A guide to interpreting discordant systematic reviews

Alejandro R. Jadad, MD, DPhil; Deborah J. Cook, MD, MSc; George P. Browman, MD, MSc

Abstract

SYSTEMATIC REVIEWS ARE BECOMING prominent tools to guide health care decisions. As the number of published systematic reviews increases, it is common to find more than 1 systematic review addressing the same or a very similar therapeutic question. Despite the promise for systematic reviews to resolve conflicting results of primary studies, conflicts among reviews are now emerging. Such conflicts produce difficulties for decision-makers (including clinicians, policy-makers, researchers and patients) who rely on these reviews to help them make choices among alternative interventions when experts and the results of trials disagree. The authors provide an adjunct decision tool — a decision algorithm — to help decision-makers select from among discordant reviews.

Résumé

LES EXAMENS CRITIQUES SYSTÉMATIQUES DEVIENNENT des outils importants pour guider les décisions relatives aux soins de santé. Comme le nombre des examens critiques systématiques publiés augmente, on en trouve de plus en plus souvent plusieurs qui portent sur une question thérapeutique identique ou très semblable. Même si les examens critiques systématiques promettent de trancher des résultats contradictoires d'études primaires, on commence à constater des divergences entre les examens. Ces divergences causent des difficultés aux décideurs (y compris les cliniciens, les chercheurs et les patients) qui comptent sur ces études pour faire des choix entre des interventions possibles lorsqu'il y a désaccord entre des experts et des résultats d'études. Les auteurs fournissent un outil d'aide à la décision — un algorithme de décision — afin d'aider les décideurs à faire un choix parmi des études divergentes.

Systematic reviews appraise critically, summarize and attempt to reconcile the published evidence concerning a particular problem. They have gained prominence as useful tools for evidence-based decision-making.[1–3] Systematic reviews can be quantitative or qualitative. In a quantitative systematic review, or meta-analysis, the results of 2 or more primary studies evaluating a health care intervention are statistically combined to produce an overall estimate of the treatment effect. When the results of primary studies are summarized but not statistically combined, the article can be considered a qualitative systematic review. Although they are important, such reviews are not the subject of this article. In the following, the terms "meta-analysis," "systematic review" and "review" are used interchangeably.

The number of meta-analyses published every year has increased at least 500-fold during the past 10 years.[4–6] This increase has led to a rise in the number of reviews addressing the same or a very similar therapeutic question, with a concomitant increase in the conflicts among reviews.[7–9] These conflicts may be trivial or may trigger passionate academic disputes spanning decades, such as the current debate over the association between corticosteroid use and peptic ulcers.[10–14]

Such discordance causes difficulties for decision-makers (including clinicians, policy-makers, researchers and patients, depending on the context) who rely on these reviews to help them make choices among alternative health care interventions when experts and the results of trials disagree. Furthermore, discordant rig-

Education

Éducation

Dr. Jadad is with the Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, and Drs. Cook and Browman are with the Departments of Clinical Epidemiology and Biostatistics and of Medicine, McMaster University, Hamilton, Ont.

This article has been peer reviewed.

Can Med Assoc J 1997;156:1411-6

orous reviews can confuse rather than clarify. They thus challenge the evidence-based approach to clinical and health care policy decision-making. In this article, we provide an adjunct decision tool to help decision-makers interpret and choose among discordant systematic reviews.

Interpreting discordance in quantitative systematic reviews

Reviews can disagree in 2 ways: their results can diverge, or the interpretations and inferences made by the review authors can be discordant. In this article, we will concentrate on reviews with discordant results. Given that the clinical, methodologic and administrative aspects of designing and conducting systematic reviews are complex and evolving, it is unrealistic to expect that a single factor invariably explains discordance among reviews.[9,15] Table 1 summarizes the potential sources of discordance among meta-analyses.

In this article, we identify the issues to be considered in determining which of 2 or more discordant reviews is most appropriate as a basis for decision-making in a given situation. We differentiate between ideal methods for resolving discordance, which may not always be feasible, and more feasible approaches. A decision-making algorithm, which summarizes the process for identifying and resolving causes of discordance, is given in Fig. 1.

Are the reviews valid?

