Clin Oral Invest (2012) 16:781–788
DOI 10.1007/s00784-011-0557-4

ORIGINAL ARTICLE

A new instrument for assessing the quality of studies on prevalence

Nikolaos Nikitas Giannakopoulos · Peter Rammelsberg · Lydia Eberhard · Marc Schmitter

Received: 27 September 2010 / Accepted: 18 April 2011 / Published online: 19 May 2011

Abstract There are numerous scientific articles of studies on the prevalence of disorders with non-standardised examination and diagnostic protocols. Because their quality is heterogeneous, a new instrument has been developed for the assessment of such studies. The new instrument is based mainly on statistical criteria. The points assigned for each of the main criteria according to the information gained from each paper are summed up to form a Total Quality Score (TQS). The interrater reliability of the instrument was tested by employing Kappa and Interrater Correlation Coefficient (ICC) statistics. The latter was assessed on the results of three independent investigators. The new quality instrument appeared to be easy to use, and the instructions were comprehensible. The $ICC_{(2,1)}$ for the TQS ranged between 0.94 and 1.00 indicating almost perfect agreement between the investigators. The reliability of the new instrument enables its use for scientific review purposes. In this way, its validity will also be tested. The instrument could be adopted for assessment of scientific articles of studies on the prevalence of disorders in many, similar, scientific areas.

Keywords Prevalence studies · Quality assessment · Reliability · Epidemiology

Introduction

In recent years, much effort has been expended in attempts to base scientific studies in medicine on better evidence and, thereby, to increase their quality and relevance. Organisations and individuals have set out to design principles and quality standards specifically for randomized controlled trials and systematic reviews. In the field of studies that investigate the prevalence of different disorders and diseases which do not yet have a single standardised method for them to be diagnosed or assessed with (e.g. temporomandibular disorders or TMDs [1], low back pain [2, 3], and shoulder pain [4, 5]), there are usually different methods and sample populations used. There is good reason to question the reliability of the results of these studies. The potential conclusions from the results are limited by methodological shortcomings. The great heterogeneity of such studies is the reason that the statistical combination of data derived from them seems impossible and should not be a prominent component of systematic reviews of observational studies [6]. Thus, great scientific effort remains useless, as no statistical combination or comparison can be conducted, and hence, no conclusions about the general picture can be drawn. Closer examination reveals that for disorders like TMDs, despite the huge volume of studies on their prevalence, it is not even possible to account for the prevalence of diagnostic subgroups or signs and symptoms in different populations across the globe, because the results from the studies cannot be reliably compared. The main reasons are lack of agreement about diagnostic criteria and examination procedures [1, 5], flaws in methodological information given in the papers, and lack of standardised qualitative assessment. An objective evaluation of the quality of prevalence studies could perhaps not directly enable the statistical combination of their data, but could, in any case, allow derivation of the impact of such a study for the clinicians, patients, and policy makers.

In order to raise the quality of observational studies, the "strengthening the reporting of observational studies in

N. N. Giannakopoulos (✉) · P. Rammelsberg · L. Eberhard · M. Schmitter
Department of Prosthodontics, University of Heidelberg,
Im Neuenheimer Feld 400,
69120 Heidelberg, Germany
e-mail: nik.giannakopoulos@med.uni-heidelberg.de

epidemiology" (STROBE) statement was developed [7]. The STROBE statement guidelines constitute a checklist which helps to improve the quality of reporting for observational studies. The STROBE checklist focuses on the reporting quality of studies on prevalence and is not an instrument to evaluate the quality of observational research [7]. Moreover, in a recent systematic review [8], 96 scales and checklists dealing with the assessment of the quality of observational studies were evaluated regarding their applicability. Only nine tools that had been created for studies of prevalence and incidence were found. A general remark was that the available tools did not discriminate poor reporting from the general quality of the studies and did not give separate conclusions about external and internal validity [8].

