhere.⁴ Reliability of a diagnostic protocol is a function of: 1) the reliability of the tests that are used to make the diagnosis, 2) the training of the examiner to perform these tests, and 3) the characteristics of the subjects on whom the tests are performed. With regard to item 3, if the test diagnoses have a low prevalence in the subject sample, or if subjects are selected whose clinical signs are minimal, reliability estimates will generally be lower. Subject selection for the reliability component of the Validation Project was designed to parallel subject selection required for rigorous testing of the validity of the RDC/TMD. For the latter, putative case status was assigned to individuals who reported minimum or mild TMD symptoms. It is for this reason that cases and controls for reliability testing were selected based on the presence or absence of TMD, but irrespective of the severity of the TMD condition in the cases. Furthermore, no attempt was made to selectively enrich this study sample for the less common diagnoses (IIb, IIC, IIIb, and IIIc). Thus, this study design was not intended to produce the highest possible estimates of reliability, but rather to deliver point estimates of reliability that would be pertinent to the rigorous conditions required for the validity testing of the RDC/TMD.

A total of 9 clinicians served as the examiners for the RDC/TMD Validation Project, including 2 Criterion Examiners (CEs) and 1 Test Examiner (TE) at each of three study sites: University at Buffalo, University of Minnesota and University of Washington. The CEs performed the criterion data collections that led to establishing the reference (gold) standard diagnoses. The TEs, on the other hand, represented the RDC/TMD at its best, and they performed only the clinical tests specified by the RDC/TMD. All 6 CEs were TMD and orofacial pain experts with between 12 to 38 years of experience in research and treatment of TMD. The 3 TEs were dental hygienists who were trained and calibrated to perform the RDC/TMD examination.

Inter-rater reliability for the published RDC/TMD examination protocol was assessed throughout the Validation Project. One baseline and four follow-up sessions were conducted annually that involved examiners from the three study sites (intersite calibration). In addition, inter-rater reliability assessment was performed continually within sites (intrasite calibration) as will be described below. All five intersite calibrations were conducted at Minnesota, and they included all 3 TEs, but just 1 CE representing each study site. At each session, 36 calibration subjects each underwent 3 examinations that were strictly based on the RDC/TMD protocol. Typically, the study sample included 3 normal subjects and 33 TMD cases. Given the requirement that calibration subjects should resemble as much as possible the subjects in whom validation testing of the RDC/TMD was performed, all participants were recruited using the same inclusion and exclusion criteria as employed for the formal validation study. Most of the 180 subjects (total of 540 exams) seen during the five annual calibration sessions were drawn from the validation study sample during the years following completion of their data collection.

Inter-rater reliability for the published RDC/TMD examination protocol was also performed within each study site, and this assessment employed the entire validation subject sample. The validation subjects were drawn from a total of 1244 candidates that were screened across the 3 study sites over the period of August 2003 to September 2006. Of these, 732 met all study requirements and were consenting. Eight of these subjects still had incomplete assessments at study closure, and could not be included in the final analyses. For five of the remaining 724, evidence lacked for their clear classification as a case or control, and they were excluded from the analyses. An additional 14 subjects were found to have co-morbid conditions that are not recommended for inclusion in the initial validation of a test protocol.⁵ The co-morbidities included chondromatosis ($n = 2$), fibromyalgia ($n = 9$) and other rheumatologic disorders ($n = 3$). Therefore, the final validation study sample was 705, including 614 cases and 91 controls.⁶ Apart from their diagnostic ambiguities and co-morbidities, the 19 excluded cases with complete data did not differ from the 614 included cases relative to study covariates such as gender distribution, mean age, number of concurrent TMD diagnoses, duration of TMD symptoms,¹ characteristic pain intensity,^{1, 7} pain-related disability,^{1, 7} nonspecific physical symptoms,^{1, 8, 9} and depression^{1, 8, 9} (all P values ≥ 0.12).

As noted above, one CE from each site was absent from the annual intersite calibrations. However, intrasite procedures were established to monitor continually the inter-rater reliability of the three examiners at each study site. This was made possible since one of the CEs and the TE each performed examinations on the same validation subject the same day while blinded to the other's findings. The CE examination was a much-expanded set of diagnostic tests to establish the criterion diagnoses. These criterion tests are summarized in RDC/TMD validity section (fourth summary) of this paper, and are described in detail elsewhere.⁶ Interspersed among these tests were all the RDC/TMD exam items that could then be abstracted out of the criterion data collection and mathematically submitted to the RDC/TMD diagnostic algorithms, the same as were also the exam data collected by the TE. At each study site, the diagnostic reliability of the TE was compared to both CEs since, by design, the CEs alternately performed the second criterion data collection the same day that the TE examination was done. Intrasite reliability monitoring was thus performed with 705 subjects (data from 1410 exams). In addition to the eight RDC/TMD diagnoses, reliability was also evaluated for four groupings of the diagnoses: any Group I diagnosis (Ia or Ib); any Group II disc displacement diagnosis (IIa, IIb or IIc); any joint pain diagnosis (IIIa or IIIb); and any degenerative joint disease (IIIb or IIIc). A generalized estimating equations (GEE) procedure was employed to compute kappa point estimates for inter-rater agreement across multiple examiners as well as 95% confidence intervals that were adjusted for side-to-side correlations within subjects.¹⁰

Reliability results for RDC/TMD Axis I diagnoses

Study guidelines for classifying kappa coefficients were those of Fleiss et al.: > 0.75 indicates excellent reproducibility; 0.4 to 0.75 shows fair to good reproducibility; and < 0.4 is poor reproducibility.¹¹ Based on these guidelines, the reliability of the RDC/TMD Axis I diagnoses was excellent ($k > 0.75$) only for one combination diagnosis, “any Group I” (Ia or Ib), in both the intersite and intrasite assessments. Intersite reliability of the more common diagnoses, Ia, Ib, IIa, IIIa, and “any joint pain” (IIIa or IIIb), was consistently good ($k = 0.55$ to 0.63). The intrasite reliability estimates for these same diagnoses were similar with $k = 0.52$ to 0.70 . For the less common Axis I diagnoses (i.e., IIb, IIc, IIIb, IIIc, and the combined diagnosis for degenerative joint disease [IIIb or IIIc]), intersite and intrasite reliability was mostly poor or at a low level of acceptability ($k = 0.13$ to 0.43). IIb alone was found to have fair to good reliability ($k = 0.62$ intersite, $k = 0.51$ intrasite). Because of the large intrasite sample size (the entire formal validation sample, $n = 705$), the total width of the confidence intervals for the reliability estimates relative to the more common diagnoses was optimally narrow (< 0.20), and all had lower confidence bounds falling between 0.44 and 0.77 . However, the typically low prevalence of the less common diagnoses in the study samples yielded confidence intervals that were unacceptably wide. The point estimates of reliability derived from the Validation Project showed good parity with the results of the international multi-center study³ that was the most comprehensive reliability study prior to the Validation Project. For half of the diagnostic categories, our new reliability coefficients were similar to the multi-center study, being within a 0.10 range. The remaining reliability estimates were higher than for the international study.

