J Clin Epidemiol Vol. 44, No. 11, pp. 1271–1278, 1991
Printed in Great Britain. All rights reserved

0895-4356/91 $3.00 + 0.00
Copyright © 1991 Pergamon Press plc

VALIDATION OF AN INDEX OF THE QUALITY OF REVIEW ARTICLES

ANDREW D. OXMAN[1,2] and GORDON H. GUYATT[2,3]

Departments of [1]Family Medicine, [2]Clinical Epidemiology & Biostatistics and [3]Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

(*Received* 14 *February* 1991)

Abstract—The objective of this study was to assess the validity of an index of the scientific quality of research overviews, the Overview Quality Assessment Questionnaire (OQAQ). Thirty-six published review articles were assessed by 9 judges using the OQAQ. Authors reports of what they had done were compared to OQAQ ratings. The sensibility of the OQAQ was assessed using a 13 item questionnaire. Seven *a priori* hypotheses were used to assess construct validity. The review articles were drawn from three sampling frames: articles highly rated by criteria external to the study, meta-analyses, and a broad spectrum of medical journals. Three categories of judges were used to assess the articles: research assistants, clinicians with research training and experts in research methodology, with 3 judges in each category. The sensibility of the index was assessed by 15 randomly selected faculty members of the Department of Clinical Epidemiology and Biostatistics at McMaster. Authors' reports of their methods related closely to ratings from corresponding OQAQ items: for each criterion, the mean score was significantly higher for articles for which the authors responses indicated that they had used more rigorous methods. For 10 of the 13 questions used to assess sensibility the mean rating was 5 or greater, indicating general satisfaction with the instrument. The primary shortcoming noted was the need for judgement in applying the index. Six of the 7 hypotheses used to test construct validity held true. The OQAQ is a valid measure of the quality of research overviews.

Meta-analysis   Validity of results   Bias   Research design   Publishing standards   Peer review   Information dissemination

INTRODUCTION

While review articles have long offered clinicians a valuable source of information and guidelines for practice, the increasing volume of medical literature has increased their importance. At the same time, awareness of the necessity for a systematic approach to overviews of the medical literature has emerged [1–6]. Just as criteria for assessing the scientific quality of primary research have proved useful [7], guidelines for assessing the scientific quality of overviews are likely to aid interpretation by editors and readers of medical journals.

We therefore developed criteria for assessing the scientific quality of research overviews [8], and used them to prepare an Overview Quality Assessment Questionnaire (OQAQ). These criteria can, however, be used with confidence only if they are reliable (i.e. different individuals come to more or less the same conclusion when applying the criteria) and valid (they really do assess the scientific quality of the overview) [9]. We have previously described the development of our criteria, and demonstrated that they can

*All correspondence should be addressed to: Dr Andy Oxman, Department of Family Medicine, McMaster University Medical Centre, 1200 Main Street West, Room 2V10, Hamilton, Ontario, Canada L8N 3Z5 [*Tel.* (416) 521-9800 ext. 5435; *Fax* (416) 528–5337].

be reliably applied by clinical epidemiologists, clinicians with some methodological training, and by research assistants [8]. We then assessed the validity of the OQAQ. This paper describes the results of this validation exercise.

METHODS

Development of the instrument

Items were generated from a review of the literature. The inclusion criteria used to select items were that they should measure "scientific quality" and they should be applicable to overviews of practical questions in the health sciences. Items were excluded if they were redundant, irrelevant to scientific quality or were not generalizable to both quantitative and qualitative overviews of clinically relevant topics. The items were refined through an iterative process of discussions, pretesting and revision [8].

Conceptual approach

The purpose of the criteria (Table 1) is to assess certain aspects of the scientific quality of research overviews. The greater the scientific quality (i.e. methodologic rigor), the more likely an overview is to avoid bias, and its findings to reflect the truth. Other features which might be viewed as part of scientific quality, such as the importance of the question addressed or the quality of the presentation, are not considered by our criteria.

The goal of a research overview addressed by our measure of scientific quality is to approximate the truth regarding, for instance, the magnitude of effect of a treatment or the properties of a diagnostic test. To the extent that an article that scores high on our criteria achieves this goal, and that an overview which scores low on our criteria fails to achieve it, our instrument could be viewed as valid. Unfortunately, we never know, with certainty, what the truth is. Therefore, criterion validity, the comparison of an instrument with a gold or criterion standard, is not an option for validating criteria for the scientific quality of a research overview.

