RESEARCH

Evaluation of the reproducibility in the interpretation of magnetic resonance images of the temporomandibular joint

KW Butzke, KD Batista Chaves, HE Dias da Silveira and HL Dias da Silveira*

Faculty of Dentistry, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Objectives: The aim was to determine the intra- and interexaminer reproducibility in the interpretation of MRI of the temporomandibular joint among independent observers, with respect to six specific articular characteristics, and to discover which of these had greater and lesser agreement.

Methods: 30 magnetic resonance examinations of temporomandibular joints of adults were independently interpreted by 9 experienced and trained observers at 2 different times. Observers were divided into three groups according to their specialties: surgeon dentists specialized in temporomandibular dysfunction and orofacial pain, surgeon dentists specialized in radiology and medical doctors specialized in radiology. The reproducibility analysis was carried out using Cohen's kappa coefficient.

Results: The interexaminer reproducibility ranged from slight to fair. The intraexaminer reproducibility ranged from slight to no agreement. In the interexaminer evaluation, anterior disc displacement without reduction presented greater agreement, whereas change in condylar head shape showed the poorest agreement. In the intraexaminer evaluation, anterior disc displacement without reduction presented slight agreement, whereas, for the other characteristics, no agreement was observed.

Conclusion: Examiners do not demonstrate reproducibility in the interpretation of MRI of temporomandibular joints. Therefore, more efforts are necessary with respect to understanding the changes that may be detected in these images in terms of diagnosis and appropriate treatment approaches.

Dentomaxillofacial Radiology (2010) **39**, 157–161. doi: 10.1259/dmfr/16076453

Keywords: magnetic resonance imaging; temporomandibular joint; reproducibility

Introduction

Articular pathologies are commonly present in patients with temporomandibular dysfunction (TMD) and are diagnosed by clinical examination coupled with imaging diagnosis methods. In the last two decades, considerable improvement has been observed in the diagnostic potential of imaging techniques, which has significantly increased the understanding of the changes that affect the temporomandibular joint (TMJ).¹

Imaging techniques offer distinct advantages over the clinical evaluation in the diagnosis of TMDs of

articular origin. They can provide definitive evidence of pathological changes and also document the effectiveness of some treatments. Different techniques for TMJ evaluation have been described in the literature. However, some procedures are not possible when evaluating the soft tissue (for example, CT), while others impose the disadvantage of being invasive (for example, arthrography).²

MRI has been considered the technique of choice for the examination of TMJs as it provides high-quality images and excellent definition of both soft and hard articular tissues. In addition, patients are not exposed to an invasive procedure with the injection of contrast material or ionizing radiation.

The most frequent pathologies that affect the TMJ are disc interference disorders of inflammatory and degenerative origins.³ In this scenario, MRI allows the

*Correspondence to: Heraldo Luis Dias da Silveira, Rua Ramiro Barcellos, 2492. Porto Alegre, Brazil. 90035-003, Faculdade de Odontologia – Universidade Federal do Rio Grande do Sul, Departamento de Cirurgia e Ortopedia. Disciplina de Radiologia - 5° andar; E-mail: heraldods@ig.com.br
Received 3 November 2008; revised 30 January 2009; accepted 17 February 2009

diagnosis of these changes by providing essential information on the location and intensity of the signal and on the morphology of the articular disc, as well as on the amount of synovial fluid and the condition of the cortical bone, bone marrow and retrodiscal tissues. These characteristics may be related to pain and dysfunction and should be considered important in the treatment of patients with TMD.⁴

Two significant factors affect the validity and reliability of imaging diagnosis: diagnostic accuracy of the technique and performance of the examiner; therefore, the examiner plays a crucial role in imaging diagnosis.⁵ Although the evaluation of TMJ through MRI has been consistently defined with respect to diseases and to normality, some studies have shown that different interpretations may occur when presented by two or more examiners.⁶⁻⁷ Non-calibrated observers normally show an agreement ranked as slight to moderate when they evaluate these images, but better results are found when calibration is required or when an evaluation protocol is utilized.⁸ Therefore, it is known that the standardization with the use of images of reference,⁹ of criteria of description¹⁰ and of calibration⁸ considerably reduces the disagreement across examiners and increases the reliability of diagnosis. Furthermore, these studies show that correlation is high when calibrated specialists work in a group.