Three instruments for assessing the quality of systematic reviews have been published. Two checklists[16–18] have been proposed to monitor the methodologic rigour of published meta-analyses, but these checklists have not been formally validated. The other instrument[19] has been extensively validated and produces reliable scores even among individuals who have not been trained to use it.[7] Given its simplicity and validation, we recommend the latter method to assess the methodologic quality of reviews. The full instrument is published elsewhere,[7] and a modified version is included in an article in the series entitled User's guides to the medical literature.[20]

Regardless of the method used to assess its quality, if a review is judged as having few flaws, and only minor ones, it may be appropriate for use in decision-making; if the flaws are critical or numerous, the review may be unsuitable in guiding health care decisions.

Are the differences among the discordant reviews important?

Reviews of the same topic are likely to differ in some respects. Table 2 shows the types of discordance that may be observed. Discordant meta-analyses of alternative interventions may differ with respect to the direction of the estimated effect or, if the direction is the same, with respect to the effect's magnitude or statistical significance. The importance of such differences depends on whether they lead to different health care decisions, which are likely to produce different outcomes. A decision-maker may consider differences between 2 reviews to be unimportant if the estimated treatment effects are of different magnitude but in the same direction, and are statistically significant and clinically important. For example, despite differences in the population of patients, the number of studies included, the statistical methods used and the estimates of benefit, several reviews have shown that administration of low doses of heparin produces a clinically important and statistically significant reduction in the risk of deep venous thrombosis after surgery.[21–23] In such cases, the treatment effect may be so robust that it is shown consistently despite differences in the methodologic approach of each review.

Conversely, differences among reviews are regarded as important if they lead to different health care decisions, with implications for patient outcomes, costs of treatment or both. This difference is particularly relevant if one review suggests that an intervention produces statistically significant or clinically important benefits, whereas another points to lack of benefit or to harm. A decision-maker responsible for allocating health care resources, for instance, would find it difficult to decide whether to fund a program to use laser for the treatment of pain in musculoskeletal conditions. One review suggests that laser treatment is effective,[24] whereas another concludes that it is ineffective.[25]

Table 1: Sources of discordance among meta-analyses

Clinical question
Populations of patients
Interventions
Outcome measures
Settings
Study selection and inclusion
Selection criteria
Application of the selection criteria
Strategies to search the literature
Data extraction
Methods to measure outcomes
End points
Human error (random or systematic)
Assessment of study quality
Methods to assess quality
Interpretations of quality assessments
Methods to incorporate quality assessments in review
Assessment of the ability to combine studies
Statistical methods
Clinical criteria to judge the ability to combine studies
Statistical methods for data synthesis

Such discordance could also confuse patients and clinicians who are trying to decide whether to use the treatment. In such circumstances, strategies to understand why reviews disagree are particularly needed to help choose the most appropriate review to use as a basis for decision-making.

If discordance among the results of reviews is judged to be important (i.e., if the reviews lead to different decisions), the next step is to determine whether the reviews address the same question.

Do the reviews ask the same question?

The decision-maker should determine which review addresses the question most relevant to his or her situation, by comparing the populations, interventions, outcome measures and settings examined in each review (Step A in Fig. 1). If the research questions are not identical, the most appropriate review, in principle, is the one with the question closest to the problem that the decision-maker is trying to solve (Step B), and the decision-maker need not go further. If the reviews address the same question, it is important to establish whether they include the same primary trials.

Do the reviews include the same trials? (Step C)

Differences among reviews may be due to different search strategies or different criteria for selecting studies for inclusion. In general, reviews with more

Table 2: Types of discordance

Type	Example
Results	
Direction of effect	One review favours the experimental treatment and another favours the control treatment
Magnitude of effect	One review suggests that the intervention results in a 30% reduction in mortality and another suggests that it results in a 5% reduction in mortality
Statistical significance	One review shows a statistically significant difference between the experimental and the control treatments and another review shows a nonsignificant difference between them
Interpretation of the results	