Table 1. Criteria for assessing the scientific quality of research overviews*

1.	Were the search methods reported?
2.	Was the search comprehensive?
3.	Were the inclusion criteria reported?
4.	Was selection bias avoided?
5.	Were the validity criteria reported?
6.	Was validity assessed appropriately?
7.	Were the methods used to combine studies reported?
8.	Were the findings combined appropriately?
9.	Were the conclusions supported by the reported data?
10.	What was the overall scientific quality of the overview?

*The 10 items referred to in the text are briefly summarized in this table. The complete questionnaire that was used is available from the authors.

We therefore used a number of alternative strategies. While none of them is altogether satisfactory, we believe that taken together they can provide strong evidence regarding the validity of the instrument.

Three approaches were used to assess the validity of the criteria.

(i) Feinstein [10] has suggested criteria for the "sensibility" (or face validity) of an instrument. A group of clinical epidemiologists rated our instrument according to Feinstein's criteria.

(ii) The validity of our criteria would be compromised if apparent deficiencies in an article were due to the authors' not reporting their procedures in sufficient detail (or the editors insisting that details be omitted). We therefore surveyed authors to determine the extent to which methodological shortcomings identified by our criteria reflected what was actually done.

(iii) We tested the instrument's construct validity by generating seven *a priori* hypotheses concerning how the instrument should behave if it were really measuring the scientific quality of the overviews. We assumed that the extent to which these hypotheses proved correct would reflect the validity of the criteria.

We shall now describe how overviews and judges were selected and, in some detail, how these three standards were applied.

Sensibility

Thirteen items based on Feinstein's criteria for sensibility or credibility [10] were constructed. Each item was rated on a scale of 1 to 7, where 1 signified that the instrument was not meeting its goals and 7 signified goals were fully met. The item that dealt with applicability, for instance, was worded as follows.

> To what extent is the index applicable to the wide variety of research overviews that address clinically relevant questions?
>
> 1 Small Extent
> 2
> 3 Limited Extent
> 4
> 5 Fair Extent

Table 2. Sensibility of criteria: summary of quantitative ratings

Item	Mean (SD; range)
1. Wide applicability	5.4 (1.2; 2–7)
2. Use by various groups	5.1 (1.0; 4–7)
3. Clarity and simplicity	5.3 (1.1; 3–7)
4. Adequate instructions	4.9 (1.4; 1–7)
5. Information available	3.8 (1.9; 1–7)
6. Need for subjective decisions	4.0 (1.3; 2–6)
7. Likelihood of bias	5.6 (0.8; 4–7)
9. Single domain	5.7 (1.0; 4–7)
9. Redundant items	6.2 (1.1; 3–7)
10. Comprehensiveness	5.4 (1.4; 3–7)
11. Item weights	5.8 (0.8; 5–7)
12. Number of response options	5.7 (1.7; 1–7)
13. Discriminative power	5.7 (0.8; 5–7)

6
7 Large Extent

The issues dealt with in each of the 13 items are summarized in Table 2.

Fifteen clinical epidemiologists with appointments in the Department of Clinical Epidemiology and Biostatistics at McMaster were selected at random, received a copy of the overview rating instrument, and were asked to complete the sensibility survey. The mean and standard deviation of the clinical epidemiologists' ratings were calculated.

Selection of overviews and judges

The process by which we selected overviews and judges has been described in detail elsewhere [8] and will be summarized briefly here. Articles with a practical focus on questions regarding causation, prognosis, diagnosis, therapy, prevention or policy were chosen. We attempted to identify overviews with a broad spectrum of scientific quality. Six exemplary overviews published in 1985 and 1986 were identified through personal contact with investigators with a known interest in research synthesis. Four meta-analyses, identified through a textword search of the MEDLINE database (January 1985 to March 1987) using "meta-analysis" as the search term, were identified as a second source of more rigorous overviews. Four meta-analyses were selected sequentially from this listing, beginning with the most recent posting. Two overviews were selected from journals with a methodological focus, 4 from annual reviews, 8 from major American internal medicine journals, 8 from major general medical journals, and 4 from other journals. A complete list of the overviews chosen can be obtained by writing to the authors.

The primary topics of the overviews included drug therapy (8), behavioral and other non-pharmacological therapeutic interventions (6), risk associated with medical interventions (2), prevention (5), diagnosis (6), etiology (7), and prognosis (2). The specific topic areas varied widely and included cardiology, oncology, psychiatry, preventive medicine, infectious diseases, endocrinology, surgery, nephrology, nursing care, spinal manipulation and acupuncture.

