Guides for Reading and Interpreting Systematic Reviews

I. Getting Started

Terry P. Klassen, MD, MSc; Alejandro R. Jadad, MD, DPhil; David Moher, MSc

It is almost impossible for any individual to keep up-to-date with the health care literature. There are approximately 17,000 new biomedical books published every year, along with 30,000 biomedical journals, resulting in an annual increase of 7%. Physicians attempting to keep abreast of their field would need to read, on average, 19 original articles each day. Reviews offer the potential to reach that elusive goal: making life easier.

In the past few years, the Archives has published systematic reviews on the role of metered-dose inhaler accessory devices compared with nebulizers for the treatment of acute asthma management, the efficacy of bronchodilator therapy for bronchiolitis, and the behavioral and cognitive effects of methylxanthines. These are all complex clinical issues, and for any one clinician to gather and synthesize the evidence would be very time-consuming.

NARRATIVE REVIEWS

Traditionally, individuals often considered experts in the field have conducted narrative reviews of the literature, associated with a particular health condition (e.g., the treatment of children with seizures), using informal and subjective methods to collect and interpret information.

Readers of narrative reviews face at least 2 problems. First, they are not provided with a detailed description of the process that led to the review. Second, because of this lack of information, readers cannot replicate and verify the results and conclusions of the reviews. The difficulties in verifying and replicating narrative reviews have been highlighted repeatedly during the past 10 years.

SYSTEMATIC REVIEWS AND META-ANALYSES

A systematic review is a review in which there is a comprehensive search for relevant studies on a specific topic, and those identified are then appraised and synthesized according to a predetermined and explicit method. We believe this systematic approach provides the reader with a unique advantage over any other type of review: the ability to replicate it. A meta-analysis is the statistical combination of at least 2 studies to produce a single estimate of the effect of the health care intervention under consideration.

For editorial comment see page 622

The simple act of statistically combining studies does not guarantee a valid or reliable answer to a question. Indeed, recent evidence indicates that many meta-analyses are not conducted rigorously. If meta-analyses are to be reliable, however, they should involve systematic and detailed processes, which include measures to reduce bias and imprecision, that are reported clearly and precisely to allow replication by others. For the remainder of this series, we will use the term systematic quantitative review to describe this systematic approach. While systematic reviews have

From the Departments of Pediatrics (Dr Klassