



The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus

Arianne P. Verhagen,^{1,4,*} Henrica C. W. de Vet,^{1,4} Robert A. de Bie,^{1,4} Alphons G. H. Kessels,^{1,4} Maarten Boers,^{2,4} Lex M. Bouter,^{3,4} and Paul G. Knipschild^{1,4}

¹DEPARTMENT OF EPIDEMIOLOGY, MAASTRICHT UNIVERSITY, MAASTRICHT, THE NETHERLANDS; ²PREVIOUSLY: DEPARTMENT OF INTERNAL MEDICINE/RHEUMATOLOGY, MAASTRICHT UNIVERSITY, MAASTRICHT, THE NETHERLANDS; CURRENTLY: DEPARTMENT OF CLINICAL EPIDEMIOLOGY, VRIJ UNIVERSITEIT, UNIVERSITY HOSPITAL, AMSTERDAM, THE NETHERLANDS; ³DEPARTMENT OF EPIDEMIOLOGY AND BIostatISTICS, INSTITUTE FOR RESEARCH IN EXTRAMURAL MEDICINE, VRIJE UNIVERSITEIT, AMSTERDAM, THE NETHERLANDS; AND ⁴NETHERLANDS SCHOOL OF PRIMARY CARE RESEARCH, MAASTRICHT, THE NETHERLANDS

ABSTRACT. Most systematic reviews rely substantially on the assessment of the methodological quality of the individual trials. The aim of this study was to obtain consensus among experts about a set of generic core items for quality assessment of randomized clinical trials (RCTs). The invited participants were experts in the field of quality assessment of RCTs. The initial item pool contained all items from existing criteria lists. Subsequently, we reduced the number of items by using the Delphi consensus technique. Each Delphi round comprised a questionnaire, an analysis, and a feedback report. The feedback report included staff team decisions made on the basis of the analysis and their justification. A total of 33 international experts agreed to participate, of whom 21 completed all questionnaires. The initial item pool of 206 items was reduced to 9 items in three Delphi rounds. The final criteria list (the Delphi list) was satisfactory to all participants. It is a starting point on the way to a minimum reference standard for RCTs on many different research topics. This list is not intended to replace, but rather to be used alongside, existing criteria lists. J CLIN EPIDEMIOL 51;12:1235–1241, 1998. © 1998 Elsevier Science Inc.

KEY WORDS. Criteria list, quality assessment, randomized clinical trials, Delphi method, consensus technique, scale development

INTRODUCTION

In recent years, the number of available randomized clinical trials (RCTs) has grown exponentially. It is therefore almost impossible for clinicians to keep up with the increase of scientific information from original research [1]. An important aim of reviewing the literature in health care is to summarize the evidence on which clinicians need to base their care and thus to provide the empirical basis for clinical decision making. The overall conclusions of a review often appear to depend on the quality of both the individual RCTs and the review process [2,3]. A clear description of the strategies for identifying, selecting, and integrating the information distinguishes a systematic review from the traditional narrative review [4,5]. Today, many systematic re-

views rely substantially on the assessment of the methodological quality of the individual trials [6–8].

“Quality” as a concept is not easy to define. Quality of RCTs has recently been defined as “the likelihood of the trial design to generate unbiased results” [9]. This definition covers only the dimension of internal validity. Although most articles proposing a criteria list to assess the methodological quality of RCTs do not explicitly define the concept of quality [10], most lists measure at least three dimensions that may encompass the concept of quality in its broadest sense: internal validity, external validity, and statistical analysis [11–15]. Some authors distinguish an ethical component in the concept of quality as well [16,17].

The method to develop a quality criteria list is similar to that of other measurement instruments, for example, “quality of life” scales [18]. Here, consensus methods are often used to select and reduce the number of items. Consensus studies are typically designed to combine the knowledge and experience of experts with the limited amount of available evidence. From the existing consensus methods, we

*Address correspondence to: Arianne P. Verhagen, Department of Epidemiology, Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands.

Accepted for publication on 6 July 1998.

chose the Delphi technique [19,20] because of the number of the participants we wanted to involve, the written procedure, the anonymity of the comments, and the time available (approximately 2 years) to conduct the study.

