Trials and Tribulations of Systematic Reviews and Meta-Analyses

Mark A. Crowther and Deborah J. Cook

St. Joseph's Hospital, Hamilton, Ontario, Canada; McMaster University, Hamilton, Ontario, Canada

Systematic reviews can help practitioners keep abreast of the medical literature by summarizing large bodies of evidence and helping to explain differences among studies on the same question. A systematic review involves the application of scientific strategies, in ways that limit bias, to the assembly, critical appraisal, and synthesis of all relevant studies that address a specific clinical question. A meta-analysis is a type of systematic review that uses statistical methods to combine and summarize the results of several primary studies. Because the review process itself (like any other type of research) is subject to bias, a useful review requires rigorous methods that are clearly reported. Used increasingly to inform medical decision making, plan future research agendas, and establish clinical policy, systematic reviews may strengthen the link between best research evidence and optimal health care. In this article, we discuss key steps in how to critically appraise and how to conduct a systematic review or meta-analysis.

Introduction and Background

Timely, useful evidence from the biomedical literature should be an integral component of clinical decision making. If one treatment has been shown to be better than another, we need to know, so that we can recommend the treatment to the appropriate patients. The worldwide effort to develop new tests and treatments, and to determine their usefulness, has never been stronger, and our patients and their families expect us to be fonts of the knowledge that results from this effort. Unfortunately, it is easy for current best research evidence to pass us by. We may lack the time, motivation, and basic skills needed to find, critically appraise, and synthesize information, all of which we must do if we are to integrate the results of original studies into our practice. We often need a concise, current, rigorous synthesis of the best available evidence on each of these topics: in brief, a systematic review. Fortunately, several methods are emerging that can greatly enhance our ability to interpret and apply research evidence; foremost among them is the systematic review.

Types of Review Articles

The term "systematic review" describes a specific type of review article with a methods section. A systematic review involves the searching, selecting, appraising, interpreting, and summarizing of data from original studies.[1] These original studies may be observational studies or randomized trials. The study summaries may be qualitative or quantitative. If a systematic review involves a quantitative summary of results, it is called a "meta-analysis." The term "overview" is sometimes used to denote a systematic review, whether quantitative or qualitative.

Summaries of research that lack explicit descriptions of systematic methods are often called narrative reviews. Most narrative review articles deal with a broad range of issues related to a given topic rather than addressing a particular issue in depth. For example, a narrative review on anemia (such as that which might be found in a textbook chapter) might include sections on its physiology and pathophysiology; the epidemiology of and prognosis-associated anemia of different etiologies; diagnostic and screening approaches; and preventive, therapeutic, rehabilitative, and palliative interventions. Thus, narrative reviews are less often useful in furnishing quantitative answers to specific clinical questions.

Systematic review articles are one type of "integrative publication"; practice guidelines, economic evaluations, and clinical decision analyses are others. These other types of integrative articles often incorporate the results of systematic reviews. For example, practice guidelines are systematically developed statements intended to assist practitioners and patients with decisions about appropriate health care for specific clinical circumstances. Evidence-based practice guidelines are based on systematic reviews of the literature, appropriately adapted to local circumstances and values. Economic evaluations compare both the costs and the consequences of different courses of action; the knowledge of consequences that are considered in these evaluations is often generated by systematic reviews of primary studies. Decision analyses quantify both the likelihood and the valuation of the expected outcomes associated with competing alternatives.

