Community Dent Oral Epidemiol 2006; 34: 41–6




Community Dentistry and Oral Epidemiology

# Inter-examiner reliability in the clinical examination of temporomandibular disorders: influence of age

**Alexander J. Hassel, Peter Rammelsberg and Marc Schmitter**

Poliklinik für Zahnärztliche Prothetik, Im Neuenheimer Feld 400, Heidelberg, Germany

Hassel AJ, Rammelsberg P, Schmitter M. Inter-examiner reliability in the clinical examination of temporomandibular disorders: influence of age. Community Dent Oral Epidemiol 2006; 34: 41–6. © Blackwell Munksgaard, 2006

Abstract – *Objective:* The aim of this study was to investigate the influence of the age of the subject on inter-examiner reliability of the clinical signs of temporomandibular disorder (TMD). *Methods:* Forty-three elderly (ES) and 44 younger adults (YS) were selected. The female/male distribution was almost the same in the two groups. All participants underwent clinical examination according to the Research Diagnostic Criteria for TMD, performed successively by two clinicians. *Results:* For metric measurements – with the exception of unassisted opening – the ES gave a significantly lower range of motion with both examiners and significantly worse percentage agreement between the examiners. A remarkable inter-examiner disagreement in the elderly was found with laterotrusion and protrusion movements. The prevalence of joint sounds was rated inconsistently by the examiners. The reliability of detection was not different in the two groups. The prevalence of tender muscle sites was also inconsistent. The overall percentage agreement for subjects with at least one tender muscle point was not age dependent. Because of the very low prevalence in ES, further statistical assessment of reliability is not possible. *Conclusions:* The age-dependent lower range of motion and the inferior reliability of metric measurements in the elderly could lead to wrong diagnoses. The reliability of detecting joint sounds and tender muscles was not age dependent within the limitations of the study.

Key words: elderly; reliability; TMJ disorders

Alexander Hassel, Poliklinik für Zahnärztliche Prothetik, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany
Tel: +49 6221 566045
Fax: +49 6221 565371
e-mail: alexander.hassel@med.uni-heidelberg.de

Submitted 19 November 2004; accepted 21 April 2005

It is never possible completely to avoid error or uncertainty when detecting and measuring clinical signs and symptoms (1). During clinical examination of temporomandibular disorders (TMD), the goal must be to optimize the reliability of clinical findings by increasing the consistency of the measurements (2). The Research Diagnostic Criteria for temporomandibular disorders (RDC/TMD) were developed to improve the reliability of the diagnosis of TMD (3). Several studies on the reliability of clinical signs according to the RDC/TMD have been performed. These have shown that the reliability in the different clinical parameters could be low in some cases (2–5) and that calibration of the examiners improves results (4–6).

With the shift of the age pyramid to the elderly, topics concerning the aged subject have become of increasing interest. The health status of a person is strongly influenced by age. Many diseases, such as type II diabetes, heart diseases and cancer, increase in frequency with ageing. However, not only the diseases themselves, but the measurements or interpretation of the clinical signs of diseases could be influenced by age, too. For example, decreased ability to detect higher frequencies in hearing tests in the elderly is not a pathological finding, as it is in the young, but is simply due to ageing.

Several previous studies have addressed TMD in the elderly. These were mainly concerned with the prevalence of signs of TMD and their subjective

impact on the elderly (7–11). However, the influence of age on the clinical measurement of the signs and symptoms of TMD and on the inter-examiner reliability of these findings has not been yet examined. This could be related to the lower physical dexterity of the elderly subjects. It is possible that the performance and reproduction of mandibular movements could be more limited than in younger subjects. This could lead to wrong diagnoses, not because of pathologic findings but due to age.

Therefore, the aim of this study was to measure the inter-examiner reliability of clinical signs of TMD according to the RDC/TMD in a group of elderly subjects, in comparison with a group of younger adults.