Three research assistants, three clinicians with research training, and three experts in research methodology evaluated all 36 papers. Faculty members were chosen on the basis of their interest in overview methodology; research

Table 3. Comparisons of assessments of published overviews with authors responses when questioned about their methods

Criteria*	Author's response†	*n*‡	Mean§	(SD)¶	*p* Value‖
2. Comprehensive search	Unpublished research + MEDLINE	11	4.3	(1.8)	0.014
	MEDLINE (not unpublished research)	15	3.0	(1.1)	
	Neither	4	2.1	(0.3)	
4. Selection bias avoided	Explicit criteria used	10	5.6	(1.2)	0.000
	Criteria used to some extent	10	2.8	(1.0)	
	Personal judgement used	8	2.3	(0.2)	
6. Validity assessed	Explicit criteria used	11	4.7	(1.8)	0.003
	Criteria used to some extent	9	2.8	(1.6)	
	Personal judgement used	6	1.7	(0.9)	
8. Combined appropriately	Systematically for the most part	6	6.1	(1.6)	0.000
	Somewhat systematic approach	10	3.2	(1.5)	
	Not done systematically	12	2.4	(0.9)	

*Criteria used by referees to assess overviews (see Table 1).

†Responses from survey of authors of overviews (response rate 30/36 = 83%). The response options relative to each criterion are discussed in the text.

‡Number of authors who responded accordingly. Totals do not add up to 30 for each criterion because not all of the respondents answered all of the questions.

§Mean score on criterion calculated using the means of the assessments by nine judges.

¶Standard deviation.

‖Calculated using one way analysis of variance.

assistants were randomly selected from among those employed in our Department of Clinical Epidemiology and Biostatistics. Each item in the OQAQ is associated with a 7 point scale in which 7 is the highest quality and 1 is the lowest quality. The mean rating of all nine judges was used in each analysis reported in this paper. The overall rating for each overview was calculated as the mean rating for Item 10 (Table 1).

Relation between methods and reporting

A questionnaire which addressed four of our criteria (items 2, 4, 6 and 8 from Table 3) was constructed. This questionnaire was sent to the first author of each of the 36 overviews. They were asked whether they used each of 12 specific search strategies, including MEDLINE, other bibliographic databases, manual search strategies and personal communication, whether they used explicit selection criteria and whether they used explicit criteria to assess the validity of each study included in the overview. The response options for the questions regarding specific search strategies were yes and no, and for the questions regarding the use of explicit criteria there were three response options: no (personal judgement was used), to some extent, and yes (rigorously).

While no consensus has been reached regarding the appropriateness of various methods of combining studies, including both statistical and qualitative methods, it is important to consider variation among study findings relative to five potential sources: chance and differences in study design, target populations, exposures or interventions, and outcome measures. Authors were asked to specify the extent to which these sources of variation were considered. Three response options were given for these questions: no; considered, but not evaluated systematically; and yes, systematically.

For each item, responses were used to classify overviews into three categories in terms of their rigor. The overview ratings were then compared with the mean scores on the corresponding item in the OQAQ. A one-way analysis of variance was conducted to determine if there was a difference on mean item OQAQ scores in overviews which were classified in different categories according to the authors' responses to their questionnaire.

As described below, the questionnaire also included queries as to authors' background and training which were used in the tests of construct validity.

Construct validity

Prior to completing the construction of the OQAQ we made seven predictions about how the instrument should behave if it was truly measuring the scientific quality of the overviews. These predictions were as follows:

(1) The average overall quality of overviews identified as exemplary will be higher than the average score for a sample of typical review articles drawn from a spectrum of medical journals. To test this hypothesis the mean overall scores of the six exemplary articles were compared with the scores of 26 general articles (meta-analyses were excluded from this comparison) using an unpaired t-test. It could be argued that this prediction is tautologous, to the extent that the experts who selected the exemplary reviews were implicitly using criteria similar to ours. However, it is unlikely that the criteria were identical. Indeed, experts differ on what makes an exemplary overview. The exemplary articles were identified independently by experts who were not part of our group, and with whom there had been no discussion of what constitutes an exemplary overview. Meta-analyses were excluded from the comparison because we did not feel that, if OQAQ was performing as we hoped, exemplary overviews would necessarily be rated higher than meta-analyses.