However, this is not usual in everyday clinical situations, when professionals, who are involved in imaging diagnosis of TMDs, work independently. The surgeon dentist who has attended post-graduate courses on TMD and orofacial pain works directly on the treatment of these disorders and uses MRI examinations of the TMJ, which are interpreted by doctors and surgeon dentist post-graduates in radiology. Therefore, it is necessary that these professionals are trained and that a consensus is reached, since wrong information generated from these examinations may complicate the treatment, causing unnecessary discomfort to the patient.

Therefore, in an attempt to reproduce this situation, the objective of this study was to verify reproducibility when non-calibrated examiners evaluate TMJ MRI scans independently, without the use of criteria for pre-established standards, as observed in daily clinical practice. This is important because most of the treatment strategies for TMD are based on data obtained from clinical findings supplemented by TMJ imaging.

Materials and methods

This is an observational transverse study of reproducibility,¹² conducted after the examination and approval of its protocol by the commission of research and committee of ethics in research of the University.

30 MRIs of 30 adult patients of both genders were used. These examinations generated 60 image stacks of both TMJs of all patients. The examinations were retrieved from the files of a clinic specializing in image diagnosing. The inclusion criterion was the technical quality of the examinations. The images showed good resolution and sharpness without blurring, allowing the clear observation of all articular structures. The examinations were conducted in an MRI unit (Magnetom 63 SP-Siemens, Erlangen, Germany), 1.5 tesla, in sequences generated in T_1 , T_2 and PD (proton density) and with 3-mm-thick planes. TMJs were bilaterally evaluated in the oblique sagittal plane (in mouth closed and mouth open positions) and in the coronal plane (only with the mouth closed). The oblique sagittal slices were performed perpendicularly to the long axis of the condylar head and coronals, in parallel to the anatomical structures. Maximum intercuspation was used for the acquisition of images in the mouth closed position, and maximum mouth opening for the mouth open position. Cinematic images were also prepared during the examinations, allowing the sequential evaluation of the articular movement.

The 9 examiners, described in detail below, evaluated the image stacks of the 60 TMJs during their familiar work routine. For inspection of the images, the examiners received a table listing six possible articular characteristics to be assessed. Examiners were required to mark the absence or presence of the different specified characteristics: changes in the disc (shape; anterior disc displacement with reduction (DDR); anterior disc displacement without reduction (DDWR) changes in the condylar head (CH) (shape; osteophyte); and effusion. The interpretations were made independently by each examiner, in two separate stages, with a 30 day interval.

The examiners were divided into three separate groups according to their specialties. Group 1 comprised dentists who received specialized training over a 2 year full-time post-graduate course in TMD and orofacial pain with training in MRI interpretation. Group 2 comprised dentists who received specialized training over a 2 year full-time post-graduate course in oral radiology with training in MRI interpretation of the TMJ; they all had at least 2 years of clinical practice in this field. Group 3 comprised medical doctors who received specialized training over a 3 year full-time post-graduate course in medical radiology; they were all used to MRI interpretation of the TMJ as part of their clinical duties for at least 2 years. The intra- and interexaminer reproducibilities were evaluated for the three groups conjointly and for each group separately, as well as the articular characteristics that exhibited higher and lower agreement for the observations made.

Statistical analysis

The variables examined were classified as qualitative and dichotomous, because they are of the presence/

absence type. The evaluation of the agreement between the examiners (interexaminer reproducibility) and between the two examination periods of the same examiner (intraexaminer reproducibility) was conducted using Cohen's kappa coefficient. The calculation of kappa coefficients was based on the mean values of the left and right side of the TMJ. To verify whether the measurement obtained was satisfactory or not, the interpretation was based on the data of Landis and Koch,¹² which suggest the following interpretation for kappa: <0 no agreement; 0.00–0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; and 0.81–1.00 almost perfect agreement. Data analysis was performed using the SPSS program, version 15.0 for Windows (SPSS, Chicago, IL).