In the past 20 years, the publication of systematic reviews and meta-analyses has grown exponentially.[2] The methodology of meta-analyses has evolved from a simple summary of data from published studies to a specific set of statistical methods that acknowledge and reduce the impact of bias and examine heterogeneity of individual study results. Systematic reviews are generated to answer specific, often narrow, clinical questions in depth. These ques-

tions can be formulated explicitly according to four variables: a specific population and setting (such as elderly outpatients), the condition of interest (for example, hypertension), an exposure to a test or treatment (such as pharmacologic management), and one or more specific outcomes (such as cardiovascular and cerebrovascular events and mortality). A narrative review may not include a systematic literature review, rarely delineates specific reasons for inclusion or exclusion of specific articles from the review, and does not normally present the mathematic models used to combine results. Narrative reviews rarely include explicit information on the process used to identify, extract, and combine the information presented; as a result, qualitative reviews are much more prone to systematic error and bias. Unfortunately, many review articles that purport to be systematic are narrative, and determining the degree of rigor in a systematic review requires special skills.

As the volume of medical literature has increased, making it difficult or impossible for clinicians, educators, and investigators to maintain current knowledge even in highly specialized area of practice, systematic reviews and meta-analyses have proliferated. Such articles play an important role in measuring and reporting inadequacies in the underlying data, thus reducing the likelihood of systematic error. By collating data from the literature, systematic reviews combine results in a way that reduces the likelihood that chance observations unjustifiably affect clinical practice. In addition, if performed correctly, systematic reviews should more precisely estimate the impact of interventions in terms of their benefits and harms. In selected cases, summarizing the results of several studies can identify outcomes or toxicities that are not perceptible in any of the constituent studies.

Applying Systematic Review Methods to Studies Other than Randomized Trials

It is important to acknowledge that types of evidence other than randomized controlled trials of drug therapy can be systematically reviewed. Thus, the review methods for such primary articles are the same, although methods of pooling results will differ. Results from cohort studies and other nonrandomized designs may also be mathematically combined. However, for questions of prevention and treatment, the greatest degree of confidence in the results of a systematic review will be obtained when a series of randomized controlled trials are combined that exposed patients to similar interventions and that measured similar outcome variables. Inclusion of data from sources other than randomized trials reduces the reliability of the conclusions of a systematic review on issues of prevention and treatment, because other study designs, all else being equal, are prone to measured and unmeasured bias.

Assessing the Quality of a Systematic Review

Introduction

In reviewing an article that purports to be a methodologically rigorous systematic review, the reader should look for key features that identify a truly systematic approach to literature evaluation.[3] First, the review should propose a specific and focused question. Thus, a review that purports to systematically address "antineoplastic therapy for diffuse large B-cell lymphoma" is most unlikely to be systematic, since the question is so broad that it cannot be answered with a simple systematic review; neither the intervention, population, or outcome variables of interest can be determined from such a question. Second, the method of literature review should be specified with sufficient clarity to ensure that the reader can determine if important, relevant studies were likely to have been omitted from the analysis. Third, explicit criteria that define the reasons why individual papers were selected or not selected for inclusion should be presented. Fourth, the reader should be able to determine from the extracted information if the primary studies included in the review were methodologically valid. Additional criteria that may support the methodologic rigor of a systematic review include the evaluation of possible reasons for differences among study results. To the extent that primary studies came to similar conclusions, this improves inferences that can be made from the systematic review.

A table outlining the Users' Guides to the Medical Literature[3] highlights traditional critical appraisal questions for systematic reviews and meta-analyses (**Table 1**). Regarding the results of a meta-analysis, if they are presented with figures, they are usually "forest plots" (**Figure 1**) that clearly define the source of the information included in the analysis, the number of patients potentially exposed to the intervention, the number of patients experiencing the outcome of interest, the calculated individual and summary relative risks or odds ratios, a measure of the weight of the impact of individual primary studies to the overall result, and a measure of the magnitude and precision of the treat-

Table 1. Questions that should be considered in determining if the results of systematic review are valid (adapted from Oxman et al[3]).

1) Did the overview address a focused clinical question?
2) Were the criteria used to select articles for inclusion both defined and appropriate?
3) What is the likelihood that relevant studies were missed?
4) Was the validity of the included studies assessed?
5) Were the assessments reproducible?
6) Were the study-to-study results congruent?
7) How precise were the results of the overview?
8) Were all clinically important outcomes considered?

