# Critical Appraisal of Methods Used in Randomized Controlled Trials of Treatments for Temporomandibular Disorders

**James R. Fricton, DDS, MS**
Professor
University of Minnesota School of Dentistry
Minneapolis, Minnesota

**Wei Ouyang, DDS, PhD**
Assistant Professor
Renmin University of China
Beijing, China, and
former Research Assistant
University of Minnesota School of Dentistry
Minneapolis, Minnesota

**Donald R. Nixdorf, DDS, MS**
Assistant Professor
University of Minnesota
Minneapolis, Minnesota

**Eric L. Schiffman, DDS, MS**
Associate Professor
University of Minnesota School of Dentistry
Minneapolis, Minnesota

**Ana Miriam Velly, DDS, PhD**
Research Associate
University of Minnesota School of Dentistry
Minneapolis, Minnesota

**John O. Look, DDS, PhD**
Senior Research Associate
University of Minnesota School of Dentistry
Minneapolis, Minnesota

**Correspondence to:**
Dr James Fricton
Professor
School of Dentistry
University of Minnesota
515 Delaware St. SE
6-320 Moos
Minneapolis, MN 55455

***Aims:*** *To evaluate the quality of methods used in randomized controlled trials (RCTs) of treatments for management of pain and dysfunction associated with temporomandibular muscle and joint disorders (TMJD) and to discuss the implications for future RCTs.* ***Methods:*** *A systematic review was made of RCTs that were implemented from 1966 through March 2006, to evaluate six types of treatments for TMJD: orthopedic appliances, occlusal therapy, physical medicine modalities, pharmacologic therapy, cognitive-behavioral and psychological therapy, and temporomandibular joint surgery. A quality assessment of 210 published RCTs assessing the internal and external validity of these RCTs was conducted using the Consolidated Standards of Reporting Trials (CONSORT) criteria adapted to the methods of the studies.* ***Results:*** *Independent assessments by raters demonstrated consistency with a mean intraclass correlation coefficient of 0.63 (95% confidence interval). The mean percent of criteria met was 58%, with only 10% of the RCTs meeting the four most important criteria.* ***Conclusions:*** *Much of the evidence base for TMJD treatments may be susceptible to systematic bias and most past studies should be interpreted with caution. However, a scatter plot of RCT quality versus year of publication shows improvement in RCT quality over time, suggesting that future studies may continue to improve methods that minimize bias.* J OROFAC PAIN 2010;24:139–151

**Key words:** quality, randomized clinical trials, temporomandibular, tension type headache, TMD, TMJ

Many diverse treatments have been evaluated for their efficacy in the management of pain and dysfunction associated with temporomandibular muscle and joint disorders (TMJD). Most TMJD treatments have been tested by means of randomized controlled trials (RCTs). However, these trials have demonstrated a wide range of results that may be due to varying methods and study designs.[1–5] With diverse results, it is difficult for health-care providers to infer reliable and precise evidence-based clinical treatment guidelines from reviews of these RCTs. When systematic reviews and meta-analyses have been performed using low-quality studies, they have generally shown a greater magnitude of treatment benefit than those using higher-quality studies, thereby overestimating the efficacy of the intervention.[6–9] There is also evidence that significant variation exists within the academic community over the use of appropriate standards for analyzing and reporting the results of sponsored clinical research, especially clinical trials sponsored by industry.[10] In order to improve the consistency of reporting on RCTs, a collaboration of

clinical epidemiologists, biostatisticians, and journal editors published a statement called CONSORT (Consolidation of the Standards of Reporting Trials).[9,11–13] The CONSORT statement includes criteria for use by authors, journal editors and peer reviewers to assess the quality of a report on a RCT.

## Assessment of Quality

The design of a credible RCT for TMJD is feasible.[7,8,14] Appropriate reporting of these studies is also a reasonable expectation and the goal of the CONSORT criteria.[9,11] Many RCTs for TMJD continue to be weak, in part, because the investigators may be unaware or find it difficult to implement fundamental design criteria that are essential for all RCTs. RCTs should be designed to minimize bias, with bias being defined as any potential distortion of a study result such that it differs systematically from the true effect estimate of efficacy.[14] It is possible that essential design methods were implemented in some of the reviewed studies, but these procedures were not reported.[14] Other less obvious design problems relate to criteria that are important particularly for TMJD studies in which pain is the most important outcome to measure. For example, given the subjective nature of pain, assessment by examiners blinded to the intervention is essential.[1,4] Reliable, valid, and standardized measures should also be employed that take into account the multidimensional nature of pain.[15–17] Attention must be given to floor and ceiling effects associated with pain measures.[18] Little or no reported pain at baseline creates a ceiling effect where no clinically significant improvement with treatment will be detected.[19]

### Quality Assessment Criteria

A number of criteria have been proposed in the past to define and critically assess the quality of RCTs and determine the degree of bias that may influence RCT results.[11,19–25] Jadad[20] has suggested that the following five critical aspects be considered in assessing the quality of a clinical trial:

- The clinical relevance of the research question.
- The internal validity of the trial, that is, the degree to which biased results have been addressed by the study design, study sample, and methods, including the outcome measures.
- The external validity, that is, the precision and extent to which it is possible to generalize results to other populations and settings.
- The appropriateness of data analysis and presentation.
- The ethical implications of the intervention being evaluated.

The Oxford Centre for Evidence-based Medicine (www.cebm.net)[26] has also developed a set of criteria for RCTs. This collaboration by a group of international scientists has also promoted evidence-based reviews for use in making clinical decisions. They suggested the following questions for critical appraisal of an RCT quality:

- Did the trial address a clearly focused issue?
- Is a trial an appropriate method to answer this issue?
- Did the study have enough participants to minimize the play of chance? Was a sample size calculation done?
- Was the assignment of patients to groups randomized, and was the randomization concealed?
- Were reliable and valid outcome measures of known or probable clinical importance measured for at least 80% of participants who entered the trial?
- Were all of the patients who entered the trial accounted for in the end?
- Were the patients analyzed in the groups to which they were assigned, ie, an intention-to-treat analysis?
- Were the patients and rater/observer "blinded" to which treatment was being received when possible?
- Aside from the experimental treatment, were the groups treated equally?
- Were the groups similar at the start of the trial?