Results

Interexaminer reproducibility

The interexaminer reproducibility in each group is shown in Table 1. The mean kappa values in Group 1 show that reproducibility was (i) slight for DDR and change in CH shape; (ii) fair for change in disc shape, effusion and osteophyte; and (iii) moderate for DDWR. Group 2 presented the same reproducibility as Group 1, except for effusion, which showed no agreement. In Group 3 reproducibility was (i) slight for change in disc shape; (ii) fair for DDR, changes in CH and effusion; (iii) moderate for osteophyte; and (iv) substantial for DDWR. It is important to note that these figures indicate that the mean reproducibility ranged from slight (Group 2) to fair (Groups 1 and 3), and that these reproducibility values are lower than acceptable to indicate a good agreement (kappa > 0.60). In total, Groups 1 and 3 exhibited fair reproducibility and Group 2 showed no agreement.

Table 1 shows the interexaminer reproducibility among groups. This was fair in all comparisons, with mean kappa values between 0.18 and 0.41.

Intraexaminer reproducibility

The intraexaminer reproducibility in each group is shown in Table 2. The mean kappa values show that the three groups' reproducibilities ranged from slight to no agreement.

Variables that had higher and lower agreement

In the evaluation of the interexaminer reproducibility the variables change in CH shape and effusion presented the lowest agreement (slight), whereas the variable DDWR exhibited the highest agreement (moderate). In the evaluation of the intraexaminer reproducibility, osteophytes revealed the absence of agreement, but other variables showed slight agreement.

Discussion

Diagnosis by MRI has been extremely important in the evaluation of patients with TMJ disorders, in that it complements the findings of clinical examination and assists in the development of an appropriate treatment strategy. As the examiner interprets the information of the image, the diagnosis is the result of an interaction between this information and the evaluation of the observer; therefore, in addition to the quality of the image, the good performance of the examiner is essential for a correct and reliable diagnosis.

Even when a standard protocol is used for the acquisition of MRI of the TMJ, some factors, such as the presence of artefacts, blot (interference), differences in the visual field and individual anatomical aspects, may interfere with the quality of the images.¹³ The examinations that were included in this research were selected because of the technical quality they presented, and thus the bias surrounding the problems involved in the acquisition of images was eliminated. The lack of reproducibility found for the articular characteristics evaluated is, therefore, related to the examiner and his or her performance in the identification of these characteristics.

Programmes for calibration have been proposed to reduce the variability of examiners and increase the reliability of imaging diagnosis. The study conducted by Orsini *et al*⁸ showed that, after calibration, interexaminer agreement was improved from moderate to substantial in the evaluation of disc position and from fair to moderate when its configuration had been interpreted. Moreover, the intraexaminer agreement was substantial for configuration and almost perfect for disc position. According to the authors, variability is reduced when examiners are involved in a training

Table 1 Interexaminer reproducibility within each group. Mean kappa values among the groups for each variable

		Group 1		Group 2		Group 3		General mean	
		Mean	Range	Mean	Range	Mean	Range	Mean	Range
Changes in disc	Shape	0.29	0.08–0.52	0.21	0.16–0.29	0.20	0.07–0.39	0.23	0.07–0.52
	DDR	0.20	0.00–0.44	0.12	0.07–0.15	0.34	0.14–0.64	0.22	0.00–0.64
Changes in CH	DDWR	0.54	0.44–0.71	0.48	0.31–0.62	0.71	0.57–0.79	0.58	0.31–0.79
	Shape	0.19	0.12–0.27	0.04	0.00–0.16	0.36	0.16–0.56	0.20	0.00–0.56
	Osteophyte	0.35	0.17–0.53	0.24	0.13–0.39	0.52	0.41–0.57	0.37	0.13–0.57
Effusion		0.38	0.26–0.50	0.00	0.00–0.01	0.34	0.18–0.48	0.24	0.00–0.50
General mean		0.33		0.18		0.41		0.31	