(2) The average overall score for meta-analyses will be higher than the average score for a sample of typical review articles drawn from a spectrum of medical journals. To test this hypothesis the mean overall scores of the four meta-analyses were compared with the scores of 26 general articles (exemplary overviews were excluded from this comparison) using an unpaired t-test. As with the first hypothesis, it could be argued that this prediction is tautologous, to the extent that meta-analysts implicitly use the same criteria as ours. However, it is again unlikely that the criteria were identical. The fact that there is disagreement among meta-analysts regarding what makes an exemplary overview, is reflected in the variability of the methods used in published meta-analyses and their quality [3].

(3) The average overall score for overviews with summary graphical or tabular displays of data will be higher than the average score for other review articles. To test this hypothesis the mean score of articles with and without graphical or tabular displays was compared using an unpaired t-test.

(4) The overall score will correlate positively with the amount of research training of the authors. Research training was ascertained by asking first authors directly. Authors were asked both about their total research training and about their training in overview methodology. Spearman's rank order correlations between the extent of research training and overall score on the overview was calculated for both total research training and training in overview methodology.

(5) The overall score will correlate negatively with the strength of the author's opinion regarding the topic prior to preparing the overview, and with the authors' prior expertise in the content area of the overview. Authors were asked about the strength of their opinion prior to conducting their overview, and their previous expertise. A Spearman's rank order correlation between the strength of their opinion and overall score on the overview was calculated.

(6) The average overall score for overviews written by more than one author will be higher than the average score for those written by a single authors. An unpaired t-test on overall score was conducted comparing articles with one author vs those with more than one author.

(7) The average overall score for overviews that address a single substantive question will be higher than the average score for those that address multiple questions. An unpaired t-test on overall score comparing articles addressing a single question and articles addressing more than one question was conducted.

Information regarding the authors research background, knowledge of overview methodology, strength of opinion prior to preparing the overview and level of content area expertise was obtained through the questionnaire that we sent to the first author of each of the 36 overviews. They were asked to indicate their "background in clinical investigation and/or epidemiology, taking into account both formal training and practical experience" on a 7 point scale with four anchors ranging from no research experience or training to extensive background (PhD level). They were asked to indicate if they had "read any of the literature on methodology for research overviews" on a 7 point scale, ranging from no to extensively, to assess their knowledge of overview methodology. They were asked to indicate the strength of their opinion prior to preparing the overview, with respect to the primary question that the overview addressed, using a 7 point scale ranging from decided to undecided. Their level of expertise relative to the topic of the overview, prior to writing it, was assessed using another 7 point scale ranging from little background to expert, with a detailed definition provided for each of the four anchors that were used; e.g. "expert = prior to writing the review, you had already read extensively in this area, and done research or written articles on the same topic."

RESULTS

Sensibility

The mean and standard deviation associated with each of the questions from the sensibility survey are summarized in Table 2. For most items ratings were 5 or greater, indicating general satisfaction with the instrument. There are three items for which the mean score was less than 5. The first (item 4) has to do with the adequacy of the instructions. Three people scored this item 4 or less and suggested that the instructions should be simplified or clarified.

The second problematic item (item 5) has to do with how likely it is that information needed to make the ratings will be included in a published overview. For instance, how likely is it that authors will report whether or not a comprehensive search was undertaken, or how they chose the articles to be included from among those which might have been included. Eight people scored this item 4 or less.

The third problematic item (item 6) has to do with the need for judgement in making ratings. Eight people scored this item less than 5. Reference was made to criteria 2, 4, 6 and 8 in particular. To some extent this reflects the same problem as identified above for item 5; i.e. that methods are typically not reported in published review articles. Furthermore, as noted in the respondents' comments, it reflects a concern regarding the ability of readers without training in overview methodology to make appropriate assessments.

Relation between methods and reporting

Thirty of 36 authors responded to our survey concerning their methods. Comparisons between what authors reported having done when questioned about their methods with assessments using criteria 2, 4, 6 and 8 are summarized in Table 3. For each of these four criteria the mean score for overviews for which the responding authors reported using more rigorous methods were significantly higher than for

Table 4. Tests of construct validity

	Mean	Difference	Correlation	*p* Value
1. Exemplary vs	5.5	1.9		0.006
others (not meta-analyses)	2.6			
2. Meta-analyses vs	5.0	2.4		0.004
others (not exemplary)	2.6			
3. Summary tables or graphs vs	4.2	1.6		0.004
no summary table or graph	2.7			
4. Research background			+0.06	0.747
Knowledge of overview methodology			+0.43	0.017
5. Strength of opinion			−0.45	0.019
Content area expertise			−0.52	0.004
6. More than one author vs	3.7	1.2		0.005
one author	2.4			
7. Single question vs	4.1	1.8		0.000
multiple questions	2.2			

those where the respondents reported using less rigorous methods. On average, the judges assessments using the criteria were consistent with what the authors stated they did in response to direct inquiries about the methods they used.