## Materials and methods

### *Subjects*

This study was approved by the review board of the University of Heidelberg (June 2003). Eighty-seven subjects were included in this study. All elderly subjects received written information and signed an informed consent form. The elderly sample (ES; 43 subjects; 28% males, 72% females; mean age 82.5 years, from 68 to 96) was chosen from the inhabitants of four geriatric care centres needing moderate care (care level I). If this group was too large in any one centre, a random sample was taken. As comparison group, 44 younger adults were recruited (YS; 44 subjects; 32% males, 68% females; mean age 27.5 years, ranging from 18 to 45). The members of the YS were volunteers and students from the department. They were chosen with the same distribution of gender and the same range of age as ES. As no ES patient was undergoing current TMD treatment, only YS patients who were not in current treatment were selected. The study was designed as pilot study. A power estimation was not possible, as there were no other studies on the influence of age on inter-examiner reliability.

### *Clinical examination*

The clinical examination was performed by two clinicians using the RDC/TMD. One of the examiners (examiner I) was calibrated by the gold standard (calibration meeting with Prof. S. Dworkin); the second examiner (examiner II) was not calibrated by the gold standard, but worked clinically with the RDC/TMD. Before starting this investigation examiner I gave an additional theoretical and practical lesson of several hours to examiner II in order to minimize effects of different calibration. The examination was performed strictly according to the RDC/TMD and each subject was examined successively by both examiners.

Clinical measurements of the range of motion were obtained using a ruler. Unassisted opening (opening as wide as possible without pain), maximum unassisted opening (even if pain is felt), maximum assisted opening (forced by examiner, even if pain felt), maximum laterotrusion and protrusion were recorded. The results were recorded in millimetres. Joint sounds during motions were determined using digital palpation (2) and were recorded dichotomously, as joint sound detected or not. The palpation of jaw muscles was performed using the defined pressure of 900p with two fingers and was recorded for the different muscle sites dichotomously. The exact examination procedure has been described elsewhere (3).

### *Statistics*

The reliability of continuous variables was calculated with the intraclass correlation coefficient (ICC). The mean difference between the examiners in millimetres and the percentage disagreement in relation to the values of examiner I (absolute millimetre difference between the examiners divided by the measured value from examiner I) were also calculated. The values found by two examiners for categorical parameters were calculated with $\kappa$ statistics (Cohen's $\kappa$). The overall percentage agreement was also calculated. The differences between two groups were calculated with the Mann–Whitney *U*-test for independent groups; the level of significance was set at $P < 0.05$.

All statistics were performed using SPSS Version 11.5.1 (SPSS Inc., Chicago, IL, USA).

## Results

### *Metric measurements*

With the single exception of the measurement of unassisted opening, all values (millimetres of motion) of the YS were statistically greater than for the ES (Table 1). The percentage difference between examiners was significantly greater for the ES, again with the exception of unassisted opening motion. For lateral movements and protrusion of the mandibular, the ICC values were worse in

