The Relationship Between Methodological Quality and Conclusions in Reviews of Spinal Manipulation

Willem J. J. Assendelft, MD; Bart W. Koes, PhD; Paul G. Knipschild, PhD; Lex M. Bouter, PhD

Objective.—To study the relationship between the methodological quality and other characteristics of reviews of spinal manipulation for low back pain on the one hand and the reviewers’ conclusions on the effectiveness of manipulation on the other hand.

Data Sources.—Reviews identified by MEDLINE search, citation tracking, library search, and correspondence with experts.

Study Selection.—English- or Dutch-language reviews published up to 1993 dealing with spinal manipulation for low back pain that include at least two randomized clinical trials (RCTs).

Data Extraction.—Methodological quality was assessed using a standardized criteria list applied independently by two assessors (range, 0% to 100%). Other extracted characteristics were the comprehensiveness of the search, selective citation of studies, language, inclusion of non-RCTs, type of publication, reviewers’ professional backgrounds, and publication in a spinal manipulation journal or book. The reviewers’ conclusions were classified as negative, neutral, or positive.

Data Synthesis.—A total of 51 reviews were assessed, 17 of which were neutral and 34 positive. The methodological quality was low, with a median score of 23%. Nine of the 10 methodologically best reviews were positive. Other factors associated with a positive reviewers’ conclusion were review of spinal manipulation only, inclusion of a spinal manipulator in the review team, and a comprehensive literature search.

Conclusions.—The majority of the reviews concluded that spinal manipulation is an effective treatment for low back pain. Although, in particular, the reviews with a relatively high methodological quality had a positive conclusion, strong conclusions were precluded by the overall low quality of the reviews. More empirical research on the review methods applied to other therapies in other professional fields is needed to further explore our findings about the factors related to a positive reviewers’ conclusion.

Historically, spinal manipulation has been associated with chiropractic and osteopathy. Political controversies between chiropractors and related professionals and the medical establishment possibly played a role in the scientific assessment of the effectiveness of spinal manipulation. There are many different approaches to spinal manipulation, and a great variety of control treatments are used in RCTs. For spinal manipulation, heterogeneity in the reviewers’ opinions can be expected, thus making it a good subject for investiga-
Table 1—Criteria List for Assessment of Methodological Quality of Review Articles, With Weights Applied*  

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Selection: A. Description of inclusion and exclusion criteria</td>
<td>10</td>
</tr>
<tr>
<td>B. Search strategy</td>
<td>10</td>
</tr>
<tr>
<td>C. Emphasis on randomized clinical trials (RCTs)</td>
<td>10</td>
</tr>
<tr>
<td>Methodological Quality Assessment: D. Assessment of the validity RCTs</td>
<td>12</td>
</tr>
<tr>
<td>E. Number of reviewers</td>
<td>10</td>
</tr>
<tr>
<td>F. Blinding of reviewer(s)</td>
<td>2</td>
</tr>
<tr>
<td>G. Agreement of reviewer(s)</td>
<td>2</td>
</tr>
<tr>
<td>Intervention: H. Description of manipulative intervention(s)</td>
<td>8</td>
</tr>
<tr>
<td>I. Description of control intervention(s)</td>
<td>7</td>
</tr>
<tr>
<td>Data Presentation: J. Outcome presentation</td>
<td>20</td>
</tr>
<tr>
<td>K. Statistical pooling</td>
<td>3</td>
</tr>
<tr>
<td>L. Discussion power of negative RCTs</td>
<td>3</td>
</tr>
<tr>
<td>Evaluation: M. Overall conclusion</td>
<td>5</td>
</tr>
<tr>
<td>N. Discussion of heterogeneity of RCTs and outcomes</td>
<td>10</td>
</tr>
</tbody>
</table>