Conclusions on the reliability of RDC/TMD Axis I diagnoses based on clinical signs

To employ the RDC/TMD-specified clinical tests as a stand-alone criterion diagnosis for TMD would be unacceptably susceptible to diagnostic misclassification. While the more common diagnoses may show good examiner reliability, some measurement variability (lack of agreement) is clearly present, even when these procedures are performed by well-trained examiners.

Reliability of radiographic interpretations used for RDC/TMD Axis I diagnoses

The published RDC/TMD Axis I protocol was primarily based on clinical signs and symptoms. When accessible, radiographic imaging was also recommended to help differentiate the three disc displacement diagnoses in Group II, and the diagnoses of arthralgia, osteoarthritis and osteoarthrosis that make up Group III. The published protocol described briefly the use of magnetic resonance imaging (MRI) and arthrography for diagnosis of disc displacement, as well as the use of tomography for detection of osseous degenerative changes associated with diagnoses of osteoarthritis or osteoarthrosis.¹ However, no criteria for interpretation of the imaging were specified. At the time of the Validation Project, state-of-the-art methods for temporomandibular joint (TMJ) imaging had progressed to include computed tomography (CT) for diagnosis of osseous degenerative changes, with MRI still the standard for detecting disc displacements. Because panoramic radiography had also been recommended for screening of intra-articular hard tissue TMJ pathology,^{12, 13} this diagnostic method was evaluated as well in the Validation Project. To support and enhance the validity of the reference standard protocol, comprehensive criteria were compiled for image acquisition and analysis of CT, MRI and panoramic radiographs, all of which have been described in detail elsewhere.¹⁴ This was followed by training and reliability assessment of the Validation Project radiologists for the analysis of these images.

Methods for scoring radiographic diagnoses

To ensure site-to-site uniformity in the interpretation of radiographic imaging, certain decision rules were specified. If there were different findings possible based on different slices of CT or MRI, the “worst case” rule was applied for the diagnostic decision. For example, if only one slice clearly demonstrated a disc displacement when the others appeared to be normal, the diagnosis was “disc displacement.” Radiographic interpretations were scored as categorical variables. Osseous status was characterized as normal, indeterminate, or frank degenerative joint changes. Disc status was scored as normal, anterior disc displacement with reduction, anterior disc displacement without reduction, disc not visible, or indeterminate.¹⁴

During the Validation Project, one baseline and three additional calibration sessions were conducted to assess the reliability of radiographic interpretations. The radiologists independently performed their interpretations of digital images. Each session employed a minimum of 20 sets of panoramic radiographs and 25 sets each of CT and MRI images. Panoramic, CT and MR images were randomly ordered with respect to normal status versus osseous degenerative changes. Similar random ordering was applied to MR images for evaluation of disc position. For the results reported in this summary, radiographic findings were grouped with hard tissues coded as frank degenerative joint change versus a normal or indeterminate status. Disc position was categorized as displaced versus non-displaced. The disc categories of not visible, indeterminate, or other ratings were excluded. Reliability was estimated using the simple kappa (k) statistic since there was no issue of side-to-side correlation. Imaging views were selected from just one side of a subject, not both sides. The bootstrap method was employed to compute 95% confidence intervals for multiple examiners,¹⁵ and reliability estimates were interpreted according to the guidelines of Fleiss et al.¹¹

Overall reliability and validity results for radiographic diagnoses

Using panoramic radiographs for the diagnosis of degenerative joint change (osteoarthritis, OA), inter-rater reliability was poor at $k = 0.16$ (CI: 0.04 to 0.27). With the MRI, the diagnosis of hard tissue status showed fair reliability at $k = 0.47$ (CI: 0.33 to 0.58). Reliability for OA improved to $k = 0.71$ (CI: 0.63 to 0.79) with the use of CT images. Using MRI for the analysis of soft tissue components of the joint, the reliability of a diagnosis for any disc displacement was excellent with $k = 0.84$ (CI: 0.76 to 0.91). The diagnosis of disc displacement with reduction showed $k = 0.78$ (CI: 0.68 to 0.86), and disc displacement without reduction was at $k = 0.94$ (CI: 0.89 to 0.98).

Using the CT-based diagnosis of OA as the reference standard, the sensitivity and specificity of panoramic radiography and MRI were evaluated. The sensitivity of any diagnostic instrument is the probability that it will show a positive test result when the disorder is present as per the reference standard, and its specificity is the probability of a negative result when the disorder is absent as per the reference standard. Based on these criteria, panoramic radiography had very low sensitivity of 0.26 for OA, but excellent specificity at 0.99. MRI imaging showed sensitivity of 0.59 for OA with specificity of 0.98.

Conclusions on the reliability of radiographic diagnoses

Using MRI for diagnosis of soft tissue disorders and CT scans for hard tissue, reliability was good, even approaching the threshold for excellence. Diagnosis of disc displacements was good to excellent depending on the diagnosis. However, the extent of discordant interpretations would suggest that radiographic diagnoses should not be considered to be stand-alone gold standards for TMJ intra-articular disorders.

Reliability of self-report data used for RDC/TMD Axis I diagnoses

Patient self-reports relative to three questions are an essential component of RDC/TMD Axis I diagnoses.¹ These questions are included in the published RDC/TMD Axis I Questionnaire: Question #3 “Have you had pain in the face, jaw, temple, in front of the ear, or in the ear in the past month?” Question #14a “Have you ever had your jaw lock or catch so that it won’t open all the way?” Question 14b “Was this limitation in jaw opening severe enough to interfere with your ability to eat? Test-retest reliability assessment was performed using a subset of 70 subjects who presented for Axis I assessment at the University at Buffalo and the University of Washington. This test-retest evaluation included the entire RDC/TMD Axis I Questionnaire, the Supplemental History Axis I Questionnaire used for the criterion protocol, and all of the Axis II self-reports. Reliability results for the diagnostic Questions #3, 14a and 14b were excellent with kappa of 0.84, 0.76 and 0.75, respectively. The other test-retest reliability results will be reported in a future publication.