### CONSORT Criteria

It is not possible to assess either the internal or the external validity of a study if the quality of reporting is inadequate. In order to improve the consistency of reporting on RCTs, the CONSORT statement was developed for two-group parallel RCT designs.[9,11,12] This statement includes a checklist and a flow diagram for use by authors, journal editors, and peer reviewers to assess the quality of a report on a randomized trial. The CONSORT statement has been endorsed by a number of major journals such as the *British Medical Journal*, *The Lancet*, the *Journal of the American Medical Association*, *Journal of Dental Research,* and the *Canadian Medical Association Journal.* In addition to the CONSORT statement, the American Association of Medical Colleges, in collaboration

with the Centers for Education and Research in Therapeutics and the Blue Cross Blue Shield Association, has developed a set of principles, recommendations, and guidelines, rooted in sound science and sound ethics, to guide researchers in conducting and publishing clinical research.[27]

### Composite Scales

Some authors of systematic reviews have proposed composite scales to assess the relative quality of the reviewed studies. This investigator-based mathematical assessment is designed to assign differential quality weighting to study results that may, based on a quality cutoff, exclude some studies from consideration. There has been a wide range of such assessment instruments reported in the literature to evaluate the quality of RCTs for medical interventions, including a number of multi-item composite scales.[19,23–25] Composite assessment scales can be problematic, however, resulting in discordant findings.[14,23] The discriminating power of composite scales may also be reduced by the inclusion of criteria that are not necessarily the most appropriate. Even for the most commonly accepted design criteria, there are differences of opinion as to their relative value, as demonstrated by the different weights accorded to them in the composite scales. Since the composite quality score may mask a study's quality with respect to certain essential criteria, there is a need to consider these individual criteria separately as well as consider the percent of criteria met to derive a relative quality score.

Based on an initiative by the American Academy of Orofacial Pain (AAOP), a systematic review was conducted to assess the evidence base for TMJD treatments. Two hundred and ten RCTs for TMJD were identified as having been implemented from 1966 up through February, 2006. This review provided an opportunity to use the CONSORT criteria and conduct a methodological quality assessment of the reviewed RCTs to help determine reasons for the diversity of results that have been attributed to similar treatments, and the potential for Type I (false positive) and Type II (false negative) errors.

Thus, the aims of this study were: *(1)* to review the quality of the published methods for the RCTs of TMJD treatments, *(2)* to report the overall quality of the clinical trials by using the CONSORT criteria, *(3)* discuss the effects of inadequate application of quality methods in diverse results of past clinical trials, and *(4)* discuss the mathematical methods employed for the future series of articles synthesizing the meta-analyses.

## Materials and Methods

Given the need to standardize critical review methodology to the greatest extent possible, and the wide acceptance accorded to the CONSORT statement, the authors selected the CONSORT criteria that address study design issues as the basis for reviewing the quality of these RCTs. Table 1 shows the parallels between study design requirements and reporting guidelines, and presents the rationale for using the CONSORT criteria to assess TMJD RCTs.

### Interventions for TMJD and Outcome Variables

The RCTs for TMJD were divided into six general treatment types: *(1)* orthopedic appliances, *(2)* occlusal therapy, *(3)* physical medicine procedures, *(4)* pharmacologic therapy, *(5)* cognitive-behavioral and psychological therapy, and *(6)* temporomandibular joint (TMJ) surgery. Within these categories, 46 different treatment protocols for TMJD were identified.

Pain was the most common outcome measured in these RCTs, and this was the outcome selected for the mathematical syntheses in this review series. It is acknowledged, however, that other measures are also important such as functional status, disability, morbidity, quality of life, and cost of treatment, as well as other clinical parameters such as jaw function, palpation tenderness, and range of motion.

### Searching and Information Retrieval

A MEDLINE search strategy was developed to include RCT published during the years 1966 through March 2006 (Table 2), and implemented on the PubMed interface for MEDLINE at the US National Library of Medicine. The search strategy was based on the recommendations of the US Agency for Health Care Policy, the Cochrane Collaboration, and the Centre for Reviews & Dissemination, University of York, United Kingdom. It also included a manual hand search of references in each article and in systematic reviews and a review for duplicates.

The QUOROM (Quality of Reporting of Meta-analysis) statement includes guidelines produced to improve the quality and reporting of systematic reviews. Figure 1 presents the QUOROM diagram on how trials were excluded from the meta-analysis. This quality review began with 396 studies with 186 excluded as not being RCTs or were duplicates resulting in 210 trials being retained for