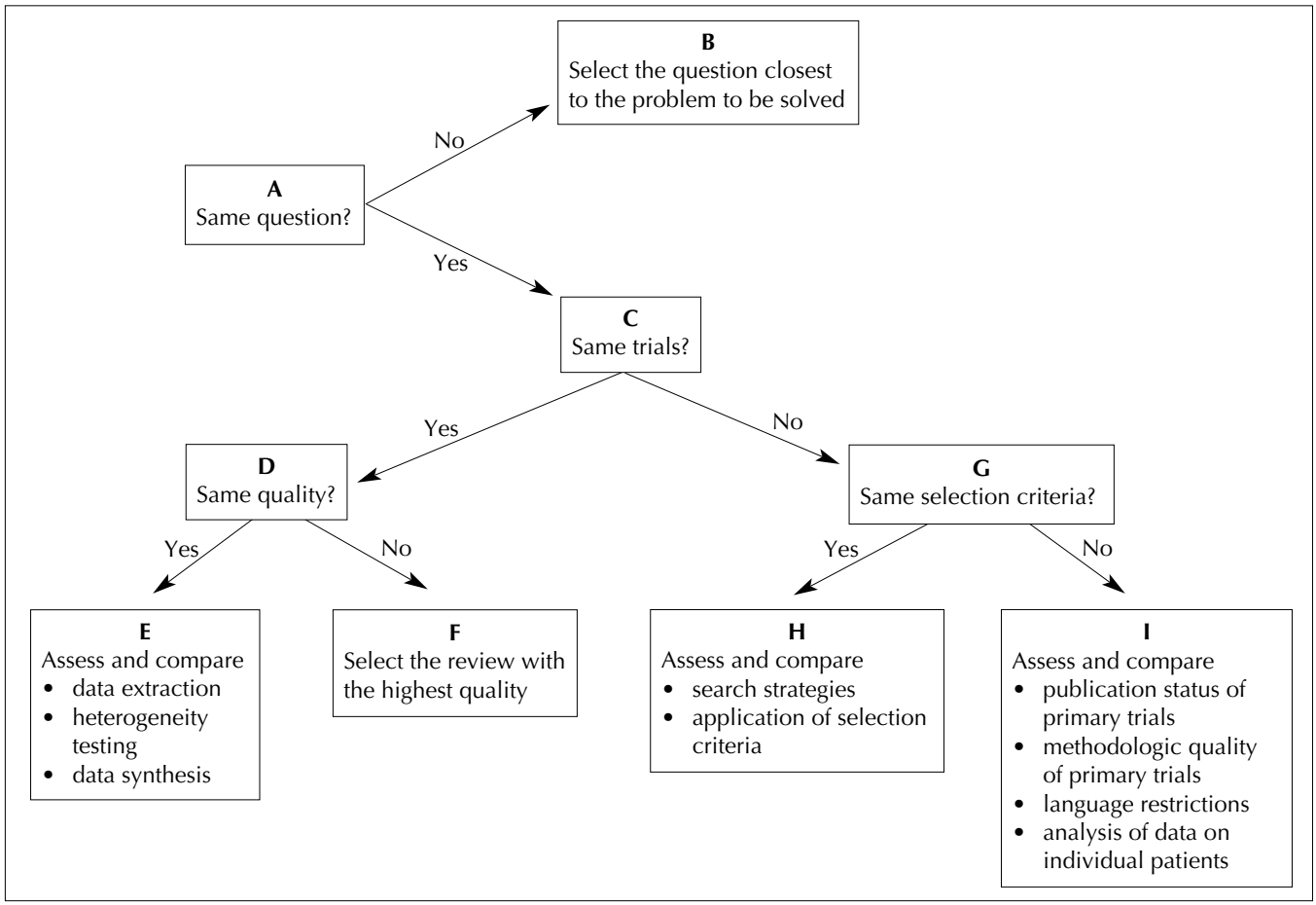

Fig. 1: A decision algorithm for interpreting discordant reviews (assuming that the reviews have few and minimal flaws).

comprehensive search strategies are less likely to be biased.[7]

We describe first the process for assessing discordance among reviews that include the same trials. A discussion of reviews that include different trials follows.

Discordant reviews containing the same trials

In the case of reviews of the same trials, the next step is to compare the scientific quality of the reviews (Step D). The review judged to be most rigorous because it minimizes opportunities for bias should be used to inform the decision (Step F).

Choosing among reviews of the same trials with the same methodologic quality (Step E)

However, the quality of discordant reviews may be similar,[7] necessitating an examination of differences in data extraction, in the ability to combine the studies (heterogeneity testing) and in data-synthesis methods.

Data extraction: Data extracted from the primary studies may differ (i.e., quality of life v. survival after myocardial infarction, or survival 3 years after diagnosis of ovarian cancer v. survival 5 years after diagnosis). If reviews differ on this basis, the decision-maker should identify the review that takes into account the outcome measures most relevant to the problem that he or she is trying to solve.

There may also be differences in the data extracted as a result of human error, biased extraction or misprints. Reviews that employ independent, replicated procedures for abstraction by observers unaware of potentially biasing information (e.g., trial results, journal, publication date and authors) may be less likely to be biased than other reviews and may therefore provide the most valid information for decision-making.[26,27] Empirical studies are needed to confirm this suggestion.