Validity of RDC/TMD Axis I diagnoses based on clinical signs and symptoms

The validity of an index test is the measure to which it correctly classifies the presence or absence of a disorder in an individual when compared to a credible diagnostic reference standard. This is most often expressed as sensitivity and specificity, both measures having been defined above. In addition to evaluating the validity of the 8 RDC/TMD-specified diagnoses for TMD, the 4 combinations of diagnoses noted above for reliability testing were also assessed for their validity: any Group I myofascial pain diagnosis (Ia or Ib), any Group II disc displacement (IIa, IIb or IIc), any Group III joint pain (IIIa or IIIb), and any Group III degenerative joint disease (IIIb or IIIc).

The validation of a diagnostic protocol is indeed a challenge because it requires a credible gold standard criterion against which the test protocol is compared. If there is no objective, incontrovertible, gold standard diagnosis available, the only alternative for evaluation of a diagnostic test protocol is to develop a reference standard that brings together all information pertinent to the disorder under consideration. For TMD, there is no objective biologic test to serve as a gold standard. As reported above, the initial studies in the Validation Project showed that self-reports, clinical measures performed by experts, and radiographic interpretations are all inadequate if used as stand-alone diagnostic methods. Therefore, credible reference standard diagnoses for validation purposes had to be based on a synthesis of patient-reported symptoms (questionnaire responses), assessment of clinical signs, and radiographic evidence. In order to reduce error in the interpretation of a potentially large amount of clinical information, the Validation Project design specified that two TMD experts (i.e., the CEs) would perform independent syntheses of all the available data for each subject. Following that, they would come together for a consensus diagnosis, including a re-examination of the subject if there were any disagreement between their independent assessments. A similar study design was used to establish the reference standards employed in the validation of diagnostic criteria for fibromyalgia.¹⁶

Criterion data collection for establishing credible reference standard diagnoses

A comprehensive list of diagnostic tests was drawn from recommendations of the 1992 RDC/TMD publication, a review of recommendations published in the TMD literature since 1992, tests recommended by the NIDCR Advisory Panel, and recommendations solicited from diverse TMD groups including members of the American Academy of Orofacial Pain. Several tests were also considered based on published diagnostic criteria from the American College of Rheumatology.

The clinical examination for the criterion protocol included all the measures specified in the RDC/TMD. A pre-eminent consideration in this study was for the original RDC/TMD tests to be evaluated side-by-side under identical conditions with the new candidate tests. Then, based on an objective assessment of the relative utility of old and new diagnostic tests, a revised RDC/TMD could be developed. The new tests that were evaluated included joint-play tests including traction, translation and distraction,^{17–19} static and dynamic orthopedic tests,^{17, 20} soft and hard end-feel,²¹ pressure pain threshold algometry,²² the bite test with unilateral and bilateral placement of cotton roles,^{21, 23} and a one-minute clench.²⁴ The list of published tests making up the criterion examination has been described in greater detail elsewhere.⁶ New tests for the criterion protocol included 3–4 pounds digital pressure for the myofascial pain exam, in contrast to the 2 pounds specified by the RDC/TMD, and a novel TMJ palpation technique for arthralgia that is described elsewhere.⁶ One particular new test that turned out to be very informative was as follows: When pain was reported, the subject was asked if this pain was a “familiar pain,” that is, pain similar to or like what had been experienced before as a result of the target condition. Subjects were also asked to indicate any possible sites of referred pain. Additional tests that were employed in the criterion examination included joint loading with opening,¹⁷ the use of a stethoscope to assess for joint noise, and a comprehensive occlusal examination that recorded the number of teeth, overbite, crossbite and midline discrepancy,^{25, 26} assessment of occlusal intercuspal contacts using Shim stock® (Almore International Inc. Portland, Oregon) in maximum intercuspal position (MIP),²⁷ and assessment of centric position (CR) as well as CR to MIP slides.²⁸ Subjects were asked to report any exam-induced joint noise, and this information was recorded. Finally, as reported above, imaging of subjects for the criterion examination included a panoramic radiograph, and bilateral TMJ MRIs and CTs. In all, more than 200 clinical variables were measured as a part of the criterion examination.

The Advisory Panel also vetted a criterion history data collection to be used along with the published RDC/TMD History. The Supplemental History Questionnaire⁶ was designed to guide the criterion examiners in their semi-structured history interview. It consisted of 61 questions assessing pain in jaw muscles, the TMJ, the ear, and the temple, TMJ noise and locking, perceived occlusal changes, and tension-type headache as defined by the criteria of the International Headache Society.²⁹

Minimization of circularity in validity assessment

Circularity in a validation study is a problem that may arise from the design. Among other things, it tends to inflate estimates of validity. It is present when cases and controls are intentionally selected based on characteristics that the test protocol is specifically designed to detect. To minimize such circularity, the following was performed:

- 1) Inclusion criteria for study eligibility differed from RDC/TMD diagnostic criteria by allowing putative case status to individuals who reported a minimum of one of the three cardinal symptoms of TMD: a) jaw pain, b) limited mouth opening or c) TMJ noise. Additionally, the study plan specified recruitment of a minimum 100 consensus-diagnosed TMD cases with minimal signs and symptoms, that is, cases that would normally not receive a TMD diagnosis based on the RDC/TMD-defined criteria. Subjects who denied having any of these symptoms of TMD were enrolled as controls.
- 2) The criterion examination was designed to assess for and diagnose an expanded TMD taxonomy that was independent of the original RDC/TMD taxonomy that is limited to 8 diagnoses. This expanded taxonomy included 6 groupings of TMD with a total of 30 separate diagnoses.⁶ Thus, TMD diagnoses beyond those specified by the RDC/TMD were considered when the consensus diagnoses were rendered.

Circularity also occurs if the reference standard examination protocol resembles too closely the test protocol. If the reference standard and the test protocol were to share no tests in common, this would constitute the cleanest separation. Carrying this principle to an extreme, one could conclude that any muscle or joint palpation, or any range-of-motion measurement, should be absent from the reference standard since these are measures employed in the RDC/TMD protocol. This, however, overlooks the fact that these procedures are standard orthopedic assessments, not only for TMD, but also for multiple domains of medicine. More important, relatively modest differences in the operationalization specified for these procedures can result in radically new diagnostic inferences as will be clear in the final report of this summary.