Table 1. Values for metric measurements in ES and YS

| | ES | YS |
|---|---|---|
| Unassisted opening | | |
| Middle in mm (examiner I) | 43.0 (40–45.9) | 46.9 (44–49.8)* |
| Middle in mm (examiner II) | 42.7 (40.1–45.4) | 45.9 (43–48.7) |
| Difference between examiners | 4.4 (3.2–5.6) | 4.0 (3–5) |
| %Difference between examiners | 0.11 (0.08–0.16) | 0.09 (0.06–0.11) |
| ICC-value | 0.88 (0.78–0.94) | 0.91 (0.83–0.95) |
| Maximum unassisted opening | | |
| Middle in mm (examiner I) | 46.5 (43.9–49.2) | 52.7 (50.3–55.1)** |
| Middle in mm (examiner II) | 46.5 (44–49.1) | 51.8 (49.2–54.3)** |
| Difference between examiners | 2.7 (2.0–3.3) | 1.7 (1.1–2.2) |
| %Difference between examiners | 0.06 (0.04–0.08) | 0.03 (0.02–0.05)** |
| ICC-value | 0.95 (0.91–0.97) | 0.98 (0.96–0.99) |
| Maximum assisted opening | | |
| Middle in mm (examiner I) | 48.2 (45.7–50.7) | 54 (51.7–56.1)** |
| Middle in mm (examiner II) | 47.9 (45.3–50.4) | 52.6 (50.2–55)* |
| Difference between examiners | 2.2 (1.6–2.8) | 1.6 (1–2.2) |
| %Difference between examiners | 0.05 (0.03–0.06) | 0.03 (0.02–0.04)* |
| ICC-value | 0.96 (0.92–0.98) | 0.98 (0.96–0.99) |
| Maximum laterotrusion | | |
| Middle in mm (examiner I) | 7.9 (7.1–8.7) | 10.0 (9.4–10.7)** |
| Middle in mm (examiner II) | 8.0 (7.1–8.9) | 10.1 (9.5–10.8)** |
| Difference between examiners | 2.6 (2.0–3.3) | 1.8 (1.3–2.3) |
| %Difference between examiners | 0.34 (0.12–0.56) | 0.16 (0.1–0.21)** |
| ICC-value | 0.71 (0.45–0.84) | 0.77 (0.57–0.88) |
| Maximum protrusion | | |
| Middle in mm (examiner I) | 3.2 (2.4–3.9) | 4.7 (4.1–5.4)** |
| Middle in mm (examiner II) | 2.4 (1.7–3.1) | 5.5 (4.9–6.2)** |
| Difference between examiners | 1.4 (1.0–1.9) | 1.1 (0.8–1.4) |
| %Difference between examiners | 0.44 (0.32–0.56) | 0.20 (0.14–0.27)** |
| ICC-value | 0.78 (0.59–0.88) | 0.90 (0.81–0.95) |

Values in parenthesis represent 95% CI. *$P < 0.05$, **$P < 0.01$.

both groups, and in some cases not acceptable, according to Lee et al. (lower bound of 95% CI <0.75) (12).

## *Joint sounds*

The prevalence of joint sounds was inconsistent between the examiners (Table 2). Only the $\kappa$-value for joint sounds during opening motion in both groups and for closing motion in ES reached acceptable levels, according to Landis and Koch (13). The overall percentage agreement was above 83% for opening and closing motion in both groups, above 65% for laterotrusion and over 76% for protrusion. A significant difference between the groups could not be found.

## *Myogenous findings*

Examiner I found 34% subjects in the YS with at least one tender intraoral and 25% with a tender extraoral muscle point. In contrast, only 11% of subjects were detected with at least one intraoral and 7% with an extraoral tender muscle point (only three tender extraoral muscle points in all subjects together) in the ES ($P < 0.05$). The other examiner found nearly the same prevalence of subjects with at least one tender extraoral muscle site in ES and YS (14%), but more single tender muscles points in

Table 2. Values for joint sounds in YS and ES

| | ES | YS |
|---|---|---|
| Opening motion | | |
| Prevalence (examiner I) | 35.7 | 12.2* |
| Prevalence (examiner II) | 27.9 | 20.5 |
| $\kappa$-value (SD) | 0.62 (0.13) | 0.44 (0.17) |
| Percentage agreement | 83.3 | 85.4 |
| Closing motion | | |
| Prevalence (examiner I) | 25 | 7.3* |
| Prevalence (examiner II) | 19 | 18 |
| $\kappa$-value | 0.43 (0.17) | 0.33 (0.20) |
| Percentage agreement | 87.2 | 90.2 |
| Laterotrusion | | |
| Prevalence (examiner I) | 35.7 | 24.4 |
| Prevalence (examiner II) | 39 | 40.9 |
| $\kappa$-value | 0.33 (0.15) | 0.23 (0.15) |
| Percentage agreement | 68.3 | 65.0 |
| Protrusion | | |
| Prevalence (examiner I) | 26.2 | 17.9 |
| Prevalence (examiner II) | 20 | 27.3 |
| $\kappa$-value | 0.35 (0.17) | 0.33 (0.17) |
| Percentage agreement | 77.5 | 76.9 |

*$P < 0.05$.