Construct validity

The results of the tests of construct validity are summarized in Table 4. The hypotheses were, in general, confirmed. Differences between categories (exemplary vs others, meta-analyses vs others, etc.) were substantial and very unlikely to have occurred by chance. The only hypothesis that was not confirmed was the anticipated correlation between research training and quality of the overview. Other anticipated correlations were in the range of 0.43–0.52 and very unlikely to have occurred by chance.

DISCUSSION

As stated above, criteria for judging the scientific quality of an overview are valid insofar as they give a high rating to an overview whose result approximates the truth, and a low rating to an overview whose result deviates substantially from the truth. Theoretically, empirical evidence could be gathered which would support the validity of individual criteria in this fashion. For example, that an overview should systematically consider the validity of a study is supported by a number of observations. These include the fact that sub-experimental study designs consistently yield larger effect sizes than do randomized trials [11, 12] and that, within randomized trials, trials in which randomization may not have been blinded yield systematically larger effects [13]. It would be desirable if there was empirical evidence that, for instance, overviews in which inclusion criteria were precisely and specifically stated yielded results that were consistently closer to the truth than overviews in which this was not the case. Opportunities for comparable validation of the OQAQ criteria are limited, although in some cases it is possible to compare meta-analyses of small studies with large co-operative trials [14] or to compare overviews that agree and disagree [15].

We have used three surrogates to validate the OQAQ: the consistency of OQAQ ratings with author's reports of their own work; the "sensibility" of the questionnaire; and the extent to which seven *a priori* hypotheses about OQAQ's performance held true. In general, each of these strategies provided strong support for the validity of the OQAQ. In the remainder of the discussion, we will deal with instances in which the criteria were not fully supported by the validation exercise.

Respondents to the sensibility survey felt that the criteria demanded an excessive degree of subjective judgement. We believe the respondents are accurate in identifying this limitation of the criteria. Although we strove to minimize the need for judgement, some judgement is inevitable. We believe the need for judgement contributed to the poorer reliability among research assistants observed for items 7, 8 and 9 (although it did not appear to be a problem on items 4 and 6) as reported in our prior investigation of the reliability of the OQAQ [8].

It is not surprising that our criteria did somewhat poorly with respect to the fifth criterion in the sensibility questionnaire regarding the likelihood that information needed to make the ratings will be included in a published overview.

Most published overviews currently do not report methods. For this reason we included four criteria (criteria 1, 3, 5 and 7 in Table 1) that simply address whether the methods were reported. In the instructions we have suggested what we feel are reasonable guidelines to use when methods are not explicitly reported for scoring the corresponding criteria regarding the strength of the methods (criteria 2, 4, 6 and 8 in Table 1). This is a somewhat unsatisfactory solution. However, the guidelines are based on the assumption that if what was done is not reported, there is a good chance that it was not done rigorously. This contention is supported by the extent to which authors' reports of what they did were related to scores on OQAQ criteria (Table 3).

That these limitations in sensibility do not markedly impair the performance of the OQAQ is suggested by both the excellent reliability of the instrument (in research assistants without training in overview methodology as well as in clinical epidemiologists) [8] and the extent to which data regarding the relation of OQAQ criteria to external measures was consistent with our *a priori* hypotheses. For the one *a priori* hypothesis that was not supported, it is worth noting that although the scientific quality of the overviews was not positively correlated with the amount of research training of the authors, scientific quality was positively correlated with knowledge of overview methodology ($r = 0.43$, $p = 0.017$).

At the time when our survey was undertaken (1987–1988), relatively few meta-analyses had been published [3]. Subsequently, there has been an increase in the number of published meta-analyses and increased recognition of the application of scientific principles to overviews [2]. Nonetheless, a small proportion of review articles that readers are likely to encounter are scientifically rigorous [2, 3, 8, 16].