The aim of this study is to achieve consensus among experts, implicitly based on both empirical evidence and personal opinion, on how the quality of RCTs can be measured best, resulting in a quality criteria list. We have considered two approaches to reach this goal: try to achieve consensus on the definition of quality of RCTs and infer the necessary items for a criteria list, or, conversely, try to achieve consensus on items that, according to the participants, measure quality of a trial and infer from those a definition, or a description of the concept, of quality. We considered the latter approach to have a higher chance of success.

To be able to measure the quality of the design and conduct of a trial, one has to rely on the quality of the report. Our point of departure is the ideal situation, that is, that the report presents an honest, accurate, and comprehensive reflection of the conduct of the study. We regard the quality criteria list resulting from this study as a starting point for a future minimum reference standard to be used in systematic reviews. As such, it is not intended to replace existing criteria lists but to facilitate comparison of reviews more easily. This article presents the Delphi procedure and the resulting criteria list in quality assessment of RCTs on which experts reached consensus.

METHOD

Staff Team

A staff team was formed to initiate this research and consisted of all authors except L.M.B. All staff team members are epidemiologists, one of whom is also a clinician and one of whom has a statistical background. The others are medical doctors or health scientists. The staff team was responsible for the procedures of the selection of items and the participants and was responsible for the construction of the questionnaires, the analysis of the responses and the formulation of the feedback.

Selection of the Items

For the development of the initial item pool, we collected all items from existing quality criteria lists for RCTs. For the search strategy four sources were used: an article by Moher *et al.* [10], the doctoral thesis of Jadad [15], information from the Methods group of the Cochrane Collaboration and a Medline search on CD-ROM using the key words: *quality, assessment, methodology, randomized clinical trials, scales, checklists, quality scores, meta-analysis, epidemiology, and methods*. Papers are included when a criteria list for quality assessment of RCTs was presented. Papers were excluded when a modification of an existing list was used.

We made headings of various aspects of a design of an RCT, (e.g., aim, study question, randomization, blinding), under which all items were ordered. A total number of 17 headings (or domains) were created. On the basis of this initial item pool, we formulated the Delphi-1 questionnaire. To generate a more complete item pool, the participants were given the opportunity in Delphi-1 to add items they missed.

Selection of Participants

The participants had to be epidemiologists or statisticians concerned with quality assessment in systematic reviews or meta-analyses. We tried to achieve a wide range of different points of view on quality assessment. First, we asked all first (or co-) authors of a publication of an original quality criteria list to participate, one (co-) author per original article. Next, after an extensive brainstorm of the members of the staff team, we generated a list of leading epidemiologists and statisticians in the field of quality assessment. This resulted in three groups of experts of roughly equal sizes: authors, epidemiologists, and statisticians.

Procedure

During the whole Delphi procedure, we used structured questions, for example: "Should this item be included into the criteria list?" or "Do you agree with the rewording of this time?" The answer options used were 5-point Likert-scales (totally agree–totally disagree) or a "yes/no/don't know" answer format. We invited participants to give reasons for their choices. After each Delphi round, a feedback report was made to inform the participants about opinions and arguments of the other participants. The staff team decided, on the basis of the answers and arguments of the participants, which items and questions would appear in the next questionnaire. Staff team decisions were presented and justified in the feedback report. The participants were given the opportunity to react to, or when necessary oppose, the arguments of other participants and the decisions made by the staff team. Three or four Delphi rounds were considered sufficient to reach consensus; consensus being defined as a "general agreement of a substantial majority."

Analysis

The analysis of the responses from the Delphi rounds was both quantitative and qualitative. Quantitatively, we presented the mean scores on the 5-point Likert scales (strongly disagree [0 points], moderately disagree [1 point], neutral [2 points], moderately agree [3 points], and strongly agree [4 points]) as a percentage of the maximum obtainable score. For example: a mean score of 1.9 is 47.5% of the maximum achievable score. For questions with a "yes/no/don't know" answer format we calculated a "yes minus no" score from the number of participants who answered a "yes" on a spe-

cific question minus the number of participants who answered “no.” The necessary cut-off points were determined based on the data of each Delphi round. Qualitatively, we summarized the suggestions and comments of the participants.