*The detailed criteria, with weights attached, are as follows:  
A. Description of study setting(s) included (ie, industry, general practice, hospital), intervention type(s) included (manipulation, mobilization, or both; specific techniques or profi-  
lers [proprioception, core strengthening, Cyriax, technique], osso-  
tic type(s) included (pain, global assessment, mobility,  
functional status, time until recovery, medical consumption),  
years covered, languages covered (2 points each).  
B. Established bibliographic database included (eg,  
Index Medicus, EMBASE) (5 points). Additional efforts to locate  
nonindexed RCTs (eg, citation tracking, correspondence with  
experts, manual searches in nonindexed journals) (5 points).  
C. Randomized clinical trials only, or results of RCTs dis-  
cussed separately from other study designs (10 points).  
D. A RCT included that is explicit (reproduc-  
bility by readers of the review) regarding the similarity of  
treatment groups, similarity of treatment characteristics,  
equity in treatment of missing values (dropouts, loss to follow-up),  
success in blinding of outcome assessment, relevance of  
outcome measures, and adequacy of statistical analysis (2 points each).  
E. At least two independent reviewers (4 points).  
F. Reviewer(s) blinded for at least the outcomes of the  
RCTs (2 points).  
G. Agreement between reviewers reported (quantitative) and  
acceptable. In the event of reviewer, use of an assess-  
ment list with established reliability (2 points).  
H. Description of manipulative intervention per RCT: manipulation  
technique (3 points), profession or training man-  
rupar (1 point), treatment frequency or number of treat-  
ments (2 points), duration of treatment period (2 points).  
I. Description of control treatment per RCT: type (3 points),  
treatment frequency or number of treatments (2 points),  
duration of treatment period (2 points).  
J. Main outcome(s) (at least one of the following: pain,  
global assessment, mobility, functional status, medical con-  
sumption) presented separately per RCT per group (5 points).  
Presentation of the mean difference or ratio of outcome(s)  
between intervention group(s) and control group(s) (3 points)  
with confidence interval or SD per RCT (3 points). Graphs  
presenting the most important outcome(s) (indicating  
outliers and distribution) per RCT (3 points). That is, the  
most important outcome(s) that are rational in relation to the re-  
search question of the review. If not explicitly stated, at least  
mean, recovery, or functional status.  
K. Statistical pooling of the most important outcome(s) or discussion  
of the reasons why pooling is not indicated or war-  
ranted or pooling of the subset considered to be valid and  
similar enough (3 points).  
L. Elaboration on the power of negative RCTs: calculation of  
the power of each negative RCT (3 points) or narrative  
elaboration on the power of each negative RCT (2 points) or  
narrative elaboration on the power of the negative RCTs (1 point).  
M. Overall conclusion on the aggregated level of available  
RCTs on the effectiveness of manipulation presented (5 points).  
N. Identification of relevant subgroup(s) (eg, study setting,  
disease or condition) with explicit motivation (4 points).  
Discussion of the variety of treatment modalities in the interven-  
tion groups (eg, mobilization, manipulation; chiropractic,  
osteopathic, Cyriax technique) (2 points). Discussion of the  
variety of treatment modalities in the control groups (placebo,  
existing modality) (2 points). Discussion of the relationship  
between the methodological quality of the RCTs and out-  
come (2 points).  

**METHODS**  

**Selection of Review Articles**  

A review article was included in our study if (1) the article dealt with the effectiveness of manipulation or mobilization for low back pain (not necessarily exclusively); (2) at least two randomized clinical trials were included in the assessment of effectiveness; (3) the article was not the original publication of an RCT itself; (4) it was published as a book chapter, report, or journal article; (5) the year of publication was up to 1993; and (6) the language was English or Dutch.  

Relevant articles included articles by using the key words back pain therapy, musculoskeletal diseases, joint diseases, manipulation-orthopedic, osteopathic medicine, chiropractic, evaluation studies, outcome, and process assessment for a computer search of Index Medicus (MEDLINE) for the period 1967 through 1993; by article citation tracking and library search; and through correspondence with experts in the field.  