YS. Subjects showed 14% prevalence for tenderness in at least one muscle site in ES and 20% in YS ($P > 0.05$). In summary, the overall percentage agreement between the examiners in subjects having at least one extraoral tender muscle site was 95% (ES) and 83% (YS); for intraoral muscle sites 86% in ES and 85% in YS. The percentage agreement was not age dependent ($P > 0.05$). Because of very low prevalence (for extraoral muscle sites three detections in over 800 tested points in ES by examiner I), meaningful comparison of reliability seemed not to be feasible.

### *Facial pain in the previous month*

As judged by both examiners, no subject in the ES and 7% of the YS reported facial pain in the previous month.

## Discussion

### *Study limitations*

The reliability of the measurement of clinical signs is closely related to the clinical sign stability. The difficulties with clinical sign stability in TMD have been described (14–16). In the present study design, it was not possible to evaluate the stability of clinical signs, especially if there were differences in stability in ES compared with YS. Because of the different calibration of the two examiners, there might be a bias with respect to the overall reliability coefficients in one group. But this possible bias occurs in both ES and YS and its influence is not of decisive importance regarding the differences between the groups. The selection of the subjects of the elderly was performed at random. As judged by both examiners, no subject in the ES reported ongoing pain in the previous month. This caused a limitation of the study population with suppression of the reliability coefficients of subjects with ongoing facial pain.

In conclusion, the results of the reliability analysis have to be interpreted within these limitations of the study design.

### *Metric measurements*

Similar to previous studies, the present study demonstrates better reliability of unassisted, maximum unassisted and maximum assisted jaw opening in both elderly and younger adults than with ICC values for lateral excursion and protrusion (2, 5, 6, 17, 18). According to Lee et al. (12) and Bland and Altman (19), values above 0.75 of the lower bound of the ICC-95% CI were accepted as good, but values under 0.42 as inadequate.

The absolute range of motion in the ES was rated as being significantly lower by both examiners, with the exception of unassisted opening motion by examiner II. This is presumably due to the general lower mobility of the elderly subjects and is in the lower maximum of the range. Not only the absolute difference in millimetres between examiners, but also the percentage disagreement was calculated. A significant lower percentage disagreement between the examiners could be found in all metric measurements, except unassisted opening for the YS. One reason for this may be that the elderly could have impaired ability to perform and to reproduce movements, especially more difficult motions like protrusion or laterotrusion. This could explain the very high values of percentage disagreement for these movements, extending up to 44%, in comparison with 11% for 'easier' motions like opening. The exception of unassisted opening may be because this does not reach the maximum range of motion; it is limited when pain is felt. Therefore, it is not influenced by age-dependent abilities.

These findings could influence diagnosis based on the classification system of the RDC/TMD combining different clinical findings. For the metric measurements, important cut-off limits are decisive. The cut-off limits for jaw opening are taken as 35 or 40 mm; the limits for lateral excursions are taken as 7 mm (3). The elderly showed both a lower range of motion and worse reliability and therefore greater clinical measurement error. With unchanged cut-off limits, this could lead to wrong diagnoses, as a consequence not of pathologic findings, but a direct result of age.

### *Joint-related findings*

The $\kappa$ statistic depends on prevalence and could lead to misinterpretation of agreement when the prevalence is low or study groups are different, as seen in this study (20–23). In order to interpret these $\kappa$ values, the percentage agreement for both groups was calculated (5). In the present study, only the $\kappa$ values for opening motion and closing motion of ES reached an acceptable level. All other categorical values were worse. The great difficulties in achieving reliability in detecting joint sounds and the consequent large range of $\kappa$ values – from poor to excellent – has been found in several studies (2, 5, 7, 17). These difficulties in

inter-examiner agreement in these clinical findings have been discussed (5). The different findings in prevalence in the present study emphasize once again the difficulties in detecting joint sounds. There was no significant difference in reliability between the ES and the YS. Consequently, in the limitation of the unknown differences of clinical sign stability in the both groups, age seemed not to influence the detection of joint sounds.