Ability to combine trial results (heterogeneity testing): Reviews of similar quality may differ in their assessment of whether the primary studies can be combined. Judgements based on clinical and biologic understanding of disease processes and the mechanisms of action of interventions can be used to determine whether it makes sense to pool the results of particular studies with those of other studies.[28]

The ability to combine studies statistically must also be considered; that is, one should consider whether the distributional assumptions of the data, on which the statistical methods for combining the results are based, have been violated in the review.[28]

Reviews that address whether results can be combined and that make efforts to test the underlying assumptions in choosing the statistical method for pooling data are probably more credible than reviews that ignore such issues.[29]

Methods of data synthesis: Differences in the methods used to synthesize the results of primary studies may be substantial, such as the difference between a qualitative and a quantitative review, or subtle (and yet important), such as the differences in the statistical methods used.

There are many analytic methods available for meta-analysis. These approaches differ in their assumptions, complexity, indications and, sometimes, results. A description of meta-analytic statistical methods is beyond the scope of this article, but the most important issues related to their indications, strengths and limitations have been discussed recently.[30]

Choosing among discordant reviews that include different trials (Step G)

For discordant reviews that contain different studies, the decision-maker's next task is to determine which review incorporates the trials most relevant to the question under consideration and which selection process is least likely to be biased. This determination requires an examination of the selection criteria used.

Reviews with the same selection criteria (Step H)

Differences in the studies included in reviews may reflect differences in the strategies used to identify them or in the application of selection criteria to individual studies.

Search strategy: Ideally, reviewers search all available sources of information to identify all relevant studies addressing a particular question.[31] In practice, because of time and cost constraints, reviewers strive to identify as many eligible studies as possible.[32] Differences in search strategies may result in inclusion of different studies in reviews, which in turn may lead to different results. If inclusion of different sets of studies with systematically different populations results in different conclusions, this may be due to bias. The inclusion of more studies may not lead to systematically different conclusions, but to increased precision (reduced random error). In a recent qualitative review of 80 meta-analyses of studies of analgesic interventions, the meta-analyses with more comprehensive (and presumably less biased) literature searches were less likely than other such meta-analyses to generate positive results (a statistically significant favourable effect).[7] The main problem for decision-makers, however, is the dearth of evidence to guide them in establishing how extensive a literature search must be (on average or in specific situations) to avoid bias or important imprecision in a systematic review.

Application of selection criteria: Reviews with the same selection criteria may include different trials because of differences in the application of the criteria, which are due to

random or systematic error. Decision-makers should regard as more rigorous those reviews with explicit, reproducible inclusion criteria. Such criteria are likely to reduce bias in the selection of studies.

Reviews with different selection criteria (Step I)

Reviews may use different criteria to select studies; specifically, criteria that may differ include publication status, methodologic quality, language of publication and availability of data on individual patients.

Publication status: A common cause of differences in the number of studies included in reviews is whether unpublished studies are included. If the results of unpublished studies are similar to those of published studies, excluding unpublished studies affects only the precision of the estimates. Reviews can be biased by exclusion of unpublished studies if these generally have nonsignificant results (this has been called "publication bias").[33] Despite concerns and some empirical evidence that exclusion of unpublished trials may bias meta-analyses in favour of experimental treatments, a recent survey showed that fewer than half of prominent medical journal editors believe that unpublished data should be included in meta-analyses.[34]

If 1 or more of the discordant reviews include unpublished studies, the discordance may well be due to the unpublished data; however, the discordance may be explained by differences between the published studies included in both reviews. More research is needed to understand the contributions of unpublished trials to the validity of meta-analyses. Meanwhile, decision-makers need to use judgement in determining how much weight to place on this factor when choosing among discordant reviews.

Methodologic quality of primary trials: Assessment of the methodologic quality of the trials included is an essential component of systematic reviews.[16,19] A difference in the quality of the primary studies is one of the factors that may explain differences among reviews. However, there is a lack of agreement on how the quality of primary studies should be assessed and how the assessments should be incorporated in reviews.[35]

Rather than assessing the quality of the individual studies included (a time-consuming enterprise), we recommend that decision-makers determine how the quality of the primary studies was assessed, by judging the appropriateness of the methods used to assess quality and to incorporate the quality assessments in the review. Reviews that address these issues are likely to be more rigorous than those that ignore trial quality.