To check for the completeness of our search strategy, all available titles of RCTs up to 1991 were selected in the Science Citation Index for 1970 through 1993. All articles citing these RCTs were scanned on title and type of publication in the Science Citation Index and subsequently in Index Medicus (MEDLINE).  

**Assessment of Methodological Quality**  

For the assessment of the methodological quality of the review articles, we used a list of standardized criteria based on previous work of Oxman and Guyatt. Light and Pillemter, and Mulrow. Further operationalization for the specific topic of spinal manipulation was based on our previous experience in this field. The list contained specified criteria covering the following domains: study selection, methodological quality assessment, description of the intervention, data presentation, and evaluation (Table 1).  

A specific, arbitrary weight was applied to each criterion, as described in detail in the footnote to Table 1. The scoring was done by two nonblinded assessors independently. By “assessors” we mean the first two authors of this article, W.J.J.A. and B.W.K., who applied the methodological criteria and performed the data extraction; by “reviewers” we mean authors who have written a review based on RCTs. The final score was determined in a consensus meeting. If disagreements were not resolved during this meeting, a third blinded reviewer (L.M.B.) was consulted to make the final judgment. The total score thus determined is further referred to as the “review score.”  

**Data Extraction**  

Characteristics of the review that were considered to be relevant were classified into two categories, methodological and subject related.  

Methodological characteristics included the comprehensiveness of the literature search, selective citing of the available positive and negative RCTs, language, inclusion of non-RCTs, type of publication (ie, journal article or book chapter), and being indexed in Index Medicus. The previously mentioned review score was also considered to be a methodological characteristic. Since there is no registry of peer-reviewed journals, being indexed in Index Medicus was included as a substitute indicator for peer review.  

Subject-related characteristics included publication in a specific spinal manipulation journal or book and the professional backgrounds of the reviewers (spinal manipulator or not). These characteristics were included to investigate whether a professional link between the reviewer or the publication medium and spinal manipulation influenced the reviewers’ final conclusions.  

For the determination of the comprehensiveness of the literature search of reviews, we assumed an RCT to be available for the reviewer(s) if it was published more than 2 years before the review.  

The dependent variable (ie, the final conclusion of the reviewer(s)) was expressed as “negative,” “neutral/no conclusion drawn,” or “positive.” The conclusion was considered to be negative if the reviewer(s) stated that the reference treatments were more effective than manipulation, and the conclusion was considered to be positive if the reviewer concluded that spinal manipulation was more effective than the reference treatments. The review was rated neutral if no conclusion was drawn or if the reviewers were neutral in their final conclusion. Reviews with the positive conclusion restricted to subgroups (eg, [sub]acute complaints) or to the duration of the effect (eg, short-term effects) were rated as positive. For these reviews, the nature of the restriction is reported separately.  

The two assessors (W.J.J.A. and B.W.K.) were not blinded for the review
Table 2.—Methodological Review Scores and Reviewers’ Conclusions