### *Tender muscle points*

Tender muscle points are an important clinical symptom of TMD (2). The results in this study are inconsistent between the examiners, demonstrating a possible difference of clinical sign stability in the groups. Both examiners found lower absolute numbers of tender muscle points in the ES, but this was only significant for examiner I. This may be due to the medication taken by the elderly for multiple diseases. This could also explain that no subject in ES felt facial pain. To get an idea if reliability of tender muscle points is age dependent, the overall percentage agreement in subjects having at least one tender muscle point was calculated. It could be demonstrated that there was no significant difference in this parameter between the groups. A reliability analysis could only be feasible if the prevalence of an event reaches a certain level. Especially for examiner I, the prevalence found for the ES was very low. Thus, direct comparison of the reliability of the clinical appearance of tender muscles seemed not to be feasible within the limitations of this study.

## Conclusion

The age-dependent lower range of motion and the inferior reliability of metric measurements in the elderly could lead to wrong diagnoses. The reliability of detecting joint sounds and tender muscles was not age dependent within the limitations of the study.

## References


1. World Health Organization. Oral health surveys basic methods, 4th edn. Geneva: WHO; 1988.
2. Dworkin SF, Le Resche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: reliability of clinical examiners. J Prosthet Dent 1990;63:574–9.
3. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders. J Craniomandib Disord 1992;6:301–55.
4. Goulet JP, Clark GT, Flack VF. Reproducibility of examiners performance for muscle and joint palpation in the temporomandibular system following training and calibration. Community Dent Oral Epidemiol 1993;21:72–7.
5. John MT, Zwijnenburg AJ. Interobserver variability in assessment of signs of TMD. Int J Prosthet 2001;14:265–70.
6. Goulet JP, Clark GT, Flack VF, Lui 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. Heft MW. Prevalence of TMJ signs and symptoms in the elderly. Gerodontology 1984;3:125–30.
8. Oesterberg T, Carlsson GE, Wedel A, Johansson U. A cross-sectional and longitudinal study of craniomandibular dysfunction in an elderly population. J Craniomandib Disord 1992;6:237–45.
9. Ow RK, Loh T, Neo J, Khoo J. Symptoms of craniomandibular disorder among elderly people. J Oral Rehabil 1995;22:413–9.
10. Serfaty V, Nemconsky CE, Friedlander D, Gazit E. Functional disturbances of masticatory system in an elderly population group. Cranio 1989;7:46–51.
11. Salonen L, Hellden L, Carlsson G. Prevalence of signs and symptoms of dysfunction in the masticatory system: an epidemiologic Study in an adult Swedish population. J Craniomandib Disord 1990;4:241–50.
12. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989;19:61–70.
13. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
14. Magnusson T, Egermark I, Carlsson GE. A longitudinal epidemiologic study of signs and symptoms of temporomandibular disorders from 15 to 35 years of age. J Orofac Pain 2000;14:310–9.
15. Sano T, Widmalm SE, Westesson PL, Yamaga T, Yamamoto M, Takahashi K et al. Acoustic characteristics of sounds from temporomandibular joints with and without effusion: an MRI study. J Oral Rehabil 2002;29:161–6.
16. Widmalm SE, Williams WJ, Ang BK, McKay DC. Localization of TMJ sounds to side. J Oral Rehabil 2002;29:911–7.
17. List T, Dworkin SF. Comparing TMD diagnosis and clinical findings at Swedish and US TMD centers using Research Diagnostic Criteria for Temporomandibular Disorders. J Orofac Pain 1996;10:240–53.
18. Celic R, Jerolimov V, Knezovic Zlataric D, Klaic B. Measurement of mandibular movements in subjects with temporomandibular disorders and in asymptomatic subjects. Coll Antropol 2003;27:43–9.
19. Bland JM, Altman DG. A note on use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990;20:337–40.

20. Maclure M, Willet WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–9.
21. Guggenmoos-Holzmann I. The meaning of kappa: probabilistic concepts of reliability and validity revisited. J Clin Epidemiol 1996;49:775–82.
22. Rigby AS. Statistical methods in epidemiology. V. Towards an understanding of the kappa coefficient. Disabil Rehabil 2000;22:339–44.
23. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol 1988;41:949–58.