Language restrictions: The systematic exclusion of studies not published in English may be explained by difficulties with translation or by the perception that studies published in other languages are of less relevance or of inferior methodologic quality. However, evidence is emerging to challenge this view. In a recent study, no differences were found in the quality and other design features of 133 randomized controlled trials published in English and 96 published in French, Italian, German or Spanish during the same period in the same type of journal.[36] Therefore, language-restricted searches are likely to reduce the precision of the results of reviews and may introduce systematic error.

Analyses of data on individual patients: Meta-analyses of data on individual patients are considered the yardstick against which other reviews of randomized controlled trials should be measured.[37] Data on individual patients afford the opportunity for reviewers to measure the same outcomes more uniformly, to compare outcomes measured at different times, to use intention-to-treat analysis, to conduct flexible subgroup analyses based on *a priori* hypotheses, and to update follow-up information.[38,39] Balanced against these advantages are concerns about the quality of data stored years earlier and about the time-consuming and costly process of collecting, verifying, "cleaning" and reformatting data on individual patients to create a new database.[40]

Meta-analyses based on data on individual patients are unavailable for the evaluation of most health care interventions, and will continue to be so.[39] However, if a decision-maker must choose among several high-quality reviews, an available meta-analysis based on data on individual patients is preferred.

In summary, decision-makers trying to choose among discordant reviews with similar questions answered by different studies should choose the review that includes the most comprehensive literature search, explicit and reproducible selection of studies, quality assessment of the primary studies and incorporation of data on individual patients, when possible. However, more empirical evidence is required to establish the effect of these elements on the validity of the review process, their relative importance and their effect on the results of a review.

Joint research efforts to solve discordance

An alternative approach for researchers to explore discordance among systematic reviews is to collaborate to ascertain why reviews yielded different results and to generate a new review that reconciles the discrepancies.[8] Such reviews could obviate the need for clinicians or other decision-makers to select just 1 review to use in practice. It is unlikely, however, that authors of every set of discordant reviews will cooperate to solve their disagreements.

Conclusions and future directions

The pace of medical research, our increasing need for valid, relevant health care information and our limited resources to find, appraise and apply this information underscore the need for rigorous reviews to guide health care decisions.

The exponential growth in systematic reviews has led to an increase in the number that address similar therapeutic problems and that yield discordant results. The very tools that have been promoted as arbiters have, at times, confused rather than clarified the situation.

The need to interpret discordant reviews motivated us to develop a simple guide to help decision-makers select the most relevant and valid of the conflicting reviews. Our decision algorithm is based, whenever possible, on empirical evidence.

A future research agenda should include these important goals: the development of efficient and comprehensive databases of reviews, highlighting discordant cases; the encouragement of reviewer collaboration to solve discordance; and the generation of the empirical evidence required to answer fundamental methodologic questions. In the meantime, the publication of discordant reviews is likely to continue. We hope that the guide we propose will help decision-makers to understand discordance among reviews, guide their decisions and serve as a platform for discussion and research. For evidence-based health care to achieve its initial promise, the input of decision-makers is needed to identify and resolve these new challenges.

We thank Susan Marks for reviewing the manuscript and for helping us shorten it without losing its meaning. This work was supported in part by the Ontario Cancer Treatment Practice Guideline Initiative of the Ontario Ministry of Health.