| Source, y | A 10 | B 10 | C 12 | D 12 | E 4 | F 4 | G 8 | H 7 | I 7 | J 14 | K 3 | L 5 | M 5 | N 10 | Total | Conclusion† | Restriction‡ |
|-----------|------|------|------|------|----|----|-----|-----|----|------|----|----|-----|------|-------|---------|-------------|-------------|
| Shekelle et al., 1992 | 6 | 10 | 10 | 12 | 4 | ... | 3 | 3 | 14 | 3 | ... | 5 | 6 | 76 | Positive | Yes | Yes |
| Koes et al., 1991 | 4 | 10 | 10 | 12 | 4 | 2 | 2 | 4 | 3 | 5 | 3 | 2 | 2 | 6 | 72 | Neutral | ... | ... |
| Assendelft et al., 1992 | 4 | 10 | 10 | 12 | 4 | 2 | 2 | 4 | 3 | 5 | ... | 5 | 6 | 61 | Positive | No | Yes |
| Anderson et al., 1992 | 6 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Conlon, 1992 | 2 | 10 | 10 | 4 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Shekelle et al., 1991 | 4 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| DiFabio, 1992 | 2 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Lyno and Bygren, 1987 | 2 | 10 | 10 | 6 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Ottenbacher and DiFabio, 1985 | 4 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Winer, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| O’Donoghue, 1982 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Auldenkemper, 1991 | 2 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Bronfort, 1992 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Dayo, 1983 | 2 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Evans, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Jacobs et al., 1991 | 2 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Curtis, 1988 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Brunski, 1982 | 2 | 10 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Haldeman and Phillips, 1991 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| DiFabio, 1986 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Branson, 1984 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Haldeman, 1980 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Montz, 1979 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Spitzer, 1987 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Manga et al., 1993 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Coullman, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Haldeman, 1983 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Haldeman, 1986 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Greenland and Haldeman, 1980 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Lawrence et al., 1992 | 4 | 5 | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Jafson, 1986 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Bouter, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Yates, 1980 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Tan, 1992 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Deyo, 1991 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Quartz and Hadler, 1979 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Grahame, 1980 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| LaBor and Taylor, 1992 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Gatterma, 1990 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Lankhorst, 1987 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Gilbert, 1986 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Kukurin, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Paris, 1983 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Inglis, 1979 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Nachenison, 1977 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Geinner et al., 1993 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Nachenison, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Nachenison, 1985 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Flor and Turk, 1984 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Nachenison, 1980 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Maximum score per item, % | 12 | 25 | 94 | 11 | 6 | 4 | 4 | 24 | 23 | 5 | 12 | 8 | 25 | 5 | 22 | ... | ... | ... |

*See the footnote to Table 1 for an explanation of the methodological scoring criteria.
†Positive conclusion indicates spinal manipulation was better than control treatments; and neutral, spinal manipulation was equal to control treatments or no conclusion drawn. Ellipses indicate not applicable.
‡(Sub) acute restriction indicates restriction of reported positive effects to patients with (sub) acute complaints (mainly <6 weeks); and short-term, restriction of duration of reported positive effects of spinal manipulation to a short duration (the exact period was mostly not specified).

Characteristics when they determined the value of the independent variable (the final conclusion). Therefore, the final conclusion per review was recorded by both the two assessors (W.J.J.A. and B.W.K., pair A) and also by two other trained clinical epidemiologists (A.F. de Winter, MSc, and D.A.W.M. van der Windt, MSc, pair B). Since the opinion of pair B had the least risk of bias, as they were not authors of a review themselves and had no knowledge of the other
items recorded (including the methodological criteria), their recordings were used in the final analysis.

**Data Entry and Analysis**

All data were entered and analyzed in SPSS/PC. Data were initially analyzed in frequency tables and by cross tabulation. To calculate crude and adjusted odds ratios (ORs), the ordinal and continuous independent variables (review characteristics) were all dichotomized. For continuous data the median was taken as the cutoff point. Univariate analysis was reported as ORs with corresponding 95% confidence intervals (CIs).

All characteristics with a statistically significant univariate OR and the methodological review score, which we considered to be essential a priori, were entered in a logistic regression model.

**RESULTS**

**Review Identification**

Initially 60 review articles were identified. Nine did not fulfill all the inclusion criteria: one was written in German, one dealt with methodological issues only, and seven did not include two or more RCTs. The number of reviews on the effectiveness of spinal manipulation for low back pain in English or Dutch we thus identified (51) (Table 2) by far exceeded the number of original RCTs (28).

**The Reviewers’ Conclusions**

The agreement between the two independent pairs of assessors regarding the final conclusions of the reviewers was high: 92%. None of the reviewers stated that spinal manipulation was less effective than reference treatments, 17 were neutral or offered no conclusion, and 34 were positive (Table 2). Of the 34 positive reviews, 15 (44%) restricted the positive conclusion to short-term results for manipulation and two (6%) restricted the conclusion to (sub)acute back pain, while nine (26%) restricted conclusions to both (sub)acute back pain and short-term results and eight positive reviews (24%) mentioned no restriction in their conclusions. Most reviewers did not define precisely what was meant by the terms (sub)acute and short-term.