**Table 1 CONSORT Criteria for Reporting RCTs Used for Measuring Quality of TMJD RCT Methodology**

| CONSORT criteria | Item | Descriptor | Interpretation of criteria in quality assessment for TMJD RCTs |
|---|---|---|---|
| Title and abstract | 1 | The title and abstract acknowledge the randomization process | Reporting criteria not applied |
| Background | 2 | Scientific background and explanation of rationale | Reporting criteria not applied |
| Participants | 3 | Eligibility criteria for participants and the settings and locations where the data were collected | Full credit is assigned when the recruitment source of subjects is clear with defined inclusion and exclusion criteria so generalizability of study results to another population is possible |
| Interventions | 4 | Precise details of the interventions intended for each group and how and when they were actually administered | Full credit is assigned when treatments in each group are well defined, standardized, and comparable between subjects and conforming, when possible, to an accepted standard of care |
| Objectives | 5 | Specific objectives and hypotheses included | Reporting criteria not applied |
| Outcomes | 6 | Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (eg, multiple observations, training of assessors) | Full credit is assigned if primary outcomes are relevant to the condition and when appropriate multidimensional, including both subjective and objective parameters of treatment success are used at pre- and postintervention with calibrated examiners |
| Sample size | 7 | How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules | Full credit is assigned if the criteria for sample size calculation are described, and if the study design has adequate safeguards to avoid Type I error (false positive conclusions) and Type II error (false negative conclusions) |
| Randomization / sequence generation | 8 | Method used to generate the random allocation sequence, including details of any restrictions (eg, blocking, stratification) | Full credit is assigned if the process for randomization of subjects into treatment groups is defined, and the treatment assignment is adequately concealed and implemented to minimize bias from influencing treatment assignment |
| Randomization / allocation concealment | 9 | Method used to implement the random allocation sequence (eg, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned | Included in Criteria 8 |
| Randomization / implementation | 10 | Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups | Included in Criteria 8 |
| Blinding (masking) | 11 | Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated | Full credit is assigned when measurement or detection of outcome is performed by an observer or rater who is blinded to treatment assignment |
| Statistical methods | 12 | Statistical methods used to compare groups for primary outcome(s); Methods for additional analyses, such as subgroup analyses and adjusted analyses | Full credit is assigned if a complete and appropriate statistical analysis is conducted respecting intention to treat principles. In these analyses, point estimates for the treatment effects and their confidence intervals should be estimated. In addition, known prognostic factors at baseline should be compared between study groups. Finally, primary and subgroup analyses should be adjusted for baseline differences if differences between groups were observed |
| Results of participant flow | 13 | Flow of participants through each stage (use of a diagram is strongly recommended). Specifically, for each group report, the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons | Full credit is provided if the flow of patients including the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome is included |
| Recruitment and follow-up | 14 | Dates defining the periods of recruitment and follow-up | Full credit is assigned when the participant recruitment and follow-up schedule is described with adequate time allowed for the detection of differences in outcome measures |
| Baseline data | 15 | Baseline demographic and clinical characteristics of each group are presented | Full credit is assigned when the baseline comparison group is used to assess the appropriateness of randomization, controlling for possible confounders |
| Numbers analyzed | 16 | Number of participants (denominator) in each group included in each analysis and whether the analysis was by "intention-to-treat." State the results in absolute numbers when feasible (eg, 10/20, not 50%). | Full credit is assigned if the total numbers of subjects included in the analyses are the subjects randomized, respecting the intent to treat principles. Number of subjects withdrawals, dropout and crossover need to be described |

**Table 1 CONSORT Criteria for Reporting RCTs was Used for Measuring Quality of TMJD RCT Methodology (continued)**

| CONSORT criteria | Item | Descriptor | Interpretation of criteria in quality assessment for TMJD RCTs |
|---|---|---|---|
| Outcomes and estimation | 17 | For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (eg, 95% confidence interval) | Included in Criteria 12 |
| Ancillary analyses | 18 | Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory | Included in Criteria 12 |
| Adverse events | 19 | All important adverse events or side effects in each intervention group | Reporting criteria not applied |
| Interpretation | 20 | Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision, and the dangers associated with multiplicity of analyses and outcomes | Reporting criteria not applied |
| Generalizability | 21 | Generalizability (external validity) of the trial findings | Reporting criteria not applied separately and is considered* by population and design issues |
| Overall evidence | 22 | General interpretation of the results in the context of current evidence | Reporting criteria not applied |

*This criterion (generalizability) is not evaluated separately but is included in other criteria including item 2 Participants, item 6 Outcomes, item 4 Interventions, and to some extent other criteria since it is dependent on these other criteria.

**Table 2 PubMed Search Strategy for Treatment of TMD***

| Database: MEDLINE 1966 to March 2006 | | |
|---|---|---|
| 1 | exp temporomandibular joint disorders/ or temporomandibular joint disorders.mp. | 8,937 |
| 2 | tmj.tw. | 3,035 |
| 3 | (myofascial adj3 pain).tw. | 591 |
| 4 | *craniomandibular disorders/ or craniomandibular disorders.mp. | 427 |
| 5 | exp myofascial pain syndromes/ or myofascial pain syndrome.mp. | 4,541 |
| 6 | 1 or 2 or 3 or 4 or 5 | 10,485 |
| 7 | limit 6 to (human and english language) | 7,358 |
| 8 | exp Temporomandibular joint/ | 6,619 |
| 9 | exp tension headache/ or "tension headache".mp. | 796 |
| 10 | psychogenic headache.mp. | 37 |
| 11 | chronic daily headache.mp | 151 |
| 12 | (chronic adj3 headache).mp. | 1,203 |
| 13 | muscle headache.mp. | 13 |
| 14 | (muscular adj3 headache).mp. | 57 |
| 15 | (muscle adj3 headache).mp | 320 |
| 16 | or/9-15 | 2,080 |
| 17 | exp facial pain/ or "facial pain".mp | 4,089 |
| 18 | "orofacial pain".mp. | 329 |
| 19 | myofacial pain.mp. | 56 |
| 20 | 7 or 8 or 16 or 17 or 18 or 19 | 17,063 |
| 21 | limit 20 to human and english language | 15,901 |
| 22 | limit 21 to randomized controlled trial | 396 |

*The search can be rerun at the reader's convenience at PubMed at the Internet uniform resource locator http://www.ncbi.nlm.nih.gov/PubMed/

the quality assessment process including 17 published abstracts. For the meta-analysis, studies were excluded because they were abstracts with insufficient methods described (17), had incompatible comparison groups (124), had incomplete data (8), had incompatible follow-up duration (6), and had incompatible outcome measures (5). This process resulted in 50 RCTs that were finally included in the meta-analyses. Although nonrandomized controlled trials, observational cohort studies, and abstracts are also important sources of clinical information, nonrandomized studies were not included in this search strategy because of their potential for introducing biased estimates of

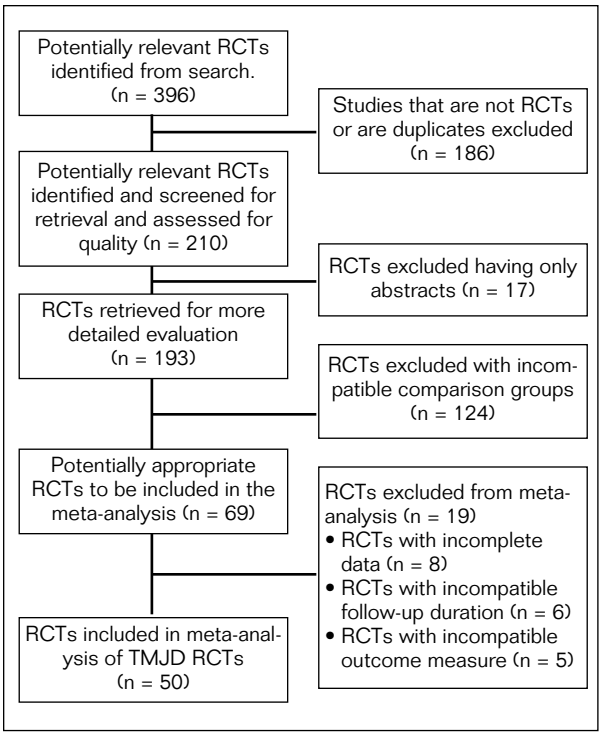


**Fig 1** QUOROM diagram for inclusion and exclusion of studies in the meta-analysis of all TMJD interventions.[37]

results. Egger et al concluded that an in-depth critical review of the methodology and quality of available articles may be better than including "gray" literature such as abstracts which is usually of lower quality.[28]