Along with the totally new orthopedic tests and the newly operationalized tests making up the criterion protocol, the Validation Project design required that the exact diagnostic tests specified by the RDC/TMD would be dispersed within this examination protocol for two reasons: 1) a credible reference standard had to be based on *all* available clinical information, and 2) the expanded set of tests that made up the criterion examination had to be tested concurrently with the RDC/TMD-specified tests in order to make a direct comparison as to their diagnostic utility. Since the validation team could not know in advance the relative weight that might be given to RDC/TMD-specified tests for the establishment of criterion diagnoses, there was a risk that this design was susceptible to a certain amount of circularity. However, as will be clear from the validation results below, the RDC/TMD-based tests did not play an important role in determining the reference standard diagnoses. The final report in this paper describing the revised Axis I diagnostic algorithms will show that the newly operationalized clinical tests were the most sensitive predictors for the reference standard diagnoses. In short, the potential for circularity in the study design did not ultimately prove to be influential in the study results.

Study population for validity assessment of RDC/TMD Axis I diagnoses

An appropriate study population for this project was recruited from the East coast of the United States (Buffalo area), the Midwest (Minnesota), and the West coast (Washington) from August 2003 to September 2006. Twenty-four percent were self-referred subjects or patients referred by local care providers (clinic cases), and 76% were respondents to study flyers and advertisements (community cases). The formal validation was designed to yield confidence limits no greater than 0.10 on either side of the point estimates for sensitivity and specificity. Inclusion and exclusion criteria for all study subjects are described elsewhere.⁶

Demographic measures for this study population included gender, age, education level, and income. Baseline Axis II measures included characteristic pain index,^{1, 7} duration of TMD symptoms,¹ depression,^{1, 8, 9} nonspecific physical symptoms,^{1, 8, 9} and pain-related disability.^{1, 7} Also recorded was current TMD treatment. Details have been published on the measurement instruments employed as well as the full spectrum and severity of TMD signs and symptoms in this study population. The prevalence of Axis II characteristics in the study population was shown to be consistent with literature reports from other population-based and clinical studies.⁶

As explained above, the Validation Project estimated the validity of the RDC/TMD in terms of its sensitivity and specificity assessed in a study sample of 705 subjects consisting of 614 cases and 91 controls, each with established reference standard diagnoses. Overall, the 614 cases presented with a total of 2,202 TMD diagnoses, or an average of 3.6 diagnoses per person.

Validation study data collection methods

The two CEs at each study site alternated between successive subjects for the purpose of completing the initial criterion examination protocol and ordering imaging. At the second visit, the TE performed the RDC/TMD test protocol, and this was followed by the second criterion examination. The TE and the second CE were both blinded to the results of the first CE as well as to each other's findings. Compared to the way the RDC/TMD protocol is typically implemented, there was one change as to how it was performed by the TEs: they were blinded to the subjects' responses to three diagnostic questions employed for the RDC/TMD algorithms. These questions query a history of facial pain, jaw locking and interference with eating. Knowing the responses to these questions could have biased a TE's data collection based on diagnostic suspicion.³⁰ Thus, the data for these questions were collected independently by the study coordinator and added to the data collection after the TE had completed the RDC/TMD examination.

The final diagnostic event was for the reference standard diagnoses to be established by the two CEs who came together with the subject still present to compare their independent findings, re-examine the subject in case of disagreement, and arrive at a consensus based on all available questionnaire, clinical and radiographic information. If either CE disagreed with the radiologist's interpretation, the radiologist also participated in the final review to establish the reference standard diagnoses.

Assessment of measurement variability for the criterion diagnoses

Three additional intersite calibration exercises were programmed during the Validation Project specifically for assessment of the reliability of criterion exams. For each session, one of the CEs from each study site came to the University of Minnesota and, over these three sessions, a total of 26 subjects were assessed by each examiner. They independently performed the criterion protocol and rendered the criterion diagnoses based on all questionnaire, clinical and radiographic data. The three examiners then came together to establish their consensus diagnosis for each subject. This study design allowed for an estimate of diagnostic agreement between their independently rendered criterion diagnoses versus the consensus diagnoses.

A second type of reliability study was performed within each site during the formal validation study. For this, diagnostic agreement was assessed between the second criterion exam and the consensus-based reference standard.

Results on the reliability of the criterion diagnoses

For the intersite ($n = 26$) criterion reliability sessions, individual criterion examinations showed excellent agreement with the consensus diagnoses for 7 of 8 diagnoses ($k = 0.82$ to 0.94). However, the diagnosis of osteoarthritis, with a sample prevalence of just 14%, showed a $k = 0.53$. The overall percent agreement between the examiners and the consensus was 94.4 %.

The intrasite agreement between the second criterion examiner and the consensus was very high with a range of kappa from 0.95 to 0.98. Percent agreement averaged 98.9%. Thus, the error associated with a single criterion exam (as opposed to a consensus between two independent examiners) would be, on average, less than 2%. All statistical computations for kappa estimates were performed using the GEE procedure described by Williamson et al. that provided adjustment for side-to-side correlation within subjects as well as estimates of agreement across multiple examiners.¹⁰

Assessment of measurement error in the test examination

The TEs were well trained to perform the RDC/TMD protocol. However, in order to ascertain measurement error associated with the test examination, the Validation Project design included plans to compare not only the agreement of the TE results with the consensus (the primary study outcome), but also to assess the TE results against both of the CEs' diagnostic findings. It is important to emphasize here that the TEs made no RDC/TMD diagnosis. They simply collected data relative to RDC/TMD-specified clinical tests. These data were then submitted to the published RDC/TMD diagnostic algorithms. All such diagnoses were algorithm-based, not examiner-based. In contrast, the CEs rendered their own criterion diagnoses but, as noted above, the criterion exam included all of the RDC/TMD examination items as part of more than 200 tests that they performed. Thus, it was possible to select out of the criterion data collections the RDC/TMD-specific tests, and submit these data to the RDC/TMD diagnostic algorithms. RDC/TMD algorithm-based diagnoses from the CE data collection were then compared to the consensus findings just like the TEs' results.