The purpose of this study was to develop a reliable instrument for qualitative assessment of the methodology of studies on the prevalence of disorders, based on strict epidemiological criteria. The latter could enable comparisons of the studies with equal quality properties described in articles with many precautions of course, leading to analyses that could help to clarify the epidemiological field of disorders with heterogeneous examination and diagnostic protocols.

Materials and methods

Scientific literature and statistical manuals were searched for properties that characterize high-quality prevalence studies in order to realize the objective of creating a disorder-oriented quality-assessment tool with a solid statistical basis. The main points considered can be assigned to three main criteria: sampling, measurement, and analysis [9].

Assessing the "sampling" literally means measuring the representative nature of the sample. Representativeness is a quality associated with the use of statistical sampling methods and careful evaluation of respondent characteristics. This criterion comprises three aspects: a clear definition of the target population, the sampling method used, and the match of respondents to the target group.

When evaluating the definition of a target group or sample for a prevalence study, the information considered most important was: age; sex; working conditions or hobbies; social, educational, or financial class; ethnicity; region of residence (urban, sub-urban, or rural), and relevant data from the health questionnaire of the sampled persons. Well-defined inclusion and exclusion criteria also ensure adequate description of the target population with regard to participants and non-participants in the study.

The term "sampling method" refers to the investigation of the way the sample was recruited, that is, whether or not probability methods were used. Probability sampling relies on the principle of randomisation to ensure that each eligible respondent has a known, most often equal, chance of selection. Thus, randomisation obviates the possibility of bias in the study. Successful randomisation enables valid statistical interpretation of "raw" results, that is, results unadjusted for other characteristics of the patients [10]. Probability sampling occurs in a variety of forms ranging from simple to complex (e.g. permuted block or cluster, stratified, multistage, multiphase, dynamic or adaptive randomisation) [9, 10].

To ensure that the characteristics of the respondents match those of the target population, two facts should be considered: response rate and description of the dropouts. Non-response is the failure to enlist sampled individuals; this can lead to selection bias and hence estimates that deviate systematically from population values. When information about non-respondents is available, methods to evaluate selection bias should be applied [9].

The approach to the second main criterion, "measurement", is by use of questions intended to clarify whether the survey yields reliable and valid measurements of the disorder. The instrument used for collecting data should be reliable and valid. Reliability establishes the extent to which an instrument can discriminate between individuals, and validity establishes the extent to which an instrument enables meaningful and useful discrimination between individuals. These are qualities that arise from the use of standardised data-collection methods and are confirmed empirically by measurement-evaluation studies [9]. Unfortunately, no commentary could be found on minimum validity standards for instruments used in prevalence studies. To guarantee validity, however, the instruments should satisfy basic standards of objectivity, specificity, and evidence. In the field of TMDs as an example, there are many examination procedures with different diagnostic criteria based on individual measurements, for example Krogh–Poulsen criteria [11], Helkimo Indexes [12], TMJ Scale [13–15], Craniomandibular Index [16, 17], Criteria of the American Academy of Orofacial Pain [18, 19], and Research Diagnostic Criteria of Temporomandibular Disorders (RDC/TMD) [20]. The examination and diagnostic procedure with the best statistical properties is currently the RDC/TMD, which has undergone many epidemiological examinations [21–24]. The reliability of the RDC/TMD can be increased if the examination is performed by examiners previously calibrated for the procedure [25, 26]. The same situation could apply to other disorders or pathologic entities with as yet no standardised or globally accepted examination and diagnostic protocol.

For the third criterion, "analysis", the scope is to examine the statistical procedure and outcomes. The reason for this is that special statistical methods are required to

obtain unbiased and precise estimates, especially when complex sampling procedures are used. Estimates in prevalence studies must, moreover, be accompanied by confidence intervals or the information needed to calculate them. Confidence intervals quantify the closeness of the unobserved value in the target population to the observed value in the sample, by telling us the chance that the value for an unobserved target population will fall within a certain range of the value for the observed sample [9].