Delphi-1

For every item, we asked the participants how strongly they agreed to include it in the final criteria list (5-point Likert scale). Participants were given the opportunity to suggest alternative wording and to add extra items. Some items basically asked for the same information but were formulated differently. Participants were able to choose the items in the wording they liked best.

Delphi-2

The Delphi-2 questionnaire provided opinions on the methods and results of procedural decisions made by the staff team and questions about the formulation of the items selected from Delphi-1 on which the participants agreed most. We decided, on the basis of the mixed responses in Delphi-1, to present all items not selected initially after Delphi-1 again in Delphi-2 for a second chance. Participants were able to choose the items they considered to be essential for the criteria list. Again, they were invited to give reasons for their decisions and opinions.

Delphi-3

We reworded the initial items based on the arguments given in Delphi-2, and we presented them in the Delphi-3 questionnaire. We asked whether the participants preferred the rewording or the original phrasing. Furthermore, we presented the items that received a second chance (based on the answers in Delphi-2) to be included into the criteria list. The participants were able to state which of these items should be added into the final list of items. Subsequently, we asked whether they agreed with the omission of domains not chosen in previous rounds (Delphi-1 and Delphi-2).

Definition of Quality

After Delphi-1, at the 3rd Cochrane Colloquium in Oslo in 1995 in a meeting with some of the participants, the issue was raised of whether we should continue talking about the “quality of RCTs” or whether we should limit ourselves to identifying a set of “parameters which may be related to effect sizes,” which implies a restriction to internal validity. Therefore, in Delphi-2 we asked the participants whether they had problems with using the word “quality” related to this criteria list. On the basis of their answers, we generated two possible definitions about quality, and the participants were asked in Delphi-3 which of the two different definitions they considered to be most accurate.

RESULTS

Participants

We were able to locate 15 of 17 identified authors (or co-authors) of original criteria lists. One of them refused to participate, and three did not respond. We located 13 of 19 epidemiologists, of whom two refused to respond and two did not respond. Of the 15 statisticians we located, one refused to respond and one did not respond. Potential participants declined mostly because they were too busy; only one declined because he did not like the Delphi method for this purpose. We started with 33 persons who agreed to participate, of whom 26 returned the first questionnaire and 21 the second and third questionnaires. One participant returned the second and third questionnaire. Reasons mostly mentioned for nonresponse was lack of time.

Delphi-1

A total of 24 papers were found presenting a criteria list [9,10,13,14,16,21–40]. Several articles used the same criteria list, namely the “Maastricht list” [33–39] or the list developed by Chalmers [13,40]. Once, a double publication of the same criteria list was found [27,28]. We started with 17 articles [9,10,13,14,16,21–33] after excluding articles in which a modification of the “Chalmers list” or the “Maastricht list” was used. From these criteria lists, we generated a large initial item pool of 206 items ordered under 17 domains. Of the 33 Delphi-1 questionnaires, 26 were returned and analyzed.

The initial item list generated intense disagreement: on 25% of the items ($n = 52$) five or more participants scored “strongly agree” to include this item, whereas five or more other participants scored “strongly disagree” to include that item (Table 1). The disagreement was in part due to different formulations of the items but also to the different priorities of the statisticians and the epidemiologists regarding the inclusion of statistical items. Epidemiologists stated repeatedly that items concerning the statistical analysis had nothing to do with the quality of RCTs, whereas the statisticians consider, for example, the performance of an *a priori* sample size calculation to be of importance to quality. Table 1 shows examples of items on which the participants disagreed strongly.

We saw no obvious difference in scoring between the authors and the epidemiologists, but we observed a difference when we divided the participants in statisticians on the one hand and epidemiologists + authors on the other. The statisticians scored 31 items greater than 70% of the maximum obtainable score, of which five items concerned statistical analysis and seven items concerned withdrawals or drop-outs.