In assessing the value of the review, it is important to consider the following questions:

1) Can the results be applied to my patients, and will the results help me care for my patients?
2) Are the benefits worth the harms and costs?

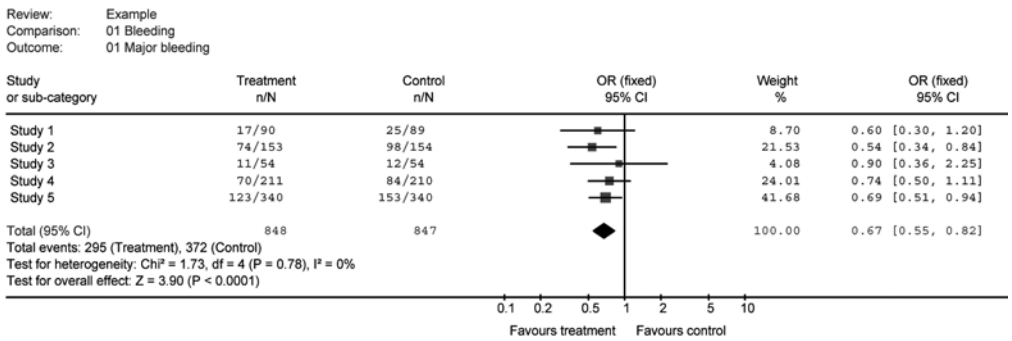

Figure 1. Sample forest plot (data are fictitious). This forest plot presents data from 5 fictional studies, two of which were "statistically significant" (studies 2 and 5) as indicated by confidence intervals that did not cross 1. Study 5 accounts for about 42% of the patients included in the systematic analysis. In sum, the meta-analysis suggests that the odds ratio in the combined analysis is 0.67 and is statistically significant, with odds ratios from 0.55 to 0.82. The advantage of systematic reviews over narrative reviews can also be inferred from this figure: a narrative review may have chosen to examine studies 1,3 and 4 and concluded that there was no difference in favour of treatment. Note that although this figure summarizes the results of the statistically pooled data, it does not explore the quality of the contributing studies (this information is usually presented in the text or an accompanying table).

ment effect overall. Such plots may also estimate the heterogeneity of the studies that contributed to the analysis. To judge the quality of a meta-analysis, a reader cannot just examine the forest plots, since these represent the seminal results of the meta-analysis but do not describe the quality of the review methods (or the quality of the contributing articles themselves).

Suggested Criteria

As the scientific methods for systematic review have developed and become more complex, there has been a need to develop criteria that allow the quality of these reviews to be determined. The key organization to help with the conduct of systematic reviews is the Cochrane Collaboration. This international, nonprofit, volunteer organization conducts and logs systematic reviews spanning the complete range of medical practice (http://www.cochrane.org/index.htm; accessed June 10, 2007). In discussing critical components in conducting a systematic review, we will follow the recommendations of the Cochrane Collaboration Handbook on Systematic Reviews and earlier documents outlining these methods.[4]

Frameworks for the evaluation of systematic reviews have been published and are beyond the scope of this manuscript (for example, see Harbour and Miller [5]). However, like evaluation frameworks from other fields, these systems assess the quality of the contributing data (in the case of systematic reviews, these are the contributing studies), the quality of the data extraction, the quality of the methodology used to combine the results, and the degree to which limitations in the analysis are identified and addressed by the authors. A critical but often overlooked step in the development of systematic reviews is a validity assessment for the included articles. In each case, this process requires assessing the validity of the primary studies using features unique to the study design. For meta-analyses of randomized trials, the validity would typically focus on the integrity of the allocation methods, blinding, and follow-up.