To supplement the quality characteristics of a prevalence study, three additional criteria were added to those discussed above. Two of these, the "recruitment procedure" and the "statistical power of the sample", are statistical; the third is ethical. The recruitment procedure must be community (general population) based to lead to a representative sample, especially if the number of participants needed is large. This applies to surveys on disorders whose prevalence is not very high in the general population. The statistical power of the sample depends on its size, and to increase the precision of the data and to enable monitoring of disease trends over time, a large sample is needed, e.g. for disorders with a low prevalence in the general population, no fewer than 600 participants. The third quality criterion added is approval of the study by an ethics commission. It may be argued that this is a soft criterion and places older studies at a disadvantage. This is, however, an argument for future studies on humans to try to guarantee the protection of the individual.

To form our quality tool, all the aspects discussed above were considered in the form presented in Fig. 1 on the example of TMDs. A first part designed to collect general information about the reader, the article, and the journal, for statistical and archiving reasons, can also be added, but is not of relevance for the instrument. The part that is shown in Fig. 1 is structured according to the above-mentioned epidemiological criteria. The points assigned to each of the main criteria, according to the information gained from each paper, are summed up to form a Total Quality Score (TQS). Detailed guidelines about how to complete and assess each item of the tool accompany the form. This quality tool assesses prevalence studies according to their quality characteristics on a TQS scale of 0 to 19 points (very bad to outstanding, respectively).

To estimate the reliability of this new instrument, we assessed the interrater agreement, i.e. the consistency among different investigators at one point in time. The reliability was assessed in two phases. At a first pilot phase, the results obtained by two dentists who had independently evaluated the same ten articles on prevalence studies on TMDs with the quality tool were compared. None of them had known or ever worked with this procedure before. No calibration was performed beforehand. The raters received no instructions or help other than the guidelines attached to the worksheet and had no contact with each other regarding the procedure of the assessment. The choice of the articles was random from a pool of about 400 articles on prevalence of TMDs. The procedure was completed within 2 months after which the raters gave their feedback about the instrument and its applicability. The majority of the data derived from the questions of the quality instrument are nominal. Thus, in order to assess the agreement between the two investigators regarding each item of the instrument, Cohen's kappa coefficient was assessed [27]. For assessing the k value of the TQS, we divided the scale into four quality subgroups: 0–4 (poor), 5–9 (moderate), 10–14 (good), and 15–19 (outstanding). In a second phase, another three dentists employed the new instrument (slightly modified according to the feedback provided during the first phase) for the same ten articles under exactly the same conditions. The TQS represents continuous data (0–19), so in order to assess the interrater agreement, the ICC was assessed [28]. The selection of the raters was random, so the ICC was assessed according to an unadjusted two-way random model for single measures (ICC 2,1) tested for absolute agreement [28]. The articles employed for the assessment in both phases are reported in the references [29–38]. Statistical tests were performed by use of the Statistical Package for the Social Sciences 16.0 for Windows (SPSS Inc., Chicago, IL, USA).

Results

The time required for evaluating each of the journal articles depended on the length of the article; the average time was approximately 8 min per article. No significant difficulties in understanding the guidelines were mentioned during the assessment time, except for one of the explanations that was modified for the second phase. After testing the interrater agreement for each of the items, in the first phase, the mean of Cohen's k was calculated for the final TQS and for each question separately. For the TQS, the mean of the k value was 0.62 ± 0.15 indicating substantial agreement between the investigators [39]. The k values for the individual questions ranged between 0.26 (fair agreement) and 1.00 (almost perfect agreement) with a mean k value of 0.78 ± 0.27 which indicates a substantial agreement [39]. In the second phase, the $ICC_{(2,1)}$ for the TQS ranged between 0.94 and 1.00 indicating almost perfect agreement between the three raters.