It was necessary to include a variety of temporomandibular search terms broad enough to include both joint and muscle disorders since most past studies have not differentiated between these different conditions. Not only was there a lack of diagnostic differentiation, but occasionally there was also ambiguity as to the diagnostic terms employed. For example, Ekberg et al[29] selected myofascial pain subjects for a clinical trial based on the Research Diagnostic Criteria for Temporo - mandibular Disorders (RDC/TMD) criteria.[30] Accordingly, they excluded subjects with TMJ pain that could be verified by interview or examination. But, in a review of stabilization splint therapy,[4] the criteria listed for a diagnosis of myofascial pain included not only muscle pain but also TMJ pain on palpation. The issue is that muscle versus joint conditions in TMJD can have different outcomes associated with a given intervention.[1] Thus, the study populations need not only to be accurately characterized as to the diagnoses present, but they must also be described using terms that have the same meaning for most readers. In addition, a diagnosis of tension-type headache (TTHA) was also included in this review for the following reasons:

1. The International Headache Society diagnostic criteria for TTHA with pericranial muscle tenderness and the AAOP diagnostic criteria for temporalis myofascial pain are overlapping sufficiently to suggest that they are the same condition.[1,15,31,32]
2. TTHA is a common symptom of the masticatory myofascial pain component of TMJD.[30,33,34]
3. Some treatments used for TTHA are also used for TMJD myofascial pain involving the temporalis muscle.[33]

### Point of Departure for Using CONSORT Criteria Guidelines

The quality of a study design is not synonymous with the quality of reporting as defined by the CONSORT statement.[9,11] However, most of the reporting criteria addressed in the CONSORT statement correspond directly to items that define the quality of the study design (Table 1). The only exceptions to this would be Items 1, 2, 5, 19, 20, 21, and 22. Since this study focused on quality of methods, these reporting criteria are important for understanding the rationale for a study but its generalizability was not assessed. Item 1 relates to the title and abstract; item 2 describes the background and rationale for the study; item 5 reports the objectives of study; item 19 reports adverse events; item 20 includes the discussion of results in light of the study hypotheses and potential systematic and random error; item 21 discusses generalizability of results; and item 22, the discussion of results in light of the current evidence. For simplicity of reviewing and presenting the criteria, items 8, 9, and 10 were combined as part of the randomization process and items 12, 17, and 18 were combined as part of the statistical analysis process. Discussing generalizability of the study is an important quality criteria that is dependent on the population and study design characteristics. Since it is difficult to score independently, it was not scored as part of the quality score.

### Evaluation of Critical Study Design Criteria to Minimize Systematic Bias

There are four internal validity design criteria as defined by the CONSORT criteria that have been

shown to provide a unique differentiation between well-designed and poorly-designed studies and are considered Level I criteria. They include randomization process (8–10), blinding of outcome methods (11), comparable groups (15), and handling of withdrawals or dropouts in the data analysis (16). These same criteria are also the most commonly recommended in quality assessment scales and are the same as Jadad et al[35] and Moher et al.[23] These authors suggested that these criteria could conveniently serve as a generic quality assessment scale for all RCTs, and thus were included in each review. This recommendation is also consistent with the content of Sections 6.2 to 6.6 in the Cochrane Handbook for Systematic Reviews of Interventions 4.2.5.[36,37]

### Interpretation of CONSORT Criteria for This Critical Appraisal of RCTs for TMJD

Applying the CONSORT criteria to evaluation of quality of RCTs is a process that is different for different disorders or interventions being studied. For RCTs of TMJD, specific criteria for outcomes, subjects, subject flow, and statistical analysis are unique and need to be defined *a priori*. For this reason, a description of how the criteria were used to score whether a study met or did not meet the criteria is included. Table 1 presents the rules for interpreting the CONSORT criteria for TMJD RCTs.

### Methods for Assessment of Rater Reliability

*Internal Reliability*. Two raters of study quality, one an expert in TMJD and orofacial pain and the other a PhD candidate in Health Services Research, first discussed and arrived at a common understanding on the definition and the threshold for meeting each of the quality criteria above. A series of 15 articles were then selected from a broad range of treatment modalities. The scoring system was simple: 0 = criterion not met, and 1 = criterion met. The overall composite score (0 – 1) was the proportion of criteria met (# criteria met/15). The two raters each scored the same articles while blind to the other's scores, and their scores were compared using the intraclass correlation coefficient (ICC) as the measure of agreement.

### External Agreement Appraisal

In order to compare the scoring of these raters to scores previously published, the selection of articles for reliability testing included 14 of the articles that were reviewed by Forssell and Kalso.[1] The two sets of scores were compared using the ICC as the measure of agreement.

### RCT Quality Appraisal Methods

The quality assessment criteria were applied to each RCT in a two-step process. First to be reviewed were the Level I criteria for minimizing systematic bias, and a notation was made as to whether these four criteria had been met. Second, each study was evaluated as to whether it met 11 remaining criteria. Finally, a quality assessment score was calculated to reflect the percentage of all 15 criteria that was met for each study, thus, permitting an overall estimate of the quality of the evidence base for the treatment of TMJD.

### Distribution of Quality Scores

In order to assess improvement in trial methodology over time, a scatter plot analysis was performed for the quality scores of the 210 TMJ studies that qualified as published RCTs.

## Results

Figure 1 presents the flow of studies that were included in the meta-analysis. Three hundred and ninety-six (396) studies were reviewed by the reviewers. Following the exclusion of studies that did not meet the definition of a randomized trial, 210 studies remained that were used for quality assessment.

