Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?

Alejandro R. Jadad, MD, DPhil; R. Andrew Moore, DPhil; Dawn Carroll, RGN; Crispin Jenkinson, DPhil; D. John M. Reynolds, DPhil; David J. Gavaghan, DPhil; and Henry J. McQuay DM

Oxford Regional Pain Relief Unit (A.R.J., R.A.M., D.C., H.J.M.); Nuffield Department of Anaesthetics (A.R.J., R.A.M., D.C., D.J.G., H.J.M.); Department of Public Health and Primary Care (C.J.); and University Department of Clinical Pharmacology (D.J.M.R.), University of Oxford, Oxford, UK

ABSTRACT: It has been suggested that the quality of clinical trials should be assessed by blinded raters to limit the risk of introducing bias into meta-analyses and systematic reviews, and into the peer-review process. There is very little evidence in the literature to substantiate this. This study describes the development of an instrument to assess the quality of reports of randomized clinical trials (RCTs) in pain research and its use to determine the effect of rater blinding on the assessments of quality. A multidisciplinary panel of six judges produced an initial version of the instrument. Fourteen raters from three different backgrounds assessed the quality of 36 research reports in pain research, selected from three different samples. Seven were allocated randomly to perform the assessments under blind conditions. The final version of the instrument included three items. These items were scored consistently by all the raters regardless of background and could discriminate between reports from the different samples. Blind assessments produced significantly lower and more consistent scores than open assessments. The implications of this finding for systematic reviews, meta-analytic research and the peer-review process are discussed. *Controlled Clin Trials* 1996; 17:1–12

KEY WORDS: *Pain, meta-analysis, randomized controlled trials, quality, health technology assessment*

INTRODUCTION

The use of reliable data to support medical and public health decisions is essential if the growing demand for health care is to be met from limited resources. Determining the effectiveness of medical interventions from clinical research data is not an easy task, especially if studies addressing the same therapeutic problem produce conflicting results. The assessment of the validity of the primary studies

has been identified as one of the most important steps of the peer-review process [1] and as one of the key components of systematic reviews [2, 3]. For more than 10 years it has been suggested that the validity or quality of primary trials should be assessed under blind conditions in order to reduce or avoid the introduction of selection bias into meta-analyses and systematic reviews [4]. Similar suggestions have been made in relation to the peer review process [5], but there is no empirical evidence to substantiate any of these claims [5, 6].

There are three methods to assess the quality of clinical trials: individual markers, checklists, and scales [7]. Scales have the theoretical advantage over the other methods in that they provide quantitative estimates of quality that could be replicated easily and incorporated formally into the peer review process and into systematic reviews. The main disadvantage of quality scales is that there is a dearth of evidence to support the inclusion or exclusion of items and to support the numerical scores attached to each of those items. In a recent search of the literature, 25 scales designed to assess the quality of primary trials were identified, but only one had been developed following established methodological procedures (8). In this paper we describe the development of such a scale and its use to evaluate the effect of blinding on the assessments of quality.

METHODS

Established methodological procedures suggested for the development and validation of any other health measurement tool were followed. They included preliminary conceptual decisions; item generation and assessment of face validity (sensibility); field trials to assess consistency, frequency of endorsement, and construct validity; and the generation of a refined instrument [9, 10].

Preliminary Conceptual Decisions

During the development of an instrument, it is important to define the entity to be measured and the framework within which the instrument will be used. In this particular case, the purpose of the instrument was to assess quality, defined as the likelihood of the trial design to generate unbiased results and approach the "therapeutic truth." This has also been described as "scientific quality" [11]. Other trial characteristics such as clinical relevance of the question addressed, data analysis and presentation, literary quality of the report, or ethical implications of the study were not encompassed by our definition. The aims of the instrument were (1) to assess the scientific quality of any clinical trial in which pain is an outcome measure or in which analgesic interventions are compared for outcomes other than pain (e.g., a study looking at the adverse effect profile of different opioids), and (2) to allow cons·stent and reliable assessment of quality by raters with different backgrounds, including researchers, clinicians, and professionals from other disciplines and members of the general public.

Item Generation and Assessment of Item Face Validity

A multidisciplinary panel of judges with an interest in pain research and/or experience in instrument development was assembled. The definition of quality

and the purposes of the instrument were discussed with each of the judges. They were given 2 weeks to produce a list with preliminary items to be considered for inclusion in the instrument. To generate the items, the judges referred to the criteria published in previous instruments and used their own judgment. Once they had generated the items, they sent them to one of the investigators (ARJ) who produced a single list with all the items nominated by each of the judges.