Results comparing the test examiners to the criterion examiners for their implementation of the RDC/TMD protocol

This investigation on measurement variability demonstrated nearly total parity between the CEs and the TEs for the performance of the RDC/TMD examination protocol. None of the 24 validation study diagnostic estimates, 12 each for sensitivity and specificity, differed by more than 0.15. Overall percent agreement with the reference standard was 84% for the TEs, and 85% for the CEs.

Assessment of covariates that could statistically influence estimates of sensitivity and specificity

Secondary study analyses were planned for 13 covariates that were measured throughout the formal validation study to assess their influence on sensitivity and specificity estimates. We have seen above that the validation study sample presented with an average of 3.6 diagnoses per TMD case. Thirty-two percent of all subjects had from 0 (normals) to 2 TMD diagnoses, and 68% had 3 to 5 concurrent diagnoses. The effects associated with these two categories were evaluated as were also the effects of appropriate categories for the remaining 12 covariates. Most categories of covariates were differentiated by their median values. For the entire list of test diagnoses, we assessed the effects of age, gender, education,¹ income,¹ number of concurrent TMD diagnoses, duration of TMD symptoms,¹ characteristic pain intensity,^{1, 7} nonspecific physical symptoms,^{1, 8, 9} depression,^{1, 8, 9} pain-related disability,^{1, 7} current or recent treatment for TMD, and study site. For Groups II and Group III diagnoses only, we assessed the effect associated with the right joint being affected as opposed to a left joint disorder. A significant effect was reported if a statistically significant difference ($P < 0.005$ taking into account multiple comparisons) was observed between the defined categories of a covariate for either sensitivity or specificity estimates.

Statistical methods for establishing the validity of the RDC/TMD diagnoses

GEE procedures were employed to account for multiple diagnoses within individuals for Group II and Group III diagnoses. Side-to-side correlations within subjects do not affect point estimates of sensitivity and specificity, but they do affect estimation of the confidence intervals. The effects of the 13 covariates were measured using separate logistic regression models. The primary validation results for this study were the overall estimates of sensitivity and specificity combining the data of the three study sites with no adjustment of point estimates for any of the multiple covariates, and with only the confidence limits adjusted for within-subject correlations. The sensitivity of the RDC/TMD, based on the TE examination

data, was estimated when a given diagnosis was determined to be present by the reference standard, regardless of what other concurrent diagnoses were present. The study sample used to estimate RDC/TMD specificity included all the normal subjects plus all the TMD cases in which a specific diagnosis was not present as per the reference standard.

In the original publication for the RDC/TMD, it was proposed that a valid diagnostic instrument should have sensitivity of ≥ 0.75 and specificity of ≥ 0.95 .¹ These specifications for validity were retained for this study. A single diagnosis or a combination diagnosis was to be declared valid if the point estimates for sensitivity and specificity fell within these bounds, even when the lower confidence intervals did not attain these thresholds.

Primary results of the formal validation assessment of the RDC/TMD

The precision of the validation study was very high. Just one confidence limit differed by as much as 0.10 from the point estimate, that being the upper bound for the sensitivity of IIb (disc displacement without reduction with limited opening). The width of all other upper and lower confidence bounds was less than 0.10. The validity results are published and discussed in detail.³¹ For this summary, we note the following: the only diagnosis that attained target validity was the combined diagnosis of Ia or Ib myofascial pain. Its sensitivity was 0.87, and specificity was 0.98. No single RDC/TMD diagnosis reached both target sensitivity and specificity. Ia was slightly deficient for both sensitivity (0.65) and specificity (0.92). Ib showed on-target sensitivity (0.79), but slightly deficient specificity (0.92). Sensitivity for joint pain (IIIa) was 0.53, and it improved only to 0.57 when assessing for any joint pain (IIIa or IIIb). Specificity of IIIa was below target (0.86), but specificity for the combination IIIa or IIIb did reach target (0.95). For all other intra-articular diagnoses (IIa, IIb, IIc, IIIb, IIIc), sensitivity was poor, while specificity ranged from slightly deficient (IIa only) to on target (≥ 0.95).

Secondary findings from the formal validation assessment of the RDC/TMD

The extent to which covariates affected the test results of the RDC/TMD has been discussed in detail, including statistically significant differences between categories for ubiquitous covariates that include gender, the number of concurrent TMD diagnoses, duration of the TMD, characteristic pain intensity, nonspecific symptoms, and depression.³¹ As an example, assessment of a subject having 0 to 2 concurrent TMD diagnoses showed significantly higher specificity ($P < 0.001$) for pain diagnoses such as Ia, Ib, and IIIa. Their specificity increased from a deficient level when 3 – 5 diagnoses were present to on target (≥ 0.95). However, sensitivity for IIa dropped significantly ($p < 0.001$) from 0.43 when 3 – 5 diagnoses were present to less than half of that coefficient (0.21) when 0 – 2 diagnoses were present.

Conclusions relative to the validity of RDC/TMD Axis I diagnoses

The RDC/TMD is a relatively simple and well-standardized diagnostic protocol that can be recommended for research involving myofascial pain, especially when there is no need to differentiate Ia from Ib. However, for the diagnosis of joint pain, this instrument is less than desirable, and for the diagnosis of intra-articular disorders including both disc displacements and degenerative joint changes, it is unacceptable. While covariates appear to influence the sensitivity and specificity of this examination protocol, more research is needed to understand their effects. The results of the Validation Project are generalizable due to the broad geographic distribution from which validation subjects were recruited. The results are credible in that they are supported by optimally narrow confidence intervals, and they demonstrate that there is a need for revision of the RDC/TMD Axis I algorithms in order to improve diagnostic validity of this instrument.

Proposed revisions of the RDC/TMD Axis I diagnostic algorithms

In the event that the published RDC/TMD procedures would be found to be deficient, NIDCR mandated the development of revised diagnostic examination protocols and diagnostic algorithms for TMD that would also be based on clinical tests for signs and symptoms. The original RDC/TMD diagnostic algorithms are decision and classification tree models.¹ The Group I, Group II and Group III RDC/TMD algorithms consist of nodes defined by a “split condition” that is either satisfied or not satisfied by the results of required clinical tests or self-reports. A node may consist of a single measure or a combination of measures. Beginning with the initial (gateway) node, a diagnostic decision is made for each node, thus leading to the terminal node and the diagnosis. The advantage of these diagnostic structures is that they are readily interpretable and intuitively consistent with theoretical constructs that describe the conditions. Our aim in this data analysis was to retain this classification tree approach that is highly desirable in medicine. Thus, revisions for the algorithms involved selecting the best evidence-based tests that, when assembled in a classification tree, would predict the reference standard diagnoses. As mentioned above, more than 200 tests were simultaneously evaluated, all of these being tests that were performed as part of the criterion examination on the 705 validation subjects.