### Interrater Reliability and Concurrent Validity Estimates

The mean ICC for interrater reliability was 0.63 (95% confidence interval), with a range from 0.59 to 0.67 for each type of TMJD treatment. This level of reliability was seen as adequate evidence that this proposed set of criteria could be consistently interpreted and applied for systematic reviews. Comparing the scoring by this study's raters to Forssell and Kalso's results[1] for the same studies, the results showed significant agreement with an ICC of 0.59 and $P < .009$. The confidence intervals ranged from 12.5% to 84.5%.

### Percent of Criteria Met by Studies for Each Treatment Group

Table 3 presents the means and ranges for the percent of criteria met by the RCTs grouped according

**Table 3 The Number, Percent, and Range of Studies Meeting CONSORT Methods Criteria for Level I and All Criteria by Treatment Type**

| Treatment | No. of RCTs | % Level I criteria met | Range of % for all criteria met | Mean ± SD % for all critera |
|---|---|---|---|---|
| Physical medicine/ exercise/injections | 80 | 14% (n = 11) | 20 to 93 | 58% ± 17 |
| Orthopedic appliances | 46 | 13% (n = 6) | 27 to 87 | 53% ± 15 |
| Pharmacologic therapy | 44 | 9% (n = 4) | 27 to 87 | 67% ± 14 |
| Cognitive-behavioral/ psychological therapy | 24 | 8% (n = 2) | 33 to 80 | 54% ± 12 |
| TMJ surgery | 7 | 14% (n = 1) | 33 to 100 | 55% ± 22 |
| Occlusal therapy | 9 | 11% (n = 1) | 33 to 67 | 45% ± 11 |
| Totals/mean | 210 | 12% (n = 25) | 33 to 100 | 58% ± 15 |

**Table 4 The Percent of Studies That Met Each of the CONSORT Criteria That Reflect Study Design and Methods Issues Versus Reporting and Interpretation Issues**

| CONSORT criteria | No. of studies | 3 | 4 | 6 | 7 | 8,9,10* | 11* | 12,17,18 | 13 | 14 | 15* | 16* | Mean % of criteria met |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Physical medicine/ exercise/injections | 80 | 91 | 93 | 95 | 16 | 9 | 45 | 41 | 9 | 93 | 91 | 43 | 55 |
| Orthopedic appliances | 46 | 98 | 85 | 63 | 20 | 13 | 93 | 75 | 40 | 65 | 100 | 60 | 67 |
| Pharmacologic therapy | 44 | 76 | 92 | 84 | 4 | 0 | 16 | 60 | 12 | 92 | 96 | 56 | 54 |
| Behavioral/psychological therapy | 24 | 86 | 96 | 72 | 11 | 23 | 72 | 55 | 8 | 72 | 95 | 34 | 58 |
| TMJ surgery | 7 | 86 | 71 | 100 | 14 | 14 | 29 | 29 | 14 | 100 | 71 | 57 | 55 |
| Occlusal therapy | 9 | 78 | 56 | 78 | 0 | 22 | 44 | 11 | 11 | 89 | 67 | 67 | 45 |
| % met for all studies | | 88 | 90 | 78 | 13 | 15 | 60 | 54 | 16 | 79 | 93 | 46 | 58 |

*Considered Level 1 criteria as essential to minimize bias.
3 = Participants; 4 = Interventions; 6 = Outcomes; 7 = Sample size; 8, 9, 10 = Randomization; 11 = Blinding; 12, 17, 18 = Statistical methods; 13 = Participant flow; 14 = Recruitment and follow-up; 15 = Baseline data; 16 = Numbers analyzed.

to their treatment type. The mean percent of criteria met was comparable between the treatment types, ranging from 67% for pharmacological treatments to 45% for occlusal therapies. Over the six treatment types, the percent of criteria met by individual studies ranged from 20% to 100%. Taking into account all studies, the mean percent of criteria met was 58% with only 10% of the RCTs meeting the four most important criteria.

### Percent Totals for Studies Meeting Each of the Fifteen Criteria

Table 4 presents the individual review criteria with the percent of studies that met them. Only 10% of RCTs for TMJD met all of the Level I criteria. The most common problems and the percent of the studies involved were as follows: only 15% described an adequate concealed randomization process; 13% presented sample sizes calculation required for predefined outcome measures; 16% monitored adherence; 46% took into account withdrawals and crossovers in an intent to treat analysis; and 54% adequately described their data analysis. The most common criteria met included the use of baseline measures (93%), interventions being equalized between groups with baseline variables compared (93%), relevant and reliable multidimensional measures employed (78%), treatment well-defined and standardized (90%), and clear recruitment procedures with inclusion/exclusion criteria (79%).

### Change in RCT Quality Scores Over Time

Figure 2 represents a scatter plot of quality scores according to the year in which the RCT was published. The overall mean quality score for all the RCTs was 0.58 (SD 0.16). A regression line was fitted to these data and had a slope of 0.010. This indicates that for each unit (year) increase on the X-axis, the average quality score increased by 0.01, or one point on a 1 to 100 scale. Over any

period of one decade, quality scores are estimated to have improved by 10 points on the average, and this trend in quality improvement is highly statistically significant ($F_{1,192} = 40.09$, $P < .0001$).

## Discussion

The purpose of this study was to compare the quality of the methods of TMJD RCTs against known standards for RCTs as described by the CONSORT criteria. As Antczak and colleagues[22] have stated, "evaluating quality of RCTs is the first step in efforts to combine data from a number of similar trials in meta-analysis." This was implemented by focusing on the study quality as objectively as possible, rather than on the study results and recommendations where there could be differences of opinion. The 11 study design criteria for this quality assessment, scored as present (1) or absent (0), were determined to be the most common sources of differential (systematic) and nondifferential (random) error in the study designs for TMJD RCTs. Additional "reporting only issues" as identified in the CONSORT statement were not scored since the intention was to evaluate study methods and not reporting, but their importance cannot be overstated. Since each study was evaluated only on the published report, it is possible that the specific criteria may have been scored as having not been met but not reported. Furthermore, the problem of publication bias has not been addressed in this study. Publication bias arises from the tendency for researchers and editors to publish experimental results that are positive while results that are negative or inconclusive are left out or unpublished. This contributes to the overwhelming percent of published articles that demonstrate positive outcomes and thus, systematic reviews may not allow a true indication of the efficacy of a specific treatment. Regardless, quality reviews can still be useful to help investigators design and publish RCTs with biasing factors considered.

### RCT Quality Criteria Inadequately Applied in TMJD RCTs

The results of this study suggest that many of the universally accepted criteria for clinical trials are often not applied in RCTs of TMJD. Since 1979, the quality of RCTs in TMJD, or the reporting of such, has improved significantly, thus lending more validity to more recent studies (Fig 2). This suggests that there may be an increasing awareness of what is required to conduct and properly report RCTs in TMJD. However, only 10% of the reviewed study sample met all four criteria to minimize systematic bias. The mean percent of criteria met was 58%, based on the 15 criteria used in this review. Such inadequacies would predictably have contributed to bias in the study designs, thereby resulting in the heterogeneity in the observed results.

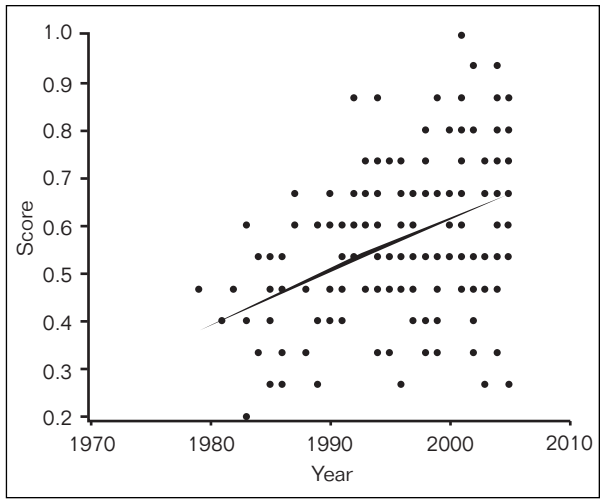


**Fig 2** Scatter plot of RCT quality versus year of publication. The regression line shows a significant ($P < .0001$) improvement in RCT quality over time with an average increase of 0.01 points per year.

Of the four essential Level I criteria, the most common problem was criterion 8 through 10: lack of a defined and concealed randomization process to minimize *selection bias*. Based on the study reports, this was met by only 15% of the studies. Concealment implies that both the investigator and the subject are blind to and unable to influence the treatment assignment and, thus, the treatment results as well. Since a defined and concealed randomization process is a well-known requirement, some studies may have met this criterion but did not report it.

Another essential criterion to minimize *measurement bias* by blinded measurement of outcome was met by just 60% of the studies. It has been shown that treatment effect estimates from studies with inadequate concealment of treatment assignment may be exaggerated by 40%.[13] A double-blind study is preferred in which both the rater and the subject are blinded. However, for many TMJD treatments such as surgery, the blinding of the patient/subject to their treatment group is not possible or ethical. In these studies, it is best to control all variables possible and consider that some subject bias towards better efficacy may be present.

*Performance bias*, or *unequal comparison group bias*, must be minimized by the subject's active participation in the control group being equal to that of the active treatment group. These groups should be comparable with regard to clinician contact, medication use, and time of follow-up. It is difficult for the reader to know these details unless they are specifically reported. Ninety-three percent of studies attempted to minimize comparison group bias by controlling for baseline differences in prognostic factors. As noted earlier, randomization cannot guarantee the absence of chance-related baseline imbalances between treatment groups that influence results, especially with sample sizes of 40 or less. It is important to measure at baseline whether groups are comparable with regard to known prognostic factors, such as gender, duration of pain, and depression, and to take between-group differences into account for the analysis.

Only 46% of the studies considered the effect of *attrition bias* by performing an intent-to-treat analysis that included drop-outs or crossovers to a different treatment than what had been randomly allocated particularly with treatments for chronic pain that often used multiple modalities. Such protocol deviations or subjects that are lost to follow-up may produce a distortion of the estimated treatment effects. An intent-to-treat analysis needs to include all subjects randomized independent of any protocol violation so that the inherent statistical assumptions based on the randomized treatment allocation are valid.

Many of the other quality criteria were also not met by most studies. For example, only 13% of studies met criteria number 7 by conducting sample size calculations using pilot data. This may be a contributing factor to Type II error (false negative) in studies that would have shown an effect if an adequate sample size had been used. The impression may be that this error did not have a biasing effect and cannot be considered a problem in those studies, but this assumption is untrue. Studies with a small sample size are much more likely to have low power suffering from inflated Type II error, finding no difference when one does exist but the opposite can also be true. However, without taking into account the typical variation in the study factor, it is difficult to know whether a small sample size can accurately represent the target population.

Patient compliance with treatment, particularly when the patient plays a role in the active effects of the treatment such as using a splint or performing an exercise, also contributes to significant variability of results. Unfortunately, only 16% of studies provided some evidence of patient flow and adherence to the treatment protocol. In reviewing criteria 6, few studies considered the ceiling or floor effect in selecting subjects whose baseline symptoms were sufficiently severe enough to detect an active effect of treatment. Parallel to this matter of symptom severity are the temporal characteristics such as frequency and duration of the signs or symptoms and the need for clinically relevant outcome measures. These temporal clinical characteristics may change sooner and become more clinically relevant than pain intensity with some interventions for TMJD. Without these measurements, important changes in symptoms may remain undetected, resulting again in Type II error.

### Use of Quality Scores in Systematic Reviews

Quality scores can be used in systematic reviews in a variety of ways including weighing higher-quality studies, applying scores as a threshold for inclusion of a study in a review, as well as with an analysis and comparison of results with other reviews. For example, although weighted composite scores were not used in this review, a good example of the composite scoring approach is presented by Antczak and colleagues.[22] They published a method for assigning a greater weight to more important individual criteria, and then computing a composite score for the quality assessment of periodontal treatment RCTs. Their proposed quality score included three separate sections: *(1)* basic identification of the paper for classification purposes, *(2)* quality of the study protocol, and *(3)* data analysis and presentation of the paper. These criteria were also used, with minor modifications, in a review by Forssell and Kalso[1] of TMJD RCT evidence for the efficacy of occlusal treatments. Splints and occlusal adjustments were the two types of occlusal treatments that they examined. These authors determined that the overall quality of these RCT studies was fairly low and the results were equivocal. Although Antczak and colleagues held that meta-analysis was justifiable as the next step after their narrative synthesis of the periodontal treatment evidence, Forssell and Kalso did not, due to the heterogeneity of the TMJD studies that they reviewed.

When the present results are compared to those of Forssell and Kalso,[1] scoring agreement showed a mean ICC of 0.59 and demonstrated good agreement. Although the present study included more well-defined criteria, many of the weaknesses of the RCTs found by Forssell and Kalso were also consistent with those identified in the present study. For example, few studies had appropriate

randomization, many did not have blinded measurement of outcome, few measured adherence to treatment, and some did not consider the issues of sample size requirements, attention to dropouts, or the use of co-interventions not defined for the study protocol. In contrast to the review by Forssell and Kalso, the present study found that most studies had defined their study population (88%), their treatments were adequately defined (90%), and they had an adequate follow-up period (79%). This difference between review results may be due to the Forssell and Kalso review being limited to occlusal treatments, whereas the present results were based on a review of six types of treatments and 210 RCTs. Another design concern found by both reviews was the lack of a run-in period relative to prior treatments, self-care, and medications, with few studies satisfying this quality criterion. Any extraneous treatment such as analgesic medications not defined as part of the experimental or control interventions may influence outcomes and confound the treatment effects. They need to be matched between groups, eliminated before the study begins during the run-in period, or measured and controlled for in the statistical analysis.

### Limitations of this Review

There are several limitations to this quality review study. It is acknowledged that several scales are available to assess quality of RCT methods and this study used only the CONSORT criteria as explained. However, as noted above, the use of the criteria of Antczak et al[22] to compare the present findings with those of Forssell and Kalso[1] showed good agreement between both studies. Second, the searches used in the present study identified 210 RCTs published in the English language but excluded studies in other languages. Thus, while this study attempted to capture the majority of the published literature, it missed some literature that would have had relevance to this review.

### Towards Improvement of the Evidence Base for TMJD Interventions

Based on this review, there are recommendations for researchers planning to develop RCTs for TMJD. In general, following the CONSORT criteria as discussed here will lead to results that minimize bias.

*(1) Use of the RCT as a Gold Standard.* Most of the identified studies in this present study's search failed to meet even the essential criteria to minimize bias, thus bringing into question the credibility of their findings. Thus, their contribution relative to clinical treatment guidelines and recommendations is questionable.

*(2) Improving Quality and Reporting of RCTs of TMJDs.* Studies are needed to test TMJD interventions both against placebo groups and other treatments to determine their true relative efficacy. Unfortunately, TMJD RCTs have been often underfunded, conducted within a too-limited time frame, and underpowered. Funding agencies need to insist on standardized methodologies in the review process and ensure that funds are sufficient to conduct high-quality studies. Emerging information systems involving national registries may be appropriate for standardizing design and data collection for multicenter RCTs. More emphasis should be made on multicenter studies to ensure adequate sample sizes and broad generalizability of the results. Inclusion of the subject flow diagram will provide a description of the progress of participants throughout the study, from the number of potentially eligible individuals for inclusion in the trial, to the number of trial participants in each treatment group who complete the trial. Editors of journals need to require quality standards in their review processes. This not only encourages investigators to report their methods clearly, but also helps reviewers to assess bias in the study designs accurately. If appropriate design criteria are not met, the investigators should be prepared to justify why they were not applied.

*(3) Use of Standardized Outcome Measures for Improved Comparability Between Studies.* Many measures have been already developed and are being used across studies. More research needs to be conducted on effective tools to improve quality and ease of conducting RCTs. An attempt to standardize measures in chronic pain clinical trials has been initiated by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT), which is a collaboration between pain researchers and industry and government agencies.[16,17] However, disease-specific outcome measures need to be developed for TMJD such as reliable and valid outcomes for assessing jaw function, oral habits, and jaw-related disability.

*(4) Avoiding Diagnostic Misclassification.* Since it is possible that the specific subtype of TMJD, duration of pain, and comorbid conditions such as fibromyalgia and migraine may be important in determining outcomes of a particular intervention, it is recommended that future studies of TMJD should control for the specific diagnostic subtypes, the duration of pain, and comorbid conditions in

the study sample. Many of the reviewed studies failed to address these issues. Antczak and colleagues have emphasized this issue under Item 2.1: Selection Description.[22] The subjects must be well characterized as to their accession diagnoses and baseline prognostic factors such as duration of pain in order for the results to be interpretable and clinically meaningful. It is recommended that future studies of TMJD use the RDC/TMD to control for subtype diagnostic groups as well as a history of comorbid conditions and duration of pain, as suggested by the IMMPACT guidelines.[15–17,30]

*(5) Addressing the Inherent Susceptibility to Bias of Pain Measurement in TMJD.* Pain is the major outcome variable in most TMJD studies but is often susceptible to bias because it is dependent on subject self-report. Many of the therapies cannot be concealed from the patient, and it is well understood that the patients' preconceived idea of the therapy's benefit can influence their response with self-report instruments. It is recommended that standardized clinical examination protocols be performed by a blinded examiner that can be compared to the subject's self-report. However, even some examination items may not be completely objective since a patient must endorse pain in response to measures such as palpation pressure. A concurrent multidimensional data collection is another means for supporting the validity of a self-report. These may be standardized measures for the same construct to establish concurrent validity, or the assessment of other factors that can explain the perception of pain such as emotional factors, subject disposition, and global improvement.

*(6) Better Use of Evidence to Conduct Meta-analyses of RCTs.* It is expected that future RCTs in TMJD will be improved due to a growing awareness of the essential study design criteria and study reporting requirements. This is also supported by the trend towards such improvement occurring over time, as illustrated in Fig 2. The view of the authors is that quantitative estimates of benefit should be performed using emerging meta-analysis procedures and pooling the results of RCTs with similar interventions and comparison groups. The relative contribution of each study can be weighted by its respective standard error and its heterogeneity relative to the overall body of evidence. The alternative of attempting to assess the influence of multiple flaws in design—with an average of just 58% of criteria met—is problematic. Thus, it is reasonable to explore existing quantitative evidence using current mathematical methods to obtain information that can then be compared to the qualitative narrative syntheses.