Using a modified nominal group approach to reach consensus [12], the judges assessed the face validity of each of the items according to established criteria [9]. Those items associated with low face validity were deleted. An initial instrument was created from the remaining items.

The initial instrument was pretested by three raters on 13 study reports. The raters identified problems in clarity and/or application of each of the items. The panel of judges then modified the wording of the items accordingly and produced detailed instructions describing how each of the items should be assessed and scored. The items were classified by their ability to reduce bias (direct or indirectly) and individual scores were allocated to them by consensus.

Assessment of Frequency of Endorsement, Consistency, and Validity

Frequency of Endorsement

The frequency of endorsement was calculated by dividing the number of times each item was scored by the maximum possible number of times each of the items could have been scored, multiplied by 100. Items with very high or very low endorsement rates were eliminated because they provided little discriminative power. Items which scored similarly on excellent quality reports and poor quality reports would not help to separate excellent from poor and would just make the test more time consuming. Items with frequency of endorsement below 15% or above 85% were excluded. These values were selected *a priori* from the recommended range [10].

Consistency and Binding of the Assessments

Consistency (also known as reliability), the prime requirement of scientific information [9], refers to the level of agreement between different observations of the same entity by the same rater (intrarater consistency) or different raters (interrater consistency), or under different conditions. In this study, interrater reliability was evaluated by assessing the degree to which different individuals agreed on the scientific quality of a set of reports.

Raters were included in three categories defined *a priori*: researchers, clinicians, and others. An individual was considered a researcher if she/he had participated as an investigator in five or more randomized controlled trials (RCTs) in pain relief. Clinicians were defined as individuals involved in the management of patients with acute and chronic pain conditions for more than a year but who had participated in fewer than five RCTs in pain relief. Those raters who were neither defined as researchers nor clinicians were described as "other." All raters were recruited from the staff of the Oxford Regional Pain Relief Unit, visiting staff, and related profes-

sionals. The selection was made on the basis of interest in the study and time availability. Each rater was given the same set of reports as follows.

The raters were allocated randomly (by using a random numbers table) to open or blind assessment of the quality of the reports. Those raters allocated to blind assessments were given reports in which the authors' names and affiliation, the names of the journals, the date of publication, the sources of financial support for the study, and the acknowledgments were deleted. The raters were asked to assess the quality of the reports independently. No special training was given in how to score the items. Raters were told that there were no right or wrong answers and that it should take them less than 10 minutes to score each report.

Intraclass correlation coefficients (ICCs) and their 95% confidence intervals (95% CI) were used to measure the agreement between raters. ICCs and 95% CI were calculated according to the method described by Shrout and Fleiss [13]. Values for ICCs range from 0 to 1. The closer the values to 1 the better the agreement. Although any cutoff value is arbitrary, it was decided *a priori* that the value of ICCs should be greater than 0.5 for the criterion to be considered sufficiently reliable and greater than 0.65 to represent a high level of agreement [11].

Validity

Validity was defined as the ability of the instrument to measure what it is believed it is measuring. Assessing the accuracy with which an instrument measures a construct such as quality involves making predictions and testing them [10].

Study reports were selected from three different samples. Efforts were made to locate studies previously judged as excellent or poor, through personal contact with members of the panel of judges, external researchers, and clinicians. Given the lack of a single standard, the decision on the judged quality was made *a priori* using the definition of quality just described. The rest of the studies were selected randomly from a set of controlled studies published between 1966 and 1991, which had been identified by a high yield MEDLINE strategy [15]. The articles were presented to the raters in an order determined using a random number table.

Three different overall scores were calculated for each study report: the first score was obtained by adding the individual scores of all the items of the initial instrument. The second value was obtained by adding the scores of items with frequency of endorsement between 15 and 85%. The third score was calculated by adding the scores only of those items directly related to bias reduction. The primary outcome was the score obtained with items directly related to bias reduction and whose frequency of endorsement was between 15 and 85%.

Two predictions were made before the reports were given to the raters in order to test construct validity: (1) the mean overall scores for reports judged as excellent would be higher than for those selected randomly, and (2) the mean overall scores of reports regarded as excellent and those selected randomly would be higher than the overall score of studies regarded as poor.

The mean overall scores were compared using an unpaired *t* test. Probability values of less than 0.05 were considered significant. Data were expressed as mean and standard error of the mean.