**Review Score**

The two nonblinded reviewers initially agreed on 94% of all subitems scored (total, 51 × 35 = 1785 subitems). Nearly all the disagreements were due to reading errors or to a difference in interpretation of the methodological criteria. Since all the disagreements were resolved during the consensus meeting, the third reviewer did not have to be consulted.

The results of the assessment of the methodological quality of the review articles are presented in Table 2. The median score was 23 (range, 5 to 76), indicating the overall poor methodological quality of the reviews. Only three reviews had a score higher than 50 points of the 100 points possible. Recently published reviews tended to have higher scores.

The Figure presents the relationship between the review score and the reviewers’ conclusions. Nine of the 10 highest-ranking (methodologically best) reviews had a positive reviewers’ conclusion (see also Table 2).

**Review Characteristics**

The prevalence of the selected characteristics in the review articles included is presented in Table 3. In 16 (31%) of the 51 reviews, treatments other than spinal manipulation were also included. In 18 (40%) of the 45 reviews presenting data on this characteristic, a spinal manipulator was (co)author of the review. One third of the reviews were published in journals or books dealing exclusively with spinal manipulation. A majority of the reviews (22%) also included non-RCTs. Four reviews were in Dutch. Only a few reviews included statistical pooling. Half of the reviews included fewer than 60% of the RCTs available at the time of writing. Remarkably, the mean percentage of available RCTs included did not differ between the 12 reviews that reported an extensive search (use of a database as well as additional efforts (10 points on item B)) and the 39 reviews that had no description of such a comprehensive search (59% and 60%, respectively) (data not given in Table 3).

The results of the univariate analysis are presented in Table 3. If a review dealt with spinal manipulation only (n = 35), it had a positive reviewers’ conclusion much more often. Of these 35 reviews, 29 (82.9%) had a positive conclusion. Of the 16 reviews not dealing with spinal manipulation exclusively, only five (31.3%) had a positive conclusion. The resulting OR was 10.6 (95% CI, 2.68 to 42.1). The presence of a spinal manipulator in the review team also provided a strong association with a positive conclusion. This review characteristic was unknown for seven reviews. A third characteristic, a comprehensive literature search, also had a statistically significant OR. Also positively but not significantly related were publication in a spinal manipulation journal or book, an explicit validity assessment in the review, and a high review score. There were no clear associations with the reviewers’ conclusion for selective citing of positive RCTs, language, inclusion of non-RCTs, publication in a journal (as opposed to publication in a book or report), or being indexed in Index Medicus.

In the logistic regression analysis we included the characteristics “review of manipulation only,” “spinal manipulators in the review team,” and “completeness of search,” which were the characteristics with a statistically significant crude OR, and “methodological review score.” Statistical analysis revealed a strong association between the characteristics spinal manipulator in review team and review of spinal manipulation only. Therefore, in the logistic regression analysis we present the data for two models, alternately including these colinear variables. In these two models, only “review of manipulation only” still had a statistically significant OR.

**COMMENT**

Potential Biases

As in other kinds of empirical research, reviewing the literature is also prone to several types of bias. Our first consideration is reference bias, a form of selection bias. In general, review articles are not systematically cited by others, as is the case with RCTs in review articles. Therefore, citation tracking is not a guarantee for completeness. In fact, we expect that there are more reviews existing on this subject than we have actually identified. Some reviews were obtained through “correspondence with experts,” which is an accepted policy in literature search. In most cases these experts were spinal manipulators themselves, which suggests that our sample of reviews may have a bias toward more positive conclusions. Being indexed in Index Medicus was recorded as a separate review char-
characteristic. Table 3 shows that there was no difference in conclusion between the indexed and nonindexed reviews. Therefore, we conclude that the reviews obtained through our additional search efforts did not lead to substantial bias. We did not include unpublished reviews, since unpublished reviews do not have any impact on the general opinion. As a consequence, our conclusions apply to published reviews only.