At the same time, only a small proportion of primary research reports found in the medical literature are scientifically rigorous [17]. Careful evaluation of the scientific quality of any report is required before applying its conclusions to clinical practice. Although criteria for assessing the scientific quality of primary research reports have been put forward largely on the basis of logical arguments [7, 18–22], there is some empirical evidence regarding their validity and reliability [11–13, 23–27]. Systematic overviews provide an important means of further testing the measurement properties of criteria for evaluating primary research articles [28], as well as relieving readers of some of the burden of critically appraising the primary literature on their own.

We believe the OQAQ is a valid instrument to measure the quality of research overviews. It can be productively used by readers and editors of clinical journals to identify scientifically sound overviews and thus to judge the confidence that should be placed in their conclusions.

Acknowledgements—This work was supported by the Ontario Ministry of Health, Grant No. 01969. Dr Guyatt is a career scientist of the Ontario Ministry of Health.

REFERENCES

1. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. **Can Med Assoc J** 1988; 138: 697–703.
2. Mulrow CD. The medical review article: state of the science. **Ann Intern Med** 1987; 106: 485–488.
3. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers C. Meta-analyses of randomized controlled trials. N **Engl J Med** 1987; 316: 450–455.
4. L'Abbe KA, Detsky AS, Orourke K. Meta-analysis in clinical research. **Ann Intern Med** 1987; 107: 224–233.
5. Gerbarg ZB, Horwitz RI. Resolving conflicting clinical trials: guidelines for meta-analysis. **J Clin Epidemiol** 1988; 41: 503–509.
6. Squires BP. Biomedical review articles: what editors want from authors and peer reviewers. **Can Med Assoc J** 1989; 141: 195–197.
7. Sackett DL, Haynes RB, Tugwell P. **Clinical Epidemiology, A Basic Science for Clinicians.** Boston, Mass.: Little, Brown & Co.; 1985.
8. Oxman AD, Guyatt GH, Singer J *et al.* Agreement among reviewers of review articles. **J Clin Epidemiol** 1991; 44: 91–98.
9. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. **Can Med Assoc J** 1986; 314: 889–892.
10. Feinstein AR. **Clinimetrics.** New Haven, Conn.: Yale University Press; 1987: 141–166.
11. Sacks HS, Chalmers TC, Smith H Jr. Randomized versus historical assignment in controlled clinical trials. N **Engl J Med** 1983; 309: 1353–1361.
12. Colditz GA, Miller JN, Mosteller F. The effect of study design on gain in evaluations of new treatments in medicine and surgery. **Drug Inf J** 1988; 22: 343–352.
13. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. N **Engl J Med** 1983; 309: 1358–1361.
14. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: control of bias and comparison with large co-operative trials. **Stat Med** 1987; 6: 315–325.
15. Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. II: replicate variability and comparison of studies that agree and disagree. **Stat Med** 1987; 6: 733–744.
16. Dickersin K, Higgins K, Meinert CL. Identification of meta-analyses: the need for standard terminology. **Contr Clin Trials** 1990; 11: 52–66.
17. Williamson JW, Goldschmidt PG, Colton T. The quality of medical literature: an analysis of validation assessments. In: Bailar JC, Mosteller F, Eds. **Medical Uses of Statistics.** Waltham: NEJM Books; 1986: 370–391.

18. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. **Chest** 1986; 89: 2s–3s.
19. Chalmers TC, Smith H, Blackburn B *et al.* A method for assessing the quality of a randomized control trial. **Contr Clin Trials** 1981; 2: 31–49.
20. Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiological research. **N Engl J Med** 1982; 307: 1611–1617.
21. Horwitz RI, Feinstein AR. Methodological standards and contradictory results in case–control research. **Am J Med** 1979; 66: 556–564.
22. Begg CB. Biases in the assessment of diagnostic tests. **Stat Med** 1987; 6: 411–423.
23. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. **N Engl J Med** 1983; 309: 1358–1361.
24. Miller JN, Colditz GA, Mosteller F. How study design affects outcomes in comparisons of therapy. II: surgical. **Stat Med** 1989; 8: 455–466.
25. Bhaskar R, Reitman D, Sacks HS, Smith HS, Chalmers TC. Loss of patients in clinical trials that measure long-term survival following myocardial infarction. **Contr Clin Trials** 1986; 7: 134–148.
26. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. **Contr Clin Trials** 1989; 10: 31–56.
27. Diehl LF, Perry DJ. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? **J Clin Oncol** 1986; 4: 1114–1120.
28. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical-study of the possible relation of treatment differences to quality scores in controlled randomized clinical-trials. **Contr Clin Trials** 1990; 11: 339–352.