The following goals were set for the development of the revised diagnostic algorithms: a) they had to be valid in terms of predicting the reference standard diagnoses; b) they had to consist of simple, easy to perform and reliable tests; and c) they had to be parsimonious.

Methods for revised algorithm model building

For this analysis, two data collections were used: the consensus data set (the reference standard diagnoses), and the criterion examination data collection. The latter included all examination data collected by the second criterion examiner, the occlusal data that were collected uniquely by the first criterion examiner, and the questionnaire data collection. This criterion examination and questionnaire data set was randomly divided into two nearly equal parts: the data from 352 subjects were reserved for building new algorithm models (*training or model-building data set*), and the data from the other 353 subjects were set apart for validation (*testing data set*) of the new algorithm models.

Variable selection was performed by building diagnostic algorithms that were derived with a statistical package available at <http://roadrunner.cancer.med.umich.edu/comp/docs/R/rpart.pdf>. The advantage of this package is that it outputs its resultant diagnostic algorithms so that the investigator can assess whether item selection makes clinical sense. This methodology uses techniques referred to as 10-fold cross-validation procedures that have been described by Breiman et al.³² and Hastie et al.³³ All of the more than 200 tests used for the criterion examination were evaluated simultaneously in the model-building data set by this statistical program. The sets of variables were thus selected that best predicted the reference standard diagnoses that had been established for the 352 subjects in the model-building data set. Many clinical provocation tests for muscle or joint pain were not selected as the best predictors including orthopedic tests (algometry, jaw traction, translation and compression), static and dynamic resistance tests, and 1-minute clench. Instead, the selection fell to very simple tests as will be clear from the description of the revised algorithms below. Diagnostic algorithms were thus built and tested using the model-building data set ($n = 352$) before their final validation testing was performed using the testing data set ($n = 353$). Cutoffs for validity of the revised algorithms were sensitivity ≥ 0.75 and specificity ≥ 0.95 , the same as for the RDC/TMD validation.

The reliability of the new algorithms was also tested in a total of 27 newly-recruited subjects at the University of Minnesota (n = 18) and the University of Washington (n = 9). For this study at each site, the TE was trained to perform the revised tests based on the criterion examination specifications. This training was very simple, requiring less than two hours. Following that, a single CE and the TE at each site examined the calibration subjects, alternating in the order of their examinations for successive subjects. Their data collections were then submitted to the revised diagnostic algorithms.

Validity of the revised diagnostic procedures and algorithms

A complete discussion of the results for the revised diagnostic algorithms has been published in detail elsewhere.³⁴ A summary of findings is provided here. Revised algorithm sensitivity and specificity exceeded target cutoffs of sensitivity (≥ 0.75) and specificity (≥ 0.95) for myofascial pain (Ia) with sensitivity = 0.82 and specificity = 0.99. Myofascial pain with limited opening (Ib) showed even better diagnostic accuracy with sensitivity of 0.93 and specificity of 0.97. When muscle pain diagnoses were combined (Ia or Ib), sensitivity was 0.91 and specificity 1.00. The combined joint pain diagnoses (IIIa or IIIb) showed sensitivity of 0.93 and specificity of 0.97. Disc displacement without reduction with limited opening (IIb) was also associated with target sensitivity (0.80) and specificity (0.97). The remaining intra-articular diagnoses including IIa, IIc, IIIb, and IIIc all showed sensitivity that was below target (0.35 to 0.53). Specificity ranged from deficient (0.80) to meeting target.

Reliability of the revised diagnostic procedures and algorithms

Diagnostic reliability for the revised algorithms ranged from good to excellent. The lowest reliability coefficient observed was $k = 0.63$ for the diagnosis of IIb. As for muscle pain, Ia reliability approached excellence ($k = 0.73$), Ib was excellent ($k = 0.92$) as was the combined diagnosis, Ia or Ib, at $k = 0.83$. Reliability for joint pain was also excellent with kappa of 0.81 for IIIa, and $k = 0.85$ for IIIa or IIIb. Reliability was good to excellent for diagnosis of degenerative joint changes with $k = 0.71$ for IIIb, $k = 0.79$ for IIIc, and $k = 0.87$ for the combined diagnosis, IIIb or IIIc.

Simple, parsimonious, revised diagnostic protocols and algorithms

The revised Group I, Group II, and Group III diagnostic algorithms (Figures 1, 2 and 3, respectively) are limited to just three nodes (split conditions). This is in contrast to the diagnostic algorithms for the original RDC/TMD that employed 5 nodes in the Group I algorithm, 12 nodes in the Group II algorithm, and 3 nodes in Group III algorithm. As with the original RDC/TMD algorithms, the revised Group II and III algorithms are side-specific. The most compelling evidence for the simplicity of the revised diagnostic tests is the good reliability of these procedures, and their ready transferability as demonstrated by the short training periods needed for the TEs who had not performed these tests prior to their preparation for these calibration sessions.

The revised algorithms for Group I and III both have the same initial node, that is, Question 3 taken from the RDC/TMD questionnaire: "Have you had pain in the face, jaw, temple, in front of the ear, or in the ear in the past month?"¹ A subject's positive endorsement of the pain history is then verified by a finding of *familiar pain* based on a simple clinical examination.

For Group I myofascial pain, confirmation of the pain complaint is based on a report of familiar pain that is elicited by palpation (2# digital pressure) for at least one site among a total of 12 muscle palpation sites. These sites include 6 sites bilaterally in the masseter (origin, body, insertion) and temporalis (anterior, middle, posterior) muscles. Confirmation

of myofascial pain is also made if the subject reports familiar pain in either of these muscles that is associated with maximum unassisted or assisted opening of the jaw. Differentiation of Ia (no limitation) from Ib (limitation) is based on the interincisal distance with unassisted jaw opening without pain, after correction of this measure for anterior tooth vertical overlap. The cutoff is ≥ 40 mm. (no limitation) versus < 40 mm. (limitation). There is no Group I diagnosis in the absence of a complaint of pain and its confirmation by the finding of familiar muscle pain.