Table 1 Details of Judges and Raters

Judge No.	Sex	Background	Assessment	Comments
1	F	Clinician	Open	Rater
2	F	Clinician	Blind	Rater
3	F	Researcher	Blind	Judge and rater
4	F	Clinician	Blind	Rater
5	M	Researcher	Open	Judge and rater
6	F	Other	Open	Rater
7	M	Other	Blind	Judge and rater
8	F	Clinician	Open	Rater
9	M	Researcher	Open	Judge and rater
10	M	Researcher	Open	Judge and rater
11	M	Clinician	Blind	Rater
12	M	Other	Open	Judge and rater
13	F	Other	Blind	Rater
14	M	Clinician	Open	Rater (excluded)
15	M	Other	Blind	Rater

Generation of a Refined Instrument

A refined instrument would be produced only if (1) overall agreement was good (ICC > 0.5), and (2) it was possible to differentiate between the three types of study reports.

If appropriate, the final version of the instrument would include a list of instructions to score the items.

RESULTS

Judges, Raters, and Reports

The six judges were a psychologist, a clinical pharmacologist, a biochemist, two anaesthetists, and a research nurse with full-time involvement in pain relief activities. Thirty-six reports were selected for scoring. Seven had been judged previously as excellent, 6 as poor, and the remaining 23 were chosen randomly.

Fifteen raters, eight men and seven women, were recruited to score the 36 reports. Four of the raters were regarded as researchers, six as clinicians, and five as "others." As was the case during other instrument development exercises [14], all the judges participated in the scoring process (four as researchers and two as others). Seven raters performed open assessments and eight scored the reports under blind conditions (Table 1). One rater, a clinician allocated to open assessment, was excluded from analysis because he recorded the scores incorrectly and it was impossible to determine to which report each score referred.

Initial Instrument

The judges selected separately 49 nonredundant items (Table 2). Thirty-eight items were excluded during the consensus meeting because of poor face validity. The remaining 11 items were transformed to questions and included in the initial instrument (Table 3). Each affirmative answer was given one point. If the trial was described as randomized and/or double-blind additional points could be awarded (one extra point in each case) if the method of randomization and/or

Table 2   Items Considered by Individual Judges

1.	Random allocation	(5)
2.	Blinding	(5)
3.	Clear/validated outcomes	(5)
4.	Description of withdrawals and dropouts	(5)
5.	Clear hypothesis and objectives	(4)
6.	Clear inclusion/exclusion criteria	(4)
7.	Power calculation	(4)
8.	Appropriate size	(3)
9.	Intention to treat	(3)
10.	Single observer	(3)
11.	Adequate follow-up	(3)
12.	Negative/positive controls	(3)
13.	Controlled cointerventions	(3)
14.	Appropriate analysis	(3)
15.	Randomization method explained	(2)
16.	Description of investigators and assessors	(2)
17.	Description of interventions	(2)
18.	Raw data available	(2)
19.	Compliance check	(2)
20.	Adverse effects documented clearly	(2)
21.	Comparable groups	(2)
22.	Clinical relevance	(1)
23.	Protocol is followed	(1)
24.	Informed consent	(1)
25.	Adequate analysis	(1)
26.	Appropriate outcome measures	(1)
27.	Data supporting conclusions	(1)
28.	Paper clear and simple to understand	(1)
29.	Ethical approval	(1)
30.	Appropriate study	(1)
31.	Independent study	(1)
32.	Overall impression	(1)
33.	Prospective study	(1)
34.	More than 1 assessment time	(1)
35.	Attempt to demonstrate dose response with new agents	(1)
36.	Appropriate duration of study	(1)
37.	Description of selection method	(1)
38.	Definition of method to record adverse effects	(1)
39.	Definition of methods for adverse effect management	(1)
40.	Objective outcome measurements	(1)
41.	Avoidance of data unrelated to the question addressed	(1)
42.	Representative sample	(1)
43.	Statistics, central tendency, and dispersion measures reported	(1)
44.	Blinding testing	(1)
45.	Results of randomization reported	(1)
46.	Analysis of impact of withdrawals	(1)
47.	Clear tables	(1)
48.	Clear figures	(1)
49.	Clear retrospective analysis	(1)

The number in parentheses indicates the number of judges who suggested each of the items.

double blinding was appropriate. Conversely, points could be deducted (one point in each case) if the study was described as randomized or double blind, but the methods were inappropriate. An instruction sheet was appended to the initial instrument.