STAFF TEAM DECISIONS. The aim of the staff team was a short final criteria list. On the basis of the data, we chose a rather high cut-off point of 70%, resulting in a preliminary list of seven items, to which items could be added during

TABLE 1. Some examples of items (Delphi-1) on which the participants ($n = 26$) showed strong disagreement

Items	Number of participants who answered “strongly agree”: this item <u>must</u> be included in the list	Number of participants that answered “strongly disagree”: this items <u>must not</u> be included in the list
The study design is:		
a. Poor (e.g., no comparative groups)		
b. Inadequate (e.g., comparative single blind or open)		
c. Adequate (e.g., comparative, double blind)	9	10
Is the method described used to conceal the intervention assignment schedule from participants and clinicians until recruitment was complete and irrevocable?	10	7
Was the study described as randomized (this includes the use of words such as randomly, random and randomization)?	7	6
Dates of starting and ending accession?	6	6

the procedure. The feedback report of the Delphi-1 presented all items with their scores in percentage and all comments made by the participants (anonymously). We decided to present all items of Delphi-1 again in Delphi-2 so that participants were able to reconsider their first decisions, before any definite decision on inclusion or exclusion was taken.

Delphi-2

Of the 33 Delphi-2 questionnaires sent to all initial participants, 21 were returned and analyzed. Nonresponse was mainly in the authors/epidemiologists group. The most reported reason was lack of time, and one participant was on maternity leave. Eight participants agreed with the cut-off point of 70%, whereas nine participants answered “don’t know.” The majority of the participants ($n = 15$) accepted the seven initial items to be included, but all considered rephrasing of most items necessary. Most participants chose some of the items from Delphi-1 that had a score below the 70% (second-chance items) to be included also.

STAFF TEAM DECISIONS. The data showed a large group of “second-chance items” that were never chosen or were chosen by only one or two participants; that is, most participants did not believe those items were essential. We decided to give the items chosen at least four times a final chance to be included. Table 2 presents the reworded preliminary items and the extra items receiving a final change to be included.

Delphi-3

All 21 Delphi-3 questionnaires sent to the participants of Delphi-2 were returned and analyzed. The majority of the participants accepted the rewording of the initial items. One second-chance item was added to the final criteria list because 19 participants regarded this item as essential. On

the other items, the opinion on whether or not to include was divided with roughly equal “yes” and “no” responses. We decided these items to be important but not essential and, thus, did not include them in the final criteria list. This final list is called the Delphi list (Table 3) and includes a description about the interpretation of the items as well (available upon request from the first author).

In Table 4, we present in detail the items and domains per Delphi round.

Definition of Quality

According to the majority of the participants, restriction of “quality” to “internal validity” does not capture the concept

TABLE 2. All items selected for the definitive criteria list

Items selected and reworded in Delphi-2
1. Treatment allocation
a) Was a method of randomization performed?
b) Was the treatment allocation blinded?
2. Are the groups similar at baseline regarding the most important prognostic indicators?
3. Eligibility criteria
a) Are the inclusion criteria operationalized?
b) Are exclusion criteria operationalized?
4. Was the outcome assessor blinded?
5. Was the therapist care/provider blinded?
6. Is the numerical information regarding the primary end point sufficient to enable statistical pooling?
7. Does the analysis include an intention-to-treat analysis?
Items receiving a final chance in Delphi-3 to be included also
1. Is the withdrawal/drop-out rate unlikely to cause bias?
2. Are therapeutic and control regimens/interventions operationalized?
3. Is the compliance rate (in each group) unlikely to cause bias?
4. Is controlled for cointerventions which could explain the results?
5. Was the patient blinded?
6. Is a sample size justification described?

TABLE 3. Final Delphi List after three Delphi rounds

1. Treatment allocation	
a) Was a method of randomization performed?	Yes/No/Don't know
b) Was the treatment allocation concealed?	Yes/No/Don't know
2. Were the groups similar at baseline regarding the most important prognostic indicators?	Yes/No/Don't know
3. Were the eligibility criteria specified?	Yes/No/Don't know
4. Was the outcome assessor blinded?	Yes/No/Don't know
5. Was the care provider blinded?	Yes/No/Don't know
6. Was the patient blinded?	Yes/No/Don't know
7. Were point estimates and measures of variability presented for the primary outcome measures?	Yes/No/Don't know
8. Did the analysis include an intention-to-treat analysis?	Yes/No/Don't know