Discussion

The reliability of this new research instrument, as determined by interrater testing, was high at least in the second

DATA EXTRACTION SHEET

1) Study characteristics (type of study): _____ *Questionnaire*: ___ *Clin.Exam*: ___

2) *Ethic commission approval*: _0_ No _1_ Yes

3) Verification of eligibility: _____

_____

4a) Inter-rater agreement (k/ICC-value): _____ **4b)** Enough? : _0_ k<0.60/ ICC<0.70 _1_ k≥0.60/ ICC≥0.70

5a) Confidence intervals: _____ **5b)** Enough? : _0_ _1_ CI≥90%

6) Population characteristics

6a) Target (source) population: _____

6b) Inclusion criteria: _____

6c) Exclusion criteria: _____

_____ **6d)** Proportion of target population? : _____

6e) Recruitment procedure: _____ 0 2

6f) Total population size: _____Sample: ____

6g) Response rate: _____ **6h1)** Enough? : _0_ _1_ _2_ RR≥70%(80%)

6h2) (if **0** or **1**) Is the rate taken into account? : _0_ No _1_ Yes

7) Participant's Characteristics								**9)** Prevalence	
7a) Age	**7b)**Sex	**7c)**Job/Hob.	**7d)**Class	**7e)**Ethnicity	**7f)**Region	**7g)**Anamn	**8)** Number	**9a)**	**9b)**
						8a) Total			
						8b) Power	0 2		

10) Quality assessment

10A. Sampling (*Representativity*)

1) Target pop. clearly defined? : _0_ <5 _1_ 5-9 _2_ >9

2) Probability sampling used? : _0_ No _1_ simple _2_ sophisticated

3) Do the respondents match the target? : ___ **B.6h1) +B.6h2)**

10B. Measurement (*Reliability*)

1) Standardised data-collection methods? : _0_ not identical _2_ identical

2) Reliable survey instruments? : _0_ noth. referred. _1_ B.4b) =1/RDC _2_ RDC+calibrated

3) Valid survey instruments? : _0_ questionnaire _1_ Helkimo etc. _2_ RDC/TMD

10C. Analysis

1) Were special features accounted for? : _0_ No/Not clear _1_ Yes

2) Satisfactory confidence intervals? : ___ **question B.5b)**

10D. QUALITY SCORE: Σ10A. +10B. +10C.

C. TOTAL QUALITY SCORE: **[B.2) +B.6e) +B.8b) +B.10D)] =**

Fig. 1 The new quality tool adopted for assessment of scientific articles of studies on the prevalence of disorders with heterogeneous examination and diagnostic protocols on the example of TMDs

round, indicating that the operational definitions were sufficiently precise. The only exception (k=0.26) during the first phase had to do with insufficient explanations given about probability sampling, which was accordingly corrected from the author after the first reliability test. After this modification, the instrument was again tested, and the high $ICC_{(2,1)}$ indicated the increase in the instrument's reliability. The use of different statistical methods in the two rounds may be discussed controversially, but one should consider the properties of the statistical procedures. Cohen's k is more appropriate for nominal data sets and agreement between two raters [40], whereas ICC is more appropriate for continuous data assessed by multiple raters [28]. Thus for the pilot phase, where the focus was set on the agreement on each item, the use of Cohen's k enabled the identification of the disagreement. In the second phase, the high $ICC_{(2,1)}$ indicated almost perfect agreement on the TQS which was interpreted as confirmatory for the positive impact of the modification.

Despite the fact that guidelines for assessment of the quality of prevalence studies have been published in medical journals [9], there seems to be no other tool for the assessment of the quality of papers reporting studies on the prevalence of disorders with heterogeneous examination and diagnostic protocols in a standardised and objective way. One of the first checklists developed for the evaluation of many different research articles appeared in 1994 [41]. This checklist was very detailed and not easy to apply, as it was constructed to be applicable for four different study types (from cross-sectional to case-control studies). Moreover, nothing is mentioned about its statistical properties. The checklist developed by Downs and Black [42] has been mainly decided for non-randomised studies of health interventions and not for population-based epidemiological studies. It also mixes reporting and methodological quality items and does not assess the statistical properties of the assessed studies in depth. In a systematic review that used this instrument for the quality assessment of population-based epidemiological studies [43], the authors added two new criteria in order to make it more complete. A further development of the checklist from Downs and Black was the epidemiological appraisal instrument (EAI) [44]. It consists of 43 items and was again constructed as a general purpose instrument for many different study designs. The answers to the items of the EAI were in a "yes–partial–no–not applicable" manner, and the statistical properties of this checklist have been reported to be very good. Nevertheless, it is not likely that this checklist is applicable for prevalence studies [8].