References

1. Chalmers I, Altman DG. Foreword. In: Chalmers I, Altman DG, editors. *Systematic reviews*. London: BMJ Publishing Group; 1995.
2. Kassirer JP. Clinical trials and meta-analysis: What do they do for us? *N Engl J Med* 1992;327:273-4.
3. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *JAMA* 1992;268:240-8.
4. Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. *Stat Methods Med Res* 1993;2:161-72.
5. Chalmers I, Haynes RB. Reporting, updating, and correcting systematic reviews of the effects of health care. In: Chalmers I, Altman DG, editors. *Systematic reviews*. London: BMJ Publishing Group;1995:86-95.
6. Moher D, Olkin I. Meta-analysis of randomized controlled trials: a concern for standards. *JAMA* 1995;274:1962-4.
7. Jadad AR, McQuay HJ. Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. *J Clin Epidemiol* 1996; 49:235-43.
8. Cook DJ, Reeve BK, Guyatt GH, Griffith LE, Heyland DK, Tryba M. Stress ulcer prophylaxis in the critically ill: resolving discordant meta-analyses. *JAMA* 1996;275:308-14.
9. 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;6:733-44.
10. Conn HO, Blitzer BL. Nonassociation of adrenocorticosteroid therapy and peptic ulcer. *N Engl J Med* 1976;294:473-9.
11. Messer J, Reitman D, Sacks HS, Smith H, Chalmers TC. Association of adrenocorticosteroid therapy and peptic-ulcer disease. *N Engl J Med* 1983;309:21-4.
12. Conn HO, Poynard T. Adrenocorticosteroid administration and peptic ulcer: a critical analysis. *J Chronic Dis* 1985;6:457-68.
13. Chalmers TC. Meta-analysis in clinical medicine. *Trans Am Clin Climatol Assoc* 1987;99:144-50.
14. Conn HO, Poynard P. Adrenocorticosteroids and peptic ulcer: meta-analysis of adverse events during steroid therapy. *J Intern Med* 1994;236:619-32.
15. Hasselblad V, Mosteller F, Littenberg B, Chalmers TC, Hunink MG, Turner JA, et al. A survey of current problems in meta-analysis. Discussion from the agency for health care policy and research interPORT work group on literature review/meta-analysis. *Med Care* 1995;33:202-20.
16. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. *N Engl J Med* 1987;316:450-5.
17. Sacks HS, Berrier J, Reitman D, Pagano D, Chalmers TC. Meta-analysis of randomized controlled trials: an update of the quality and methodology. In: Bailar JC III, Mosteller F, editors. *Medical uses of statistics*. 2nd ed. Boston: NEJM Books; 1992:427-42.
18. Thacker SB, Peterson HB, Stroup DF. Meta-analysis for the obstetrician-gynecologist. *Am J Obstet Gynecol* 1996;174:1403-7.
19. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991;44:1271-8.
20. Oxman AD, Cook DJ, Guyatt GH for the Evidence-Based Medicine Working Group. User's guides to the medical literature. VI. How to use an overview. *JAMA* 1994;272;1367-71.
21. Colditz GA, Tuden RL, Oster G. Rates of venous thrombosis after general surgery: combined results of randomized clinical trials. *Lancet* 1986;2:143-6.
22. Clagett GP, Reisch JS. Prevention of venous thromboembolism in general surgical patients. Results of meta-analysis. *Ann Surg* 1988;208:227-40.
23. Imperiale TF, Speroff T. A meta-analysis of methods to prevent venous thromboembolism following total hip replacement. *JAMA* 1994;271:1780-5.
24. 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;72:483-91.
25. Gam AN, Thorsen H, Lonnberg F. The effect of low-level laser therapy on musculoskeletal pain: a meta-analysis. *Pain* 1993;52;63-6.
26. Felson DT. Bias in meta-analytic research. *J Clin Epidemiol* 1992;45:885-92.
27. Chalmers TC. Problems induced by meta-analysis. *Stat Med* 1991;10:971-80.
28. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. In: Chalmers I, Altman DG. *Systematic reviews*. London: BMJ Publishing Group; 1995:48-63.
29. Thompson SG, Pocock SJ. Can meta-analysis be trusted? *Lancet* 1991;338: 1127-30.
30. Hasselblad V, McCrory D. Meta-analytic methods for medical decision making: a practical guide. *Med Decis Making* 1995;15:81-96.
31. Peto R. Why do we need systematic overviews of randomized trials? *Stat Med* 1987;6:233-40.
32. Jadad AR, McQuay HJ. Searching the literature: be systematic in your searching [letter]. *BMJ* 1993;307:66.
33. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990;263;1385-9.
34. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, et al. Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA* 1993;269:2749-53.
35. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clin Trials* 1995;16:62-73.
36. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, LeLorier J, et al. Assessing the completeness of reporting of trials published in languages other than English: implications for the conduct of systematic reviews. *Lancet* 1996;347:363-6.
37. Chalmers I. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Ann N Y Acad Sci* 1993;703:156-63.
38. Clarke MJ, Stewart LA. Obtaining data from randomized controlled trials: How much do we need for reliable and informative meta-analyses? In: Chalmers I, Altman DG, editors. *Systematic reviews*. London: BMJ Publishing Group; 1995:37-47.
39. Oxman AD, Clarke MJ, Stewart LA. From science to practice: meta-analyses using individual patient data are needed. *JAMA* 1995;274:845-6.
40. Easterbrook P, Berlin J. Meta-analysis [letter]. *Lancet* 1993;341:965.

Reprint requests to: *Dr. Alejandro R. Jadad, Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton ON L8N 3Z5; fax 905 546-0401; Jadada@fhs.mcmaster.ca*