CH, condylar head; DDR, disc displacement with reduction; DDWR, disc displacement without reduction

Table 2 Intraexaminer reproducibility within each group. Mean kappa values among the groups for each variable

		<i>Group 1</i>		<i>Group 2</i>		<i>Group 3</i>		<i>General mean</i>	
		<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>	<i>Mean</i>	<i>Range</i>
Changes in disc	Shape	0.00	0.00–0.10	0.05	0.00–0.25	0.09	0.00–0.25	0.03	0.00–0.25
	DDR	0.04	0.00–0.18	0.00	0.00–0.09	0.00	0.00–0.11	0.01	0.00–0.18
	DDWR	0.12	0.00–0.33	0.20	0.02–0.44	0.02	0.00–0.06	0.11	0.00–0.44
Changes in CH	Shape	0.08	0.00–0.33	0.02	0.00–0.19	0.00	0.00–0.02	0.05	0.00–0.33
	Osteophyte	0.00	0.00–0.01	0.06	0.00–0.14	0.00	0.00–0.01	0.02	0.00–0.14
Effusion		0.05	0.00–0.13	0.15	0.07–0.31	0.08	0.00–0.16	0.09	0.00–0.31
General mean		0.05		0.08		0.03		0.05	

CH, condylar head; DDR, disc displacement with reduction; DDWR, disc displacement without reduction

programme in which they work in groups and establish clear evaluation criteria. Nebbe *et al*⁶ evaluated the reproducibility in the interpretation of articular disc location through MRI among four independent examiners, and the correlation obtained ranged from moderate to substantial. In that study, the examiners were calibrated and received a protocol outlining six classification criteria for the disc location. Therefore, the data from these studies, although they are encouraging, do not represent the reality in current clinical practice, in which MRI examinations of the TMJ frequently are interpreted by professionals from different areas who work individually and use their own evaluation criteria in an independent fashion.

The model presented in this research reflects the results to be expected in a clinical environment and, contrary to what is desired, it shows low inter- and intraobserver agreement (when evaluating the same image at different times), in spite of the fact that these results were produced by experienced evaluators used to interpreting these examinations.

Schmitter *et al*¹⁴ showed that generalist surgeon dentists had better reproducibility than medical doctors specialized in radiology in the evaluation of disc position and morphology and of the condylar head when high-quality images were used. According to the authors, this may have been the result of different training. Surgeon dentists are familiar with the TMDs in their studies because TMJ is part of the stomatognathic system, whereas in the medical area less emphasis is placed on this structure. Similar results could be expected for this research; however, where there were differences, this was not important.

Widmalm *et al*¹⁵ tested the hypothesis that the agreement among non-calibrated examiners and those who were experienced in the evaluation of 13 articular characteristics would be high, with kappa values above 0.60. Yet, the hypothesis was not supported by the results of the study because the mean kappa values were very low in all characteristics. In our opinion, one reason for the limitations in the diagnosis may have been because some examinations used in that research did not involve T_2 images. However, this study showed that even with a standard test, using T_1 , T_2 and PD images, which allow a better contrast among different articular tissues, the reproducibility was low. Since the highest value found for the kappa coefficients among all variables evaluated was

0.58, the increase in the number of available resources has not improved the performance of the observers.

Few studies in the literature evaluate intraexaminer reproducibility, and those that show good agreement included programs for calibration in the respective methodology.⁷⁻⁸ The main reasons for the good performance of examiners would be the use of well-defined criteria for the interpretation of the images and the relatively extensive training. In this research, the observers were given a table indicating the articular characteristics to be investigated, and the results were contradictory. In this scenario, the lack of agreement with regard to some variables and the low values for others raise concerns and are grounds for further investigation.

The goal of this study was to evaluate the agreement of the diagnosis in the interpretation of different signals commonly reported in the evaluations of MRI of the TMJ. The number of parameters suggested for these assessments is very high, with as many as 52 in a publication.¹⁶ The number of characteristics evaluated in this research was six; therefore, the results cannot be extrapolated to the other changes. However, the disc position and the morphological changes of the articular structures investigated are the signs of primary interest when we evaluate patients with TMJ disorders. In this sense, the correct diagnosis is extremely important, as it specifies the treatment requirements.