There has been much controversy on the use of quality scores in the area of clinical trials [45–48] and diagnostic accuracy studies [49]. The issue is yet not clear, and there are yet no data regarding the use of quality scores specifically for prevalence studies. The use of a numerical score for our new instrument may prove to have some drawbacks when employed for systematic reviews, but it nevertheless enables a clear, objective classification of the studies assessed, based on strict epidemiological criteria. On the other hand, the use of "yes–no–partial" answers would create a relatively big "grey zone" of "partially" good or bad studies, which does not help to clarify the general picture.

Another checklist constructed for evaluation of the quality of reporting of observational longitudinal research [50], alike to STROBE, focuses on the reporting quality of studies on prevalence and is not an instrument to evaluate the methodological quality of observational research. Focusing exclusively on reporting quality may not be enough for the qualitative assessment of a study, as reporting quality can differ from study quality [51]. Moreover, it is recommended to discriminate reporting vs. methodological quality of studies [8]. The latter issues are not yet completely clear regarding prevalence or cohort studies. The new instrument attempts to evaluate the methodological quality of articles on prevalence studies. The quality score of the article may not reflect the quality of the study, but the latter can usually not be tested. Nevertheless, the study results reach the scientific community mainly through scientific articles, and this underlines the need for good reporting practice. The newly developed instrument contains very few reporting items (such as "ethics commission approval") that are not subjective and should be considered as major quality markers also when methodological quality is being assessed.

The list developed from Nguyen et al. [52] aimed to assess the methodological quality of selected studies, but was not recommended for further use by its authors. For each of the 18 items of this list, a numerical score is given according to accompanying guidelines. Unfortunately, some of the items of the list are to be subjectively assessed or are not directly related to the quality of the study. Similar to the latter is a tool developed by Ariëns et al. [53], which intended to assess the methodological quality of observational studies, but the authors do not suggest its use for further studies despite the good agreement (84%) reported. Another remarkable tool for critically appraising studies of prevalence or incidence was developed by Loney et al. [54]. It is very short (eight items) and gives a quick overview of the studies. However, the questions do not go into much detail regarding the statistical properties of the assessed article. Very similar to the latter tool (actually a replication of it with one additional item) was the one reported by Woodbury et al. [55]. Both the last tools are indeed very much similar to ours, but none of them was constructed for broader use, and there is no information regarding their statistical properties.

There is, at last, another tool for standardising prevalence studies which should be mentioned, as it assesses the

usefulness of prevalence studies [56] and focuses mainly on their content. It consists of four questions which set conditions for proceeding to the next 15 questions and assesses, on a 100-point scale, the usefulness of a paper on a medical prevalence study, using hypertension as an example. If modified, it could theoretically be used for studies on the prevalence of other disorders too. We decided to develop a new tool, however, because the questions of the usefulness tool described in the paper of Silva et al. are very general and could not easily be adapted for studies on the prevalence of disorders with heterogeneous examination and diagnostic protocols. In other fields of medicine, for example hypertension which Silva also uses as an example, measurement procedures are standardised with well-established and globally accepted cut-off levels. The same cannot be claimed for TMDs, low back pain, or other similar pathological entities, so this new tool was created; this, in contrast with that of Silva et al., focuses on elementary morphological and content features of the prevalence studies. This approach could qualify the newly developed assessment tool to be further used for assessing papers on prevalence studies. A summary of the instruments discussed above can be found in Table 1. Often, in general medicine and in dentistry in any case, there are many papers on prevalence studies which are presenting different results. This new instrument combined with the STROBE statement guidelines could give deeper information on the quality of papers on prevalence studies and thus help in the direction of controlling and perhaps comparing such papers on a more stable basis.