Table 3   Initial Instrument and Frequency of Endorsement

Related Directly to the Control of Bias	
Items	Endorsement Frequency (%)
1. Was the study described as randomized?[a]	63
2. Was the study described as double-blind?[a]	35
3. Was there a description of withdrawals and drop outs?	54

Other Markers Not Related Directly to the Control of Bias	
Items	Endorsement Frequency (%)
4. Were the objectives of the study defined?	91
5. Were the outcome measures defined clearly?	88
6. Was there a clear description of the inclusion and exclusion criteria?	71
7. Was the sample size justified (e.g., power calculation)?	10
8. Was there a clear description of the interventions?	87
9. Was there at least one control (comparison) group?	92
10. Was the method used to assess adverse effects described?	41
11. Were the methods of statistical analysis described?	73

[a]The endorsement frequency for the appropriateness of the method to generate the sequence of randomization was 15%, and for double-blinding it was 34%. The frequency of endorsement for concealment of randomization was evaluated separately, and it was 6%.

Field Trial

Frequency of Endorsement

Each item was scored 504 times. The frequency of endorsement of individual items ranged from 10 to 92% (Table 3). Four items (definition of the objectives of the study, definition of the outcome measures, description of the interventions, and presence of a control group) were excluded because of the high frequency of endorsement. Only one item was excluded because of the low frequency of endorsement (justification of sample size). The remaining six items had frequencies of endorsement ranging from 15 to 73%. Three of those items, randomization, double blinding, and description of withdrawals and dropouts, were considered as directly related to bias reduction (Table 3).

The maximum possible score produced was 13 points by the initial instrument (11 items); 8 points by the 6 items with adequate frequency of endorsement; and 5 points by the 3 items directly related to bias reduction (Table 3).

Scores were calculated with all the items in the initial instrument (11-item score), with the 6 items selected after assessment of frequency of endorsement (6-item score), and with the 3 items directly related to bias reduction (3-item score).

Interrater Consistency

The overall agreement among the 14 raters was high for scores calculated with either 11, 6, or 3 items (Table 4). All groups of raters produced reliable scores. However, researchers produced more consistent scores than clinicians. Both of these groups were more consistent than the "others" (Table 4). The 3-item scale showed the highest levels of agreement, overall and within groups.

Table 4 Interrater Agreement and Construct Validity

Interrater Agreement [ICC (95% CI)]			
Raters	11 Items	6 Items	3 Items
Researchers ($n = 4$)	0.69 (0.48, 0.83)	0.75 (0.58, 0.84)	0.77 (0.60, 0.86)
Clinicians ($n = 5$)	0.63 (0.44, 0.78)	0.66 (0.47, 0.80)	0.67 (0.48, 0.80)
Others ($n = 5$)	0.50 (0.26, 0.71)	0.56 (0.32, 0.76)	0.56 (0.36, 0.75)
All ($n = 14$)	0.59 (0.46, 0.74)	0.65 (0.51, 0.77)	0.66 (0.53, 0.79)

Construct Validity [Mean (Standard Error of the Mean)]			
	Overall Score		
Report Sample	11 Items	6 Items	3 Items
Previously judged as excellent ($n = 7$)	9.9 (0.2)[a]	5.7 (0.2)[a]	3.4 (0.1)[a]
Seleted at random ($n = 23$)	8.3 (0.1)[b]	4.5 (0.1)[b]	2.7 (0.1)[b]
Previously judged as poor ($n = 6$)	5.0 (0.2)	2.0 (0.1)	0.7 (0.1)
All ($n = 36$)	8.0 (0.1)	4.3 (0.1)	2.5 (0.1)

[a] Significantly higher than randomly selected and poor study reports ($p < 0.001$).
[b] Significantly higher than poor study reports ($p < 0.001$).

Construct Validity

The mean overall score for the 36 reports using the 3-item instrument was 2.5; the measurements ranged from 0 to 5. The scores for reports regarded as excellent were significantly higher than for reports selected at random and both groups of studies received significantly higher scores than those reports judged as poor, with the 11-, 6-, and 3-item instruments (Table 4). The individual scores given to the reports covered the whole spectrum, from 0 to the maximum possible, regardless of the total number of items used.

All the reports judged as poor scored four points or less with the 6-item instrument and 99% scored two points or less on the 3-item scale. For reports judged excellent, 77% scored more than four points with the 6-item instrument and 71% more than two points with the 3-item tool.