of “quality,” and, consequently, a definition of quality should also contain elements of external validity and the statistical analysis. But during the process, we noticed inconsistencies, even within participants within one Delphi round. For example, a participant stated explicitly on one page that quality was only concerned with internal validity. But on the next page, the same participant suggested the inclusion of three items into the final criteria list that clearly concerned the external validity. Therefore, the staff team generated two different definitions based on the answers of Delphi-2. The first definition was: *Quality is a set of parameters in the design and conduct of a study related to effect sizes*. This definition had emerged from a workshop with some of the participants at the 3rd Cochrane Colloquium

in Oslo. The second definition was generated from the remarks in the Delphi-2 questionnaire: *Quality is a set of parameters in the design and conduct of a study that reflects the validity of the outcome, related to the external and internal validity and the statistical model used*.

The majority (n = 17) of the participants in Delphi-3 were in favor of the second definition of quality, but most of them did not like the phrasing. Only two participants preferred the first definition, and two participants answered “don't know.” The participants achieved consensus on quality being more than internal validity alone, but the staff team was not able to capture this consensus into an acceptable definition.

DISCUSSION

After three Delphi rounds, the participants achieved consensus on a generic core set of items for quality assessment in RCTs. Because of the chosen Delphi consensus procedure, we will call this list the Delphi List. In our effort to develop a criteria list, we chose not to define the word “quality” beforehand because a well-accepted definition does not exist. We assumed that the participants (all experts in the field of quality assessment) would have their own clear picture of what quality is. The advantage of a consensus method such as the Delphi approach is that the different ideas of the concept of quality integrate in the resulting criteria list, thus determining the content validity. During the process, most participants appeared to have difficulties with this approach, and we decided to try to formulate a definition of quality.

In a consensus procedure, the choice of the participants is crucial [19,20]. In the process of selecting the partici-

TABLE 4. Items and domains per Delphi round

Domains	Number of items in Delphi-1 and -2	Number of items in Delphi-3	Number of items in final Delphi list
1. Study question	2	—	—
2. Population	15	1	1
3. Sample size and power calculations <i>a priori</i>	9	1	—
4. Treatment allocation	12	1	1
5. Study design	2	—	—
6. Ethics	4	—	—
7. Intervention	19	1	—
8. Outcome measures	21	—	—
9. Follow-up/withdrawals	14	1	—
10. Blinding	28	3	3
11. Cointervention	5	1	—
12. Side-effects	5	—	—
13. Compliance	6	1	—
14. Prognostic comparability	6	1	1
15. Analysis	41	2	2
16. Conclusion	10	—	—
17. Presentation	7	—	—

pants, our aim was to achieve a broad representation of all different points of view on quality assessment using three different groups of roughly equal sizes.

In a Delphi consensus procedure, the staff team has to decide about the procedural steps [19,20]. Their decisions can vary from fully autocratic to fully democratic. Because of the expected fundamental differences, we assumed that a too-directive role would be ineffective. Therefore, we decided to allow all Delphi-1 items for a second chance. The data of Delphi-2 showed much more agreement, and we considered that a consensus could be achieved. After Delphi-3, the participants seemed satisfied with the resulting criteria list, and we believed no new arguments were given, so a fourth round would probably not add new or different information.

Based on the comments and remarks of the participants during the whole procedure, an Appendix has been constructed on the interpretation of the items. The reviewers have to decide, depending on the topic of the review, whether enough information is provided to score a "yes" on certain items. As long as these decisions are stated explicitly in the review, it will be clear for the reader how the items are scored and a comparison with reviews on other topics using the same criteria list can be made.

Empirical research concerning assessment of the methodological quality of RCTs is relatively new. Awaiting empirical research, we think it is useful to prioritize items using a group of experts. All different opinions in this field of research should be respected at this stage. Starting this research, we were well aware of the different views on quality and quality assessment of RCTs. Despite this knowledge, we were surprised by the many initial differences between the participants. Notwithstanding these differences, the participants achieved consensus on the final Delphi List. New in the ongoing discussion about quality is that we achieved broad consensus concerning the need for inclusion of three dimensions of quality into any definition of the "concept" of quality: "internal validity," "external validity," and "statistical considerations." In the feedback of Delphi-3, in which we presented the Delphi List to the participants as a result of this research project, we asked participants to react to the final result. No negative and four positive reactions or comments were received.