For Group III, a pain endorsement is confirmed as joint related (arthralgia) based on a report of familiar pain that is elicited by digital joint palpation using either of the following methods: 1# pressure applied to the lateral pole of the joint, or 2# pressure applied around the lateral pole of the joint. Joint pain is also confirmed if the subject reports familiar joint pain that is associated with maximum unassisted or assisted opening of the jaw. Joint pain with normal osseous status (IIIa) is differentiated from joint pain that is associated with osseous degenerative changes (IIIb) using one finding: the presence or absence of crepitus. Degenerative joint change with no pain (IIIc) is also differentiated from a normal joint by the finding of crepitus. Typically, a diagnosis of crepitus when using the original RDC/TMD examination operationalization has showed just fair reliability ($k = 0.53$ in the Validation Project). In contrast, the revised method for crepitus detection has excellent reliability at $k = 0.85$. The revised test is positive when crepitus is detectable with palpation and audible at 6 inches from the subject, *or* if the subject reports crepitus during the course of the exam. There is no IIIa or IIIb diagnosis in the absence of familiar TMJ pain, and no IIIb or IIIc diagnosis in the absence of crepitus. This algorithm is side-specific, that is, exam findings of joint pain and/or crepitus are determined to be related to a specific joint.

For Group II disc displacements, the algorithm is also very simple. The initial test is based on a minimum of one reciprocal (both opening and closing) disc click during any of three repetitions of the vertical jaw movements. This node is also positive if just a single opening or closing click occurs, and there is a second click that occurs during any of three repetitions of excursive or protrusive movements. Like the Group III algorithm, the Group II algorithm is side-specific; the finding of a joint click must be related to a given joint. A positive finding of disc click is sufficient for a diagnosis of disc displacement with reduction (IIa) for that joint. The second node is defined by questions 14a and 14b of the RDC/TMD Questionnaire. 14a: "Have you ever had your jaw lock or catch so that it won't open all the way?" 14b: "Was this limitation in jaw opening severe enough to interfere with your ability to eat?" The third node is defined by a 40 mm. cutoff for interincisal distance based on maximum assisted jaw opening, corrected for anterior vertical overlap. A diagnosis of disc displacement without reduction with limited opening (IIb) is made if there is no disc click, the subject responds positively to Questions 14 a & b of the RDC/TMD questionnaire, and the corrected interincisal measurement is less than 40 mm. The diagnosis of disc displacement without reduction without limited opening (IIc) is rendered if there is no disc click, a positive history of interference, and the corrected jaw opening measurement is at least 40 mm. There is no Group II diagnosis when in the presence of no click, no history of interference as per Question 14b, and jaw opening of 40 mm. or greater.

Conclusions and recommendations relative to the new Axis I examination protocols and diagnostic algorithms

The most important reason for which TMD patients seek care is the pain associated with these disorders.^{35, 36} The 1996 NIH Technology Assessment Conference Statement on the Diagnosis and Management of Temporomandibular Disorders noted that an ideal diagnostic classification system for TMD should be based on etiology.³⁷ In order for this goal to be achieved, future epidemiologic studies are required in which the subjects will receive valid and reliable phenotypic classifications using simple clinical tests based on signs and

symptoms. These revised diagnostic procedures provide simple, transferable, reliable and valid Axis I diagnostic methods for both muscle pain and joint pain that will help facilitate the studies needed to develop a diagnostic taxonomy for TMD pain that is based on mechanism and etiology.

Acknowledgments

We thank the following personnel of the RDC/TMD Validation Project: at the University of Minnesota – Gary Anderson, Quentin Anderson, Mary Haugan, Amanda Jackson, Wenjun Kang, Pat Lenton, Wei Pan and Feng Tai; at the University at Buffalo – Richard Ohrbach (Site PI), Leslie Garfinkel, Yoly Gonzalez, Patricia Jahn, Krishnan Kartha, Sharon Michalovic and Theresa Speers; and at the University of Washington – Lars Hollender, Kimberly Huggins, 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 Craniofac Pain*. 1992; 6:301–355.
2. Smith, TW. Measurement in health psychology research. In: Friedman, HS.; Silver, RC., editors. *Foundations of Health Psychology*. New York: Oxford University Press; 2007. p. 19-51.
3. John MT, Dworkin SF, Mancl LA. Reliability of clinical temporomandibular disorder diagnoses. *Pain*. 2005; 118:61–69. [PubMed: 16154702]
4. Look JO, John MT, Tai F, Huggins KH, Lenton PA, Truelove EL, Ohrbach R, Anderson GC, Schiffman EL. Research diagnostic criteria for temporomandibular disorders: Reliability of Axis I diagnoses and selected clinical measures. *J Orofac Pain*. 2010; 24(1):25–34. [PubMed: 20213029]
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for reporting of diagnostic accuracy. *Clin Chem*. 2003; 49:1–6. [PubMed: 12507953]
6. Schiffman EL, Truelove EL, Ohrbach R, Anderson GC, John MT, List T, Look JO. Assessment of the validity of the research diagnostic criteria for temporomandibular disorders: Overview and methodology. *J Orofac Pain*. 2010; 24(1):7–24. [PubMed: 20213028]
7. VonKorff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. *Pain*. 1992; 50:133–149. [PubMed: 1408309]
8. Derogatis L. SCL-90-R: Symptom Checklist-90-R. Administration, Scoring and Procedures Manual. *Psychopharmacol Bull*. 1994; 9:12–28.
9. Derogatis LR, Lipman RS, Covi L. SCL-90: an outpatient psychiatric rating scale--preliminary report. 1973; 9:13–28.
10. Williamson JM, Lipsitz SR, Manatunga AK. Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*. 2000; 1:191–202. [PubMed: 12933519]
11. Fleiss, JL.; Levin, B.; Paik, MC. *Statistical Methods for Rates and Proportions*. Hoboken, NJ: Wiley-Interscience; 2003.
12. Habets LL, Bezuur JN, Naeiji M, Hanson TL. The orthopantomogram, an aid in diagnosis of temporomandibular joint problems. II. The vertical symmetry. *J Oral Rehabil*. 1988; 15:465–471. [PubMed: 3244055]
13. Ludlow JB, Davies KL, Tyndall DA. Temporomandibular joint imaging: a comparative study of diagnostic accuracy for the detection of bone change with biplanar multidirectional tomography and panoramic images. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. 1995; 80:735–743. [PubMed: 8680983]
14. Ahmad M, Hollender L, Anderson Q, Kartha K, Ohrbach RK, Truelove EL, John MT, Shiffman EL. Research diagnostic criteria for temporomandibular disorders (RDC/TMD): Development of image analysis criteria and examiner reliability for image analysis. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. 2009; 107(6):844–860. [PubMed: 19464658]
15. Efron, B.; Tibshirani, R. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.