Final Version of the Instrument

The final version of the instrument contained the three items related directly to the reduction of bias and whose frequency of endorsement was between 15 and 85% (Appendix). The items were presented as questions to elicit yes or no answers. Points awarded for items 1 and 2 depended on the quality of the description of the methods to generate the sequence of randomization and/or on the quality of the description of the method of double blinding. If the trial had been described as randomized and/or double blind, but there was no description of the methods used to generate the sequence of randomization or the double-blind conditions, one point was awarded in each case. If the method of generating the sequence of randomization and/or blinding had been described, one additional

Table 5   Open vs. Blind Assessments (6- and 3-Item Instruments)

	Overall Score		ICC (95% CI)	
Assessment	6 Item	3 Item	6 Item	3 Item
Open	4.6 (0.1)[a]	2.7 (0.1)[b]	0.58 (0.44, 0.73)	0.56 (0.39, 0.58)
Blind	4.1 (0.1)	2.3 (0.1)	0.72 (0.61, 0.84)	0.76 (0.65, 0.86)

Overall score is expressed as mean (standard error of the mean) and the numbers have been rounded.
[a] $p < 0.001$.
[b] $p < 0.01$.

point was given to each item if the method was appropriate. A method to generate randomization sequences was regarded as adequate if it allowed each study participant to have the same chance of receiving each intervention, and if the investigators could not predict which intervention was next. Double blinding was considered appropriate if it was stated or implied that neither the person doing the assessment nor the study participant could identify the intervention being assessed. Conversely, if the method of generating the sequence of randomization and/or blinding was described but not appropriate, the relevant item was given zero points (Appendix). The third item, withdrawals and dropouts, was awarded zero points for a negative answer and one point for a positive. For a positive answer, the number of withdrawals and dropouts and the reasons had to be stated in each of the comparison groups. If there were no withdrawals, it should have been stated in the report (Appendix).

Open vs. Blind Assessments

Blind assessment of the reports produced significantly lower and more consistent scores than open assessments using either the 6- or the 3-item scale (Table 5).

DISCUSSION

This study describes the development of an instrument to assess the quality of clinical reports in pain relief and its use to evaluate the impact that blinding the raters could have on the assessments. The instrument is simple, short, reliable, and apparently valid. Given that none of the three items included in the final version of the instrument is specific to pain reports, it may have applications in other areas of medicine. All these items are very similar to the components of a scale used extensively to assess the effectiveness of interventions during pregnancy and childbirth [16] and are also part of most of the other available scales. There is empirical evidence to support the role of randomization and double blinding in bias reduction. It has been shown, for instance, that nonrandomized trials or RCTs that do not use a double-blind design are more likely to show advantage of an innovation over a standard treatment [17]. In a more recent study, the analysis of 250 trials from 33 meta-analyses showed that RCTs in which treatment allocation was inadequately concealed, or in which concealment of allocation was unclear, yielded significantly larger estimates of treatment effects than those trials in which concealment was adequate ($p < 0.001$) [18]. In the same study, trials

not using double blinding also yielded significantly larger estimates of treatment effects ($p < 0.01$).

The instrument could be used by researchers and referees to assess study protocols; by editors and readers of journals to identify scientifically sound reports; by researchers to monitor the likelihood of bias in research reports, including their own; by individuals (not necessarily with research or clinical experience in pain relief) involved in systematic reviews or meta-analyses to perform differential analysis based on the quality of the individual primary studies; and by patients to evaluate the validity of the evidence presented to them by health professionals.

It was suggested more than 10 years ago that the quality of clinical reports should be assessed under blind conditions to reduce the likelihood of selection bias systematic reviews and meta-analyses [4]. Blind assessment of the validity of clinical research reports has also been suggested as part of the peer review process [6]. The lack of evidence to support this practice has been highlighted recently [5, 6]. We found that the blinded assessment produced significantly lower and more consistent scores than open assessment. This may be very important for editors of journals to reduce bias in manuscript selection which could be introduced by open peer review, and for researchers if cutoff scores are recommended for inclusion and exclusion of trials from systematic reviews or if quality scores are used to weight the results of primary studies for use in subsequent meta-analysis [19–21].

A major disadvantage of the instrument described in this paper and of most others is that assessments of quality depend on the information available in the reports. Space constraints in most printed journals, the referral of readers to previous publications as the sources for detailed description of the methods of the trial, and the publication of trials in abstract form could lead to the assumption that the trial was methodologically deficient, even when the trial had been designed, conducted, and analyzed appropriately. Such a disadvantage could be avoided in the future if journals adopted more uniform requirements for trial reporting [7, 22].