A drawback of the new instrument is its "tailoring" to the most recent studies on the prevalence of disorders with heterogeneous examination and diagnostic protocols. Older studies may have an a priori disadvantage compared with newer studies, as for example almost no older studies refer to approval by an ethics commission. Another point of debate regarding the use of TMD prevalence studies as an example could also be the selection of the RDC/TMD as the ultimate examination procedure when other procedures also have very good statistical properties [14, 16]. The RDC/TMD has also been repeatedly criticized for not furnishing adequate statistical values for a few diagnoses [57–59]. The same problem will be faced when the instrument is used for other disorders with heterogeneous examination and diagnostic protocols, as no examination protocol has yet proven to have perfect statistical properties. Consequently, the new instrument should always be updated according to the best available evidence. The appearance of a new diagnostic or examination procedure (like the upcoming DC/TMD) with better statistical properties would necessitate readjustment of our instrument.

Conclusions

It is concluded that the newly developed tool for assessing the quality of scientific publications reporting studies on the prevalence of disorders with heterogeneous examination and diagnostic protocols has very good statistical properties with regard to interrater reliability. This enables it to be used further to assess the quality of scientific papers on prevalence studies. Its validity must, nevertheless, be proved further by using it for future literature reviews. This tool, slightly modified but based on the same principles, is designed to be used to assess the quality of articles about prevalence studies in different fields of medicine.

Table 1 Summary of potential instruments for quality assessment of prevalence studies

Instrument	Reliability	Validity	No. of items	Time needed	Applicability to the studies of incidence/prevalence
DuRant, 1994 [41]	N.r.	N.r.	18	N.r.	Yes [8]
Downs, 1998 [42]	r=0.88 / 0.75	SRTG used	26	N.r.	No [8]
Macfarlane, 2001 [43]	$k \geq 0.7$	N.r.	29	N.r.	Yes [8]
Genaidy, 2007 [44]	k=0.8–1.00	Int. consist. (0.83)	43	N.r.	Unlikely [8]
Tooth, 2005 [50]	75% agreement	N.r.	33	N.r.	Unlikely [8]
Nguyen, 1999 [52]	N.r.	N.r.	14	N.r.	Yes [8]
Ariëns, 2000 [53]	84% agreement	N.r.	18	N.r.	Yes [8]
Loney, 1998 [54]	N.r.	N.r.	8	N.r.	Unlikely [8]
Woodbury, 2004 [55]	N.r.	N.r.	9	N.r.	Yes [8]
Silva, 2001 [56]	N.r.	N.r.	19	N.r.	Yes
Giannakopoulos	ICC=0.94–1.0	N.r.	11	8 min.	Yes