When a consensus concerning the content of a criteria list is reached, the following issue of what to do with the results of quality assessment has to be addressed. A quality criteria list can be used in different ways [33,38,40,41]. It can provide a quality score as an estimate of the methodological quality. These quality scores can be used as a "threshold score" for inclusion of the article in a review, as a "weighting factor" in the statistical analysis [40,42,43], or as the input sequence in a cumulative meta-analysis [43–45]. Sometimes a visual plot of the effect size against a quality score is presented [40,42,43]. The next step will be to achieve consensus (based on empirical evidence) about

how to incorporate quality into the final conclusions of a systematic review or meta-analysis.

CONCLUSION

The participants in this Delphi process achieved consensus on a generic criteria list for quality assessment in RCTs: the Delphi List. The adoption of this core set by the participants and other researchers may be the first step toward a minimum reference standard of quality measures for all RCTs. It is not our intention to replace existing criteria lists, but we suggest it should be used alongside these lists. The validity of this criteria list will have to be measured and evaluated over time.

The authors thank the following persons for their participation: D.G. Altman, E. Andrew, J. Berlin, L.M. Bouter, S.A. Brown, M.K. Cho, M. Clarke, K. Dickersin, M. Evans (and A.V. Pollock), C. Friedenreich, P.C. Gøtzsche, S. Greenland, J. van Houwelingen, T.E. Imperiale, J. Lau, C. Mulrow, M. Nurmohamed, I. Olkin, P. Onghena, G. ter Riet, H. Sacks, K.F. Schultz, K. Smith, P. Tugwell, and S. Yusuf. Their participation in this project does not necessarily mean that they fully agree with the final criteria list, but the criteria list is the result of a "communis opinio."

References

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. **Clinical Epidemiology: A Basic Science for Clinical Medicine**. 2nd edition. Boston: Little-Brown; 1991: 359–378.
2. Haynes RB. Clinical review articles: Should be as scientific as the articles they review [editorial]. **BMJ** 1992; 304: 303–331.
3. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. **J Clin Epidemiol** 1991; 44: 1271–1278.
4. Dickersin K, Berlin JA. Meta-analysis: state of the science. **Epidemiol Rev** 1992; 14: 154–176.
5. Mulrow CD. Rationale for systematic reviews. **BMJ** 1994; 309: 597–599.
6. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. **Can Med Assoc J** 1988; 138: 697–703.
7. Mulrow CD. The medical review article: State of the science. **Ann Intern Med** 1987; 106: 485–488.
8. Mulrow CD, Oxman AD, Eds. **Cochrane Collaboration Handbook** [updated 1 March 1997]. In: The Cochrane Library [database on disk and CDRom]. The Cochrane Collaboration. Oxford: Update Software; 1996–. Updated quarterly.
9. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? **Control Clin Trials** 1996; 17: 1–12.
10. 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. **Control Clin Trials** 1995; 16: 62–73.
11. The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials: Special communication. **JAMA** 1994; 272: 1926–1931.
12. Assendelft WJJ, Koes BW, van der Heijden GJMG, Bouter LM. The efficacy of chiropractic for back pain: Blinded review of the relevant randomized clinical trials. **J Manipulative Physiol Ther** 1992; 15: 487–494.
13. Chalmers TC, Smith H Jr, Blackburn B, Silverman B,