The Question

Fundamentally, the quality of a systematic review and the reliability of its result are contingent on both the quality of the contributing studies and the quality of the methodology used to produce the systematic review. In considering a systematic review, the first and most important step is the development of a well-formulated question; a poorly formulated question reduces both the quality and utility of the product. The questions should specify the characteristics of the participants who will be evaluated, the nature of the interventions to which those individuals will be exposed, and what outcomes will be measured. Ideally, the type of study that will contribute to the analysis will also be specified. Thus, if one wished to refine the question presented above, one might obtain a "better" systematic review if one were to formulate a question such as "Does the addition of rituximab to CHOP chemotherapy increase the likelihood of disease-free survival at 5 years after its administration to patients with diffuse large B-cell lymphoma who are enrolled in placebo-controlled studies?" This question identifies the population (patients with diffuse large cell lymphoma), the nature of the intervention (the addition of rituximab to CHOP chemotherapy), the outcome measure (disease-free survival at 5 years), and the design (randomized trials). Consumers of this literature will be able to immediately identify its relevance to their needs by simply reviewing the question.

The Systematic Review Protocol and Analysis Plan

Prior to initiating a systematic review, a protocol of steps should be developed, since a systematic review is a scientific exercise. A key part of that is an analysis plan. The

development of an analysis plan is beyond the scope of this review and should be undertaken in concert with an epidemiologist or biostatistician.

Searching for relevant articles

Development of a well-formulated question and review protocol will facilitate data acquisition and ensure that the results of the study can be replicated by other research groups. In general, a comprehensive list of potentially eligible articles is obtained using text search strategies in large bibliographic databases such as MEDLINE or EMBASE. Search strategies that are nonspecific will generate long lists of articles, few of which will be relevant. In contrast, excessively specific search strategies are prone to missing papers of importance. The results of the search strategy should be supplemented by manual review of the references of relevant papers, contact with content experts and pharmaceutical companies, and review of other resources likely to contain additional studies missed by the review of MEDLINE or EMBASE. In general, this phase of systematic reviews will be significantly shortened through consultation with a professional librarian. Additional studies may be identified from the reference list and discussion of other review articles. Other information sources that might be included in systematic reviews include registries of clinical effectiveness, registries of toxic effects, and epidemiologic datasets within which exposure to the intervention of interest can be reliably ascertained. However, typically these study designs are not included in reviews focused on randomized trials.

Selecting relevant articles

Once a list of articles is obtained, they should be reviewed by two or more individuals and compared with a list of predeveloped inclusion and exclusion criteria. In duplicate, independently, the reviewers should track their assessments. Disagreements should be reviewed by a third person or by discussion and consensus between the two reviewers. Cases of non-concordance on the eligibility of any individual contributing paper should be reported. Flow diagrams are useful figures to illustrate the searching and ultimate yield of primary articles included in the review.[6]

Sensitivity Analysis

The rigor of a systematic review may be increased by inclusion of a sensitivity analysis. Sensitivity analyses measure the impact of the results after adjustment of one or more characteristics of the studies. The strength of inference is greater if the results are unchanged under varying conditions. Examples of sensitivity analyses include comparing the pooled results of the lower versus higher methodologically rigorous studies. Another example of a sensitivity analysis would be measuring and comparing the results using different techniques to impute missing data.

Limitations of Systematic Reviews

Heterogeneity

All systematic reviews should examine reasons for heterogeneity of the contributing studies. There are two basic types of heterogeneity. The heterogeneity could be about patients (e.g., are the populations similar across studies? [sometimes called clinical heterogeneity]), or the heterogeneity could be focused on the results (e.g., are the results consistent across studies?, [sometimes called statistical heterogeneity]). The greater the statistical heterogeneity, in general, the weaker the inferences about the estimates of effect overall. By contrast, the greater the clinical heterogeneity, in general, the greater the generalizability of the overall results. Thus, evidence of strong treatment effect in a systematic review with moderate clinical heterogeneity could actually increase inferences from the review, since the effect was evaluated across differing patient groups, drug doses, or the like. By contrast, inconsistency in study results in similar patients undergoing uniform interventions is problematic, since the results of the systematic review are more likely to be "driven" by the size of the largest contributing studies, which may not be the studies of greatest methodological rigor. In such cases, it may be difficult to interpret the results of the systematic review.