The results found in this study show that, despite the advanced MRI technology currently available to investigate the TMJ, which allows excellent visualization of articular structures, the professionals, regardless of their training (surgeon dentist post-graduates in TMD and orofacial pain, medical doctors of radiology and surgeon dentist post-graduates in oral radiology) show a lack of intraexaminer reproducibility. The results also showed low interexaminer reproducibility in the analysis of all articular characteristics. DDWR was the change that presented the best agreement, whereas the change in CH shape exhibited the poorest agreement.

Therefore, it is possible to conclude that more effort is needed towards a better understanding and a more accurate interpretation of the changes observed on MRI images of the TMJ. This is important because incorrect imaging diagnoses may jeopardize the clinical diagnosis and, therefore, the treatment of patients.

References

1. Rudisch A, Innerhofer K, Bertram S, Emshoff R. Magnetic resonance imaging findings of internal derangement and effusion in patients with unilateral temporomandibular joint pain. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 2001; **92**: 566–571.
2. Liedberg J, Panmekiate S, Petersson A, Rohlin M. Evidence-based evaluation of three imaging methods for the temporomandibular disc. *Dentomaxillofac Radiol* 1996; **25**: 234–241.
3. Schellhas, K.P. Temporomandibular joint injuries. *Radiology* 1989; **173**: 211–216.
4. Emshoff R, Brandlmaier I, Bertram S, Rudisch A. Relative odds of temporomandibular joint pain as a function of magnetic resonance imaging findings of internal derangement, osteoarthritis, effusion, and bone marrow edema. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 2003; **95**: 437–445.
5. Schmitter M, Kress B, Koob CLA, Gabbert O, Rammelsberg P. Temporomandibular joint disk position assessed at coronal MR imaging in asymptomatic volunteers. *Radiology* 2005; **236**: 559–564.
6. Nebbe B, Brooks SL, Hatcher D, Hollender LG, Prasad NGN, Major PW. Magnetic resonance imaging of the temporomandibular joint: interobserver agreement in subjective classification of disk status. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 2000; **90**: 102–107.
7. Tasaki MM, Westesson PL, Raubertas RF. Observer variation in interpretation of magnetic resonance images of the temporomandibular joint. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 1993; **76**: 231–234.
8. Orsini GM, Terada S, Kuboki T, Matsuka Y, Yamashita A. The influence of observer calibration in temporomandibular joint magnetic resonance imaging diagnosis. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 1997; **84**: 82–87.
9. Cholitgul W, Petersson A, Rohlin M, Tanimoto K, Akerman S. Diagnostic outcome and observer performance in sagittal tomography of the temporomandibular joint. *Dentomaxillofac Radiol* 1990; **19**: 1–6.
10. Panmekiate S, Petersson A, Rohlin M, Akerman S. Observer performance and diagnostic outcome in diagnosis of the disk position of the temporomandibular joint using arthrography. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 1994; **78**: 255–263.
11. Goldim JR. *Manual de iniciação à pesquisa em saúde*. Porto Alegre: Dacasa Editora, 2000, p 199.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1997; **33**: 159–174.
13. Langlais RP, Van Rensburg LJ, Guidry J, Moore WS, Miles DA, Nortjé CJ. Magnetic resonance imaging in dentistry. *Dent Clin North Am* 2000; **4**: 411–426.
14. Schmitter M, Kress B, Hähnel S, Rammelsberg P. The effect of quality of temporomandibular joint MR images on interrater agreement. *Dentomaxillofac Radiol* 2004; **33**: 253–258.
15. Widmalm SE, Brooks SL, Sano T, Upton LG, McKay DC. Limitation of the diagnostic value of MR images for diagnosing temporomandibular joint disorders. *Dentomaxillofac Radiol* 2006; **35**: 334–338.
16. Stack Jr B, Stack B. Underutilization of MRI. A suggested protocol. *J Craniomand Pract* 1998; **16**: 131–132.