16. Wolfe F, Smythe HA, Yunus MB, Bennett RM, Bombardier C, Goldenberg DL, Tugwell P, Campbell SM, Abeles M, Clark P. The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. *Arthritis Rheum.* 1990; 33:160–172. [PubMed: 2306288]
17. Steenks, MH.; deWijer, A.; Lobbezoo-Scholte, AM.; Bosman, F. Orthopedic Diagnostic Tests for Temporomandibular and Cervical Spine Disorders. In: Friction, J.; Dubner, R., editors. *Advances in Pain Research and Therapy Orofacial Pain and Temporomandibular Disorders*. New York, New York: Raven Press; 1995.
18. Lobbezoo-Scholte AM, Steenks MH, Faber JA, Bosman F. Diagnostic value of orthopedic tests in patients with temporomandibular disorders. *J Dent Res.* 1993; 72:1443–1453. [PubMed: 8408888]
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. [PubMed: 8057195]
20. Visscher CM, Lobbezoo F, Naeije M. A reliability study of dynamic and static pain tests in temporomandibular disorder patients. *J Orofac Pain.* 2007; 21:39–45. [PubMed: 17312640]
21. Okeson, JP. *Management of Temporomandibular Disorders and Occlusion*. St. Louis, MO: Mosby Year Book; 1993. History and examination for temporomandibular disorders. Anonymous.
22. Ohrbach R, Gale EN. Pressure pain thresholds, clinical assessment, and differential diagnosis: reliability and validity in patients with myogenic pain. *Pain.* 1989; 39:157–169. [PubMed: 2594394]
23. Howard, J. Clinical Diagnosis of Temporomandibular Joint Derangements. In: Moffett, BC., editor. *Diagnosis of Internal Derangements of the Temporomandibular Joint*. Seattle, Washington: Continuing Dental Education, University of Washington; 1984.
24. Wright, EF. Anonymous. Ames, Iowa: Blackwell Munksgaard; 2005. *Manual of Temporomandibular Disorders*.
25. Friction, J.; Kroening, R.; Hathaway, KM. Anonymous. St. Louis, MO: Ishiyaku EuroAmerica, Inc; 1988. *TMJ and Craniofacial Pain: Diagnosis and Management*.
26. Schiffman E, Friction J, Haley DP. The relationship of occlusion, parafunctional habits and recent life events to mandibular dysfunction in a non-patient population. *J of Oral Rehab.* 1992; 19:201–223.
27. Anderson GC, Schulte JK, Aeppli DM. Reliability of the evaluation of occlusal contacts in the intercuspal position. *The J of Prosth Dent.* 1993; 70:320–323.
28. Dawson, PE. Determining Centric Relation. In: Dawson, PE., editor. *Functional Occlusion From TMJ to Smile Design*. St. Louis, Missouri: Mosby Elsevier; 2007.
29. Headache Classification Subcommittee of the International Headache Society. The International Classification of Headache Disorders. ICHD-II Tension-type headache (TTH). *Cephalalgia.* 2004; 24(Supplement 1):37–43. [PubMed: 14687011]
30. Sackett DL. Bias in analytic research. *J Chronic Dis.* 1979; 32(1–2):51–63. [PubMed: 447779]
31. Truelove E, Pan W, Look JO, Mancl LA, Ohrbach RK, Velly A, Higgins K, Lenton P, Schiffman EL. Research diagnostic criteria for temporomandibular disorders: Validity of Axis I diagnoses. *J Orofac Pain.* 2010; 24(1):35–47. [PubMed: 20213030]
32. Breiman; Friedman; Olshen; Stone. *Classification and Regression Trees*. Wadsworth; 1984. p. 6-58, p. 221-247, p. 306-317.
33. Hastie, T.; Tibshirani, R.; Friedman, J. Section 7.10: Cross-Validation. Springer; 2001. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*; p. 214-217.
34. Schiffman EL, Ohrbach R, Truelove EL, Tai F, Anderson GC, Pan W, Gonzalez YM, John MT, Sommers E, List T, Velly AM, Look JO. Research diagnostic criteria for temporomandibular disorders: Methods for development, reliability and validity of revised diagnostic algorithms for Axis I J Orofac Pain. 2010; 24(1):63–78.
35. Al-Hasson HK, Ismail AI Jr, Ash MM. Concerns of patients seeking treatment for TMJ dysfunction. *J Prosthet Dent.* 1986; 56:217–21. [PubMed: 3463745]
36. Dworkin SF, Huggins KH, Wilson L, Mancl L, Turner J, Massoth D, et al. A randomized clinical trial using research diagnostic criteria for temporomandibular disorders-Axis II to target clinic

cases for a tailored self-care TMD treatment program. *J Orofac Pain*. 2002; 16(6):48–63. [PubMed: 11889659]

37. Proceedings Oral Surg Oral Med Oral Pathol Oral Radiol Endod; National Institutes of Health Technology Assessment Conference on Management of Temporomandibular Disorders; Bethesda, Maryland. April 29-May 1, 1996; 1992. p. 49-183.

Summaries of topics presented at the 2008 International Association of Dental Research in Toronto (2008)

- Reliability of RDC/TMD Axis I diagnoses based on clinical signs and symptoms
- Reliability of radiographic interpretations used for RDC/TMD Axis I diagnoses
- Reliability of self-report data used for RDC/TMD Axis I diagnoses
- Validity of RDC/TMD Axis I diagnoses based on clinical signs and symptoms
- Proposed revisions of the RDC/TMD Axis I diagnostic algorithms

Figure 1.

Revised Group I Muscle Disorders diagnostic algorithm. Reprinted by permission from the Journal of Orofacial Pain 2010, 24(1): p. 69.

Figure 2. Revised Group II Disc Displacements diagnostic algorithm. Reprinted by permission from the Journal of Orofacial Pain 2010, 24(1): p. 70.

Figure 3.

Revised Group III Arthralgia, Arthritis, and Arthrosis diagnostic algorithm. Reprinted by permission from the Journal of Orofacial Pain 2010, 24(1): p. 71. One change has been made to the original Figure 3 published in Journal of Orofacial Pain. For clarity and consistency with the manuscript text, the conjunction “or” follows the diagnostic test, *Palpation of the lateral pole with 1 pound pressure*.