- Schroeder B, Reitman D, *et al.* A method for assessing the quality of a randomized control trial. **Control Clin Trials** 1981; 2: 31–49.
14. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. **Stat Med** 1989; 8: 441–454.
 15. Jadad AR. Meta-analysis of randomised clinical trials in pain relief. (PhD Thesis). Oxford: University of Oxford; 1994.
 16. Andrew E. Method for assessment of the reporting standard of clinical trials with Röntgen contrast media. **Acta Radiol Diag** 1984; 25: 55–58.
 17. Lumley J, Bastian H. Competing or complementary? Ethical considerations and the quality of randomized trials. **Int J Technol Assess Health Care** 1996; 12: 247–263.
 18. Jaeschke R, Guyatt GH. How to develop and validate a new quality of life instrument. In: Spilker B, ed. **Quality of Life Assessments in Clinical Trials**. New York: Raven Press; 1990.
 19. Delbecq AL, Ven AH van de, Gustafson DH. Group techniques for program planning; a guide to nominal group and Delphi processes. Glenview: Scott, Foresman; 1975.
 20. Dalkey NC, Helmer O. An experimental application of the Delphi-method to the use of experts. **Manage Sci** 1963.
 21. Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. **Br J Surg** 1985; 72: 256–260.
 22. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. **Control Clin Trials** 1989; 10: 31–56.
 23. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. **Pediatrics** 1989; 84: 815–827.
 24. Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, *et al.* A cohort study of summary reports of controlled trials. **JAMA** 1990; 263: 1401–1405.
 25. Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? **Ann Intern Med** 1990; 113: 299–307.
 26. Spitzer WO, Lawrence V, Dales R, *et al.* Links between passive smoking and disease: A best evidence synthesis. **Clin Invest Med** 1990; 13: 17–42.
 27. Brown SA. Measurement of quality of primary studies for meta-analysis. **Nurs Res** 1991; 40: 352–355.
 28. Brown SA. Meta-analysis of diabetes patient education research. **Res Nurs Health** 1992; 15: 409–419.
 29. Nurmohamed M, Rosendaal F, Buller H, Dekker E, Hommes D, Vandenbroucke J, *et al.* Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: A meta-analysis. **Lancet** 1992; 340: 152–156.
 30. Onghena P, van Houdenhove B. Antidepressant-induced analgesia in chronic non-malignant pain: A meta-analysis of 39 placebo-controlled studies. **Pain** 1992; 49: 205–219.
 31. Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: A meta-analysis. **Am Rev Respir Dis** 1992; 145: 533–539.
 32. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. **JAMA** 1994; 272: 101–104.
 33. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM, Knipschild PG. Spinal manipulation and mobilization for back and neck pain: A blinded review. **BMJ** 1991; 303: 1298–1303.
 34. Beckerman H, de Bie RA, Bouter LM, De Cuyper HJ, Oostendorp RAB. The efficacy of laser therapy for musculoskeletal and skin disorders: a criteria-based meta-analysis of randomized clinical trials. **Phys Ther** 1992; 72: 483–491.
 35. van der Heijden GJMG, Bouter LM, Beckerman H, de Bie RA, Oostendorp RAB. De effectiviteit van ultra geluid bij aandoeningen van het bewegings apparaat. **Ned T Fysiother** 1991; 101: 169–177.
 36. Kleijnen J, Knipschild P, ter Riet G. Clinical trials of homoeopathy. **BMJ** 1991; 302: 316–323.
 37. Knipschild PG. Trials and errors; alternative thoughts on the methodology of clinical trials. **BMJ** 1993; 306: 1706–1707.
 38. Knipschild PG. Systematic reviews: Some examples. **BMJ** 1994; 309: 719–721.
 39. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: A criteria-based meta-analysis. **J Clin Epidemiol** 1990; 43: 1191–1199.
 40. Detsky AS, Naylor CD, Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta analysis. **J Clin Epidemiol** 1992; 45: 255–265.
 41. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. **JAMA** 1995; 273: 408–412.
 42. Jenicek M. Meta-analysis in medicine: Where we are and where we want to go. **J Clin Epidemiol** 1989; 42: 35–44.
 43. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: Current issues and future directions. **Int J Technol Assess** 1996; 12: 195–208.
 44. Koes BW, van Tulder MW, van der Windt DAWM, Bouter LM. The efficacy of back schools: A review of randomized clinical trials. **J Clin Epidemiol** 1994; 47: 851–862.
 45. Koes BW, Assendelft WJJ, van der Heijden GJMG, Bouter LM. Spinal manipulation for low back pain; an updated systematic review of randomized clinical trials. **Spine** 1996; 221: 2860–2873.