Int. consist. internal consistency

N.r. nothing referred

Acknowledgements The authors of this article would like to acknowledge the dentists J. Mahabadi, A. Hassel, R. Shahin, and W. Bömicke for their help with assessment of the new instrument's reliability.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Gross AJ, Rivera-Morales WC, Gale EN (1988) A prevalence study of symptoms associated with TM disorders. J Craniomandib Disord 2:191–195
2. Cherkin DC, Deyo RA, Wheeler K, Ciol MA (1994) Physician variation in diagnostic testing for low back pain. Who you see is what you get. Arthritis Rheum 37:15–22
3. Koes BW, van Tulder MW, Ostelo R, Kim Burton A, Waddell G (2001) Clinical guidelines for the management of low back pain in primary care: an international comparison. Spine (Phila Pa 1976) 26:2504–2513, Discussion, 2513–2504
4. Bamji AN, Erhardt CC, Price TR, Williams PL (1996) The painful shoulder: can consultants agree? Br J Rheumatol 35:1172–1174
5. Luime JJ, Koes BW, Hendriksen IJ, Burdorf A, Verhagen AP, Miedema HS, Verhaar JA (2004) Prevalence and incidence of shoulder pain in the general population; a systematic review. Scand J Rheumatol 33:73–81
6. Egger M, Smith G, Altman D (2009) Systematic reviews in health care: meta-analysis in context, 2nd edn. BMJ Books, London
7. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370:1453–1457
8. Shamliyan T, Kane RL, Dickinson S (2010) A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol 63:1061–1070
9. Boyle MH (1998) Guidelines for evaluating prevalence studies. Evid Based Mental Health 1:37–39
10. Beller EM, Gebski V, Keech AC (2002) Randomisation in clinical trials. Med J Aust 177:565–567
11. Krogh-Poulsen WG (1969) Management of the occlusion of the teeth. Examination, diagnosis, treatment. In: Chayes LSCM (ed) Facial pain and mandibular dysfunction. Saunders, Philadelphia, pp 251–258
12. Helkimo M (1974) Studies on function and dysfunction of the masticatory system. Ii. Index for anamnestic and clinical dysfunction and occlusal state. Sven Tandläk Tidskr 67:101–121
13. Levitt SR, Lundeen TF, McKinney MW (1988) Initial studies of a new assessment method for temporomandibular joint disorders. J Prosthet Dent 59:490–495
14. Levitt SR, McKinney MW, Lundeen TF (1988) The TMJ scale: cross-validation and reliability studies. Cranio 6:17–25
15. Lundeen TF, Levitt SR, McKinney MW (1986) Discriminative ability of the TMJ scale: age and gender differences. J Prosthet Dent 56:84–92
16. Fricton JR, Schiffman EL (1986) Reliability of a craniomandibular index. J Dent Res 65:1359–1364
17. Fricton JR, Schiffman EL (1987) The craniomandibular index: validity. J Prosthet Dent 58:222–228
18. McNeill C (1993) Temporomandibular disorders-guidelines for classification, assessment and management, 2nd edn. Quintessence Publishing, Chicago
19. Okeson J (1996) Orofacial pain: guidelines for assessment, diagnosis, and management. Quintessence Publishing, Chicago
20. Dworkin SF, LeResche L (1992) Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord 6:301–355
21. Wahlund K, List T, Dworkin SF (1998) Temporomandibular disorders in children and adolescents: reliability of a questionnaire, clinical examination, and diagnosis. J Orofac Pain 12:42–51
22. Schmitter M, Kress B, Leckel M, Henschel V, Ohlmann B, Rammelsberg P (2008) Validity of temporomandibular disorder examination procedures for assessment of temporomandibular joint status. Am J Orthod Dentofacial Orthop 133:796–803
23. Dworkin SF, Sherman J, Mancl L, Ohrbach R, LeResche L, Truelove E (2002) Reliability, validity, and clinical utility of the research diagnostic criteria for Temporomandibular Disorders Axis II Scales: depression, non-specific physical symptoms, and graded chronic pain. J Orofac Pain 16:207–220
24. John MT, Dworkin SF, Mancl LA (2005) Reliability of clinical temporomandibular disorder diagnoses. Pain 118:61–69
25. Schmitter M, Ohlmann B, John MT, Hirsch C, Rammelsberg P (2005) Research diagnostic criteria for temporomandibular disorders: a calibration and reliability study. Cranio 23:212–218
26. List T, John MT, Dworkin SF, Svensson P (2006) Recalibration improves inter-examiner reliability of tmd examination. Acta Odontol Scand 64:146–152
27. Cohen J (1960) A coeffiecient of agreement for nominal scales. Educ Psychol Meas 20:37–46
28. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428
29. Aggarwal VR, Macfarlane TV, Macfarlane GJ (2003) Why is pain more common amongst people living in areas of low socio-economic status? A population-based cross-sectional study. Br Dent J 194:383–387, Discussion, 380
30. Dworkin SF, Huggins KH, LeResche L, Von Korff M, Howard J, Truelove E, Sommers E (1990) Epidemiology of signs and symptoms in temporomandibular disorders: clinical signs in cases and controls. J Am Dent Assoc 120:273–281
31. Heikinheimo K, Salmi K, Myllarniemi S, Kirveskari P (1989) Symptoms of craniomandibular disorder in a sample of Finnish adolescents at the ages of 12 and 15 years. Eur J Orthod 11:325–331
32. Locker D, Grushka M (1987) The impact of dental and facial pain. J Dent Res 66:1414–1417
33. Nilsson IM, List T, Drangsholt M (2005) Prevalence of temporomandibular pain and subsequent dental treatment in Swedish adolescents. J Orofac Pain 19:144–150
34. Pedroni CR, De Oliveira AS, Guaratini MI (2003) Prevalence study of signs and symptoms of temporomandibular disorders in university students. J Oral Rehabil 30:283–289
35. Rantala MA, Ahlberg J, Suvinen TI, Savolainen A, Kononen M (2004) Chronic myofascial pain, disk displacement with reduction and psychosocial factors in finnish non-patients. Acta Odontol Scand 62:293–297
36. Raustia AM, Peltola M, Salonen MA (1997) Influence of complete denture renewal on craniomandibular disorders: a 1-year follow-up study. J Oral Rehabil 24:30–36
37. Sadowsky C, Muhl ZF, Sakols EI, Sommerville JM (1985) Temporomandibular joint sounds related to orthodontic therapy. J Dent Res 64:1392–1395
38. Tallents RH, Hatala M, Katzberg RW, Westesson PL (1993) Temporomandibular joint sounds in asymptomatic volunteers. J Prosthet Dent 69:298–304
39. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
40. Schouten H (1986) Nominal scale agreement among observers. Psychometrika 51:453–466
41. DuRant R (1994) Checklist for the evaluation of research articles. J Adolesc Health 15:4–8