The nonblinded recording of review characteristics and the reviewers' conclusions implies the possibility of information bias. However, we have presented our findings in a reproducible manner (Table 2). In this way, the critical reader is able to check our scores of the different items and can even execute a recalculation with application of other weights. Independent scoring by the two assessors resulted in a high initial agreement (94%), and a subsequent consensus meeting provided correction of the eventual individual biases. The dependent variable (the reviewers' conclusion) was even recorded by two pairs of assessors independently (also with an excellent agreement between the pairs of 92%). The fact that the authors of this article had a hand in writing three of the reviews introduced another form of information bias, which we might call reviewer bias. We were aware of this and considered exclusion of our own reviews. However, this would have had no significant influence on the conclusions of this article.

### Criteria and Weights

The choice of items and the weighing of the items in a methodological quality assessment are prone to subjective preferences. However, even in the more extensively researched field of methodological assessment of RCTs, a published validated list is still missing. Most lists used in the assessment of RCTs use an additive model, as we did. Therefore, in the absence of empirical evidence to the contrary, we think that use of a list based on an additive model is still justified. The choice of the items for our list closely matches similar recent publications on quality of reviews indicating at least concurrent construct validity. However, the process of weighing items provides more variation and cannot easily be compared among publications. Therefore, we also performed an alternative (sensitivity) analysis, applying equal weights to all items. In this procedure, the ranking of the better re-
views was almost identical (data not shown).

Quality of the Reviews Included

The methodological quality of reviews can be well assessed with the type of criteria list we used. The agreement reached among the assessors is acceptable, and the list is easy to use and robust for changes in the weighing of items. However, a list of this sort mainly assesses the description of the review procedures as such, not the quality of the performance of the review procedure. For item B (comprehensive search strategy), there was a clear indication that this was the case. This disadvantage is similar for the lists used to evaluate the quality of RCTs.82 The low review score for most reviews, indicating low methodological quality, was striking. This finding is consistent with other critical assessments of reviews in the medical field.81,82,83 The great majority of reviews lacked essential components (Table 2). This low review quality clearly indicates the need for the development of standardized high quality in the conducting and reporting of meta-analyses, similar to the recent efforts to improve the reporting of RCTs.84

The Relationship Between Review Characteristics and the Overall Conclusion of the Review

This study shows that a review of spinal manipulation is more likely to have a positive outcome if the review deals with spinal manipulation only, if a spinal manipulator was among the authors, and if a comprehensive search was described. Some characteristics were unlikely to produce significant results in our sample because of the low prevalence of the characteristic (eg, pooling, explicit quality assessment, inclusion of non-RCTs, and language).

It is difficult to generalize our remarkable findings to other fields in which reviews are written, because spinal manipulation and back pain combine certain specific methodological features.85,86 Back pain is not measured with “hard” outcomes, and in the RCTs numerous different outcome measures are applied. No uniform classification into prognostically homogeneous subgroups is available. Most importantly, the RCTs on spinal manipulation are, on the average, of a poor methodological quality.87,88 These aspects leave relatively much room for subjective interpretation of the results of the relevant effectiveness studies. For spinal manipulators as reviewers there is an increased chance of result bias, because there is a tight link of the profession to a single type of treatment. For example, cardiologists have several options to prevent reinfarction and therefore will not experience a professional threat if one of the available treatments proves to be ineffective. For spinal manipulators, however, proof of the effectiveness of their main therapy is of utmost importance for the validation of their professional identity. It is not clear whether this professional relationship with the treatment indeed caused bias in reviewing. It can be equally true that reviewers who were not spinal manipulators were biased against the therapy.

Because of the above-mentioned considerations, more empirical research on the review methods applied to other therapies in other professional fields is needed to further explore our findings.

References

3. Haynes RB. Clinical review articles—should be as scientific as the articles they review. BMJ. 1992;304:530-531.