Limited datasets and the strengths of conclusions based on a systematic review

A systematic review should not be viewed as a replacement for properly designed and sufficiently powered studies examining a treatment effect of interest. In fact, although systematic reviews may be the only way of obtaining reliable estimates of effect in some situations, systematic reviews of common interventions should be viewed as an opportunity to design and implement a large study to confirm the observation. Although systematic reviews are likely to detect defects in the literature that might not be seen by only examining individual studies, a systematic review that fails to support an intervention should not be seen as definitive evidence that such a treatment effect does not exist. While this may be the case, if the literature base is very small, then not only the primary studies may be underpowered, but the systematic review as well, making it prone to type II error (failure to include a treatment effect exists when one actually is present)

Systematic reviews are not entirely independent of the quality of contributing studies for their results. In areas with limited data, or when the quality of primary studies is poor, systematic reviews of these primary studies will be flawed. For example, sometimes systematic reviews are performed in disease states where there are only a few small studies. Such studies are prone to publication bias; as a result, meta-analysis of these datasets will produce a biased estimate of the treatment effect.

Inclusion of unpublished data

Whether unpublished studies should be included in a systematic review is contentious. Arguing for inclusion of unpublished data is "publication bias," which uniformly favors the publication of "positive" studies. However, routine inclusion of unpublished data may expose the systematic review to data of lesser quality, since unpublished data will not have been vetted through the peer review process and because it is difficult or impossible to identify all potential sources of unpublished data. Furthermore, unpublished data may be generated using less-rigorous techniques and thus may be more prone to bias. With the recent advent of clinical trial registries access to unpublished studies will likely increase. Prior to inclusion of data from unpublished data into a systematic review, authors are advised to compare the registered version of the protocol with the unpublished results to determine if the authors have deviated from their original analysis plan. If the version of the study registered at the outset corresponds well to the unpublished data available, it suggests that those results are less likely to have been influenced by bias, and thus their value to the systematic review is increased.

Influence of external agencies

Many systematic reviews are funded by organizations such as pharmaceutical companies or special interest groups. As in the design of randomized trials, the design of systematic reviews can be influenced (particularly through manipulation of inclusion and exclusion criteria) to select a particular set of studies. As a result, such systematic reviews may present a biased viewpoint. Careful assessment of the quality of the systematic review should reveal the flaws in their design. Another way in which bias can be introduced is through biased interpretation of the results of a systematic review funded by industry or authored by investigators who are influenced by industry.

Summary

In summary, systematic reviews are key evidence summaries. Systematic reviews can distil large volumes of literature into more manageable executive summaries. Meta-analyses provide more precise estimates of the effect of interventions on both benefits and harms of interventions. Ultimately, however, systematic reviews reflect the quality of the underlying data in the primary studies, and the rigor of the techniques used to search, identify, appraise, and combine the data to produce summary measures.

Correspondence

Mark A. Crowther, MD, MSC, St. Joseph's Hospital, 50 Charlton Ave. East, Rm L208, Hamilton, ON L8N 4A6, Canada; phone (905) 521-6024; fax (905) 540-6568; crowthrm@mcmaster.ca

References

1. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126:376-380.
2. Mulrow CD, Cook DJ, Davidoff F. Systematic reviews: critical links in the great chain of evidence. Ann Intern Med. 1997;126:389-391.
3. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature, VI: how to use an overview. Evidence-Based Medicine Working Group. JAMA. 1994;272:1367-1371.
4. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. J Clin Epidemiol. 1995;48:167-171.
5. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ. 2001;323:334-336.
6. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet. 1999;354:1896-900.