42. Downs SH, Black N (1998) The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 52:377–384
43. Macfarlane TV, Glenny AM, Worthington HV (2001) Systematic review of population-based epidemiological studies of oro-facial pain. J Dent 29:451–467
44. Genaidy A, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, Dunning K (2007) An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. Ergonomics 50:920–960
45. Juni P, Altman D, Egger M (2001) Assessing the quality of controlled trials. BMJ 323:42–46
46. Berlin J, Rennie D (1999) Measuring the quality of trials: the quality of quality scales. JAMA 282:1083–1085
47. Greenland S (1994) Quality scores are useless and potentially misleading. Am J Epidemiol 140:300–302
48. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352:609–613
49. Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5:19–27
50. Tooth L, Ware R, Bain C, Purdie DM, Dobson A (2005) Quality of reporting of observational longitudinal research. Am J Epidemiol 161:280–288
51. Soares HP, Daniels S, Kumar A, Clarke M, Scott C, Swann S, Djulbegovic B (2004) Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the radiation therapy oncology group. BMJ 328:22–24
52. Nguyen Q, Bezemer P, Habets L, Prahl-Andersen B (1999) A systematic review of the relationship between overjet size and traumatic dental injuries. Eur J Orthod 21:503–515
53. Ariëns G, van Mechelen W, Bongers P, Bouter L, van der Wal G (2000) Physical risk factors for neck pain. Scand J Work Environ Health 26:7–19
54. Loney P, Chambers LW, Benett KJ, Roberts JG, Stratford PW (1998) Critical appraisal of the health research literature: prevalence or incidence of a health problem. Chronic Dis Can 19:170–176
55. Woodbury M, Houghton PE (2004) Prevalence of pressure ulcers in Canadian healthcare settings. Ostomy/Wound Manage 50:22–38
56. Silva LC, Ordunez P, Paz Rodriguez M, Robles S (2001) A tool for assessing the usefulness of prevalence studies done for surveillance purposes: the example of hypertension. Rev Panam Salud Pública 10:152–160
57. Barclay P, Hollender LG, Maravilla KR, Truelove EL (1999) Comparison of clinical and magnetic resonance imaging diagnosis in patients with disk displacement in the temporomandibular joint. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 88:37–43
58. Schmitter M, Kress B, Rammelsberg P (2004) Temporomandibular joint pathosis in patients with myofascial pain: a comparative analysis of magnetic resonance imaging and a clinical examination based on a specific set of criteria. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 97:318–324
59. Emshoff R, Rudisch A (2001) Validity of clinical diagnostic criteria for temporomandibular disorders: clinical versus magnetic resonance imaging diagnosis of temporomandibular joint internal derangement and osteoarthrosis. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 91:50–55