## Conclusions

In certain areas of medicine, observational studies and nonrandomized controlled studies have an important application such as cohort studies comparing effectiveness, outcome quantification, and risk factor identification. However, the RCT is the study design recommended to assess the effectiveness of specific interventions and should be the design of choice to avoid susceptibility to systematic bias. Significant improvement in attention to accepted design criteria and reporting standards as defined by the CONSORT criteria are needed for all future TMJD studies. The present analysis found that the overall quality of the reviewed studies was modest, with only 58% of the quality criteria met and only 10% of the RCTs met the four most important criteria. But there has been a trend toward improvement in study quality over time. A hierarchy of steps is proposed for clinicians to evaluate critically the internal and external validity of studies of interest. It is hoped that by discussing these methods used in past RCTs, important improvements and standardization will be stimulated for future RCTs. Ultimately, improvement in RCT methods will allow better understanding and comparison of studies. The combining of their results by meta-analysis will also maximize the benefits obtained from clinical trial research. Notwithstanding the difficulty of many studies to meet essential criteria, and to minimize systematic bias, the available evidence needs to be fully analyzed using existing as well as newly emerging mathematical methods to synthesize the results. This will ultimately improve clinical guidelines and the care of TMJD patients.

## Acknowledgments

This project acknowledges support from the National Institute of Dental and Craniofacial Research's TMJ Implant Registry and Repository (NIH/ NIDCR Contract # N01-De-22635) and the American Academy of Orofacial Pain.

## References

1. Forssell H, Kalso E. Application of principles of evidence-based medicine to occlusal treatment of temporomandibular disorders: Are there lessons to be learned? J Orofac Pain 2004;18:9–22.
2. Ernst E, White AR. Acupuncture as a treatment for temporomandibular joint dysfunction: A systematic review of randomized trials. Arch Otolaryngol head Neck Surgery 1997;125:269–272.

3. Koh H, Robinson PG. Occlusal adjustment for treating and preventing temporomandiublar joint disorders. Cochrane Database Syst Rev. 2003;(1):CD003812.
4. Al-Ani MZ, Davies SJ, Gray RJ, Sloan P, Glenny AM. Stabilization splint therapy for temporomandibular pain dysfunction syndrome. Cochrane Database Syst Rev. 2004;(1):CD002778.
5. Kropmans TJ, Dijkstra PU, Stegenga B, De Bont LG. Therapeutic outcome assessment in permanent temporomandibular joint disc displacement. J Oral Rehab 1999; 26:357–363.
6. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effect in controlled clinical trials. J Am Med Assoc 1995;273:408–412.
7. Moher D, Pham B, Jones A, et al. Quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609–613.
8. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992;45:255–265.
9. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: The CONSORT statement. J Am Med Assoc 1996;276:637–639.
10. Korn D, Ehringhaus S. Principles for strengthening the integrity of clinical research. PLoS Clin Trials 2006;1:e1.
11. Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. Lancet 2001; 357:1191–1194.
12. Ioannidis J, Evans SJ, Gotzsche PC,et al. Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 2004;141:781–788.
13. Juni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Smith GD, Altman DG (eds). Systematic Reviews in Health Care: Meta-analysis in Context, ed 2. London: BMJ, 2001.
14. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. JAMA 2001;285: 1992–1995.
15. Drangsholt M, LeResche L. Temporomandibular Disorder Pain. Seattle: IASP, 1999.
16. Turk DC, Dworkin RH, Allen RR, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106:337–345.
17. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.
18. Mannion AF, Elfering A, Staerkle R, et al. Outcome assessment in low back pain: How low can you go? Eur Spine J 2005;14:1014–1026.
19. Smith LA, Oldman AD, McQuay HJ, Morre RA. Teasing apart quality and validity in systematic reviews: An example from acupuncture trials in chronic neck and back pain. Pain 2000;86:119–132.
20. Jadad A. Randomised Controlled Trials. London: BMJ, 2001.
21. Altman DG. Better reporting of randomized controlled trials: The CONSORT statement [editorial]. BMJ 1996; 313:570–571.
22. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized controlled trials in dental research I. Methods. J Periodontal Res 1986;21:305–314.
23. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Controlled Clin Trials 1995;16:62–73.
24. Jadad AR, Cook DJ, Jones AL, et al. The quality of randomised controlled trials included in meta-analyses and systematic reviews: How often and how is it assessed? Abstract presented at the 4th Annual Cochrane Colloquium, Adelaide, Australia, 1996.
25. Ohlsson A, Lacy JB. Quality assessments of randomized controlled trials: Evaluation by the Chalmers versus the Jadad method. Presented at the 3rd Annual Cochrane Colloquium, 1995.
26. Methods for Systematic Reviews. Oxford, UK: The Oxford Centre for Evidence-based Medicine, 2006.
27. Korn D, Ehringhaus S. Principles for strengthening the integrity of clinical research. PLoS Clin Trials 2006;1:1–4.
28. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Health Technol Assess 2003;7:1–76.
29. Ekberg E, Vallon D, Nilner M. The efficacy of appliance therapy in patients with temporomandibular disorders of mainly myogenous origin. A randomized, controlled, short-term trial. J Orofac Pain 2003;17:133–139.
30. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord 1992;6:301–355.
31. Headache Classification Committee of the International Headache Society. Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. Cephalalgia 1988;8(suppl 7):1–96.
32. Okeson JP (ed). Orofacial Pain: Guidelines for Assessment, Diagnosis, and Management. Chicago: Quintessence, 1996.
33. Ekberg E, Nilner, M. Treatment outcome of short- and long-term appliance therapy in patients with TMD of myogenous origin and tension-type headache. J Oral Rehabil 2006;33:713–721.
34. Fernandez-de-Las-Penas C, Alonso-Blanco C, Cuadrado ML, Gerwin RD, Pareja JA. Myofascial trigger points and their relationship to headache clinical parameters in chronic tension-type headache. Headache 2006;46: 1264–1272.
35. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports on randomized clinical trials: Is blinding necessary? Controlled Clin Trials 1996;17:1–12.
36. Higgins JPT, Green S. Sections 6.2 to 6.6. In: Cochrane Handbook for Systematic Reviews of Interventions 4.2.5, 2005.
37. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomized controlled trials: The quorom statement. Onkologie 2000;23:597602.