APPENDIX

Instrument to Measure the Likelihood of Bias in Pain Research Reports

This is not the same as being asked to review a paper. It should not take more than 10 minutes to score a report and there are no right or wrong answers.

Please read the article and try to answer the following questions (see attached instructions):

1. Was the study described as randomized (this includes the use of words such as randomly, random, and randomization)?
2. Was the study described as double blind?
3. Was there a description of withdrawals and dropouts?

Scoring the items:

Either give a score of 1 point for each "yes" or 0 points for each "no." There are no in-between marks.

Give 1 additional point if:	For question 1, the method to generate the sequence of randomization was described **and** it was **appropriate** (table of random numbers, computer generated, etc.)
and/or:	If for question 2 the method of double blinding was described **and** it was **appropriate** (identical placebo, active placebo, dummy, etc.)
Deduct 1 point if:	For question 1, the method to generate the sequence of randomization was described **and** it was **inappropriate** (patients were allocated alternately, or according to date of birth, hospital number, etc.)
and/or:	For question 2, the study was described as double blind but the method of blinding was **inappropriate** (e.g., comparison of tablet vs. injection with no double dummy)

Guidelines for Assessment

1. Randomization

A method to generate the sequence of randomization will be regarded as appropriate if it allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which treatment was next. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should be not regarded as appropriate.

2. Double blinding

A study must be regarded as double blind if the word "double blind" is used. The method will be regarded as appropriate if it is stated that neither the person doing the assessments nor the study participant could identify the intervention being assessed, or if in the absence of such a statement the use of active placebos, identical placebos, or dummies is mentioned.

3. Withdrawals and dropouts

Participants who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number and the reasons for withdrawal in each group must be stated. If there were no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given no points.

REFERENCES

1. Kassirer JP, Campion EW: Peer review: crude and understudied, but indispensable. JAMA 272:96–97, 1994
2. Chalmers I: Evaluating the effects of care during pregnancy and childbirth. In Effective Care in Pregnancy and Childbirth, Chalmers I, Enkin M, Keirse MJNC, eds: Oxford, Oxford University Press, 1989
3. Oxman AD, Guyatt GH: Guidelines for reading literature reviews. Can Med Assoc J 138:697–703, 1988
4. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A: A method for assessing the quality of a randomized control trial. Controlled Clin Trials 2:31–49, 1981
5. Irwig I, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F: Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 120:667–676, 1994

6. Fisher M, Friedman SB, Strauss B: The effects of blinding on acceptance of research papers by peer review. JAMA 272:143-6, 1994

7. Moher D, Jadad AR, Tugwell P: Assessing the quality of randomized controlled trials: current issues and future directions. Int J Tech Asses Health Care (In press), 1995

8. 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 16:62-73, 1995

9. Feinstein AR: Clinimetrics. New Haven, Yale University Press, 1987.

10. Streiner DL, Norman GR: Health measurement scales: a practical guide to their development and use. New York, Oxford University Press, 1989.

11. Oxman AD, Guyatt GH: Validation of an index of the quality of review articles. J Clin Epidemiol 44:1271-1278, 1991

12. Fink A, Kosecoff J, Chassin M, Brook RH: Consensus methods: characteristics and guidelines for use. Am J Public Health 74:979-983, 1984

13. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420-428, 1979

14. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, Streiner DL: Agreement among reviewers of review articles. J Clin Epidemiol 44:91-98, 1991

15. Jadad AR, McQuay HJ: A high-yield strategy to identify randomized controlled trials for systematic reviews. Online J Curr Clin Trials 2:(27 Feb.)Dec 33, 1993

16. Chalmers I, Enkin M, Keirse MJNC: Effective care in pregnancy and childbirth. Oxford, Oxford University Press, 1989

17. Colditz GA, Miller JN, Mosteller F: How study design affects outcomes in comparisons of therapy. I. Medical Stat Med 8:411-454, 1989

18. 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 273:408-412, 1995

19. McNutt RA, Evans AT, Fletcher RH, Fletcher SW: The effects of blinding on the quality of peer review: a randomized trial. JAMA 263:1371-1376, 1990

20. Nurmohamed MT, Rosendaal FR, Buller HR, Dekker E, Hommes DW, Vandenbroucke JP, Briet E: Low molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. Lancet 340:152-156, 1992

21. Fleiss JL, Gross AJ: Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. J Clin Epidemiol 44:127-139, 1991

22. 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 45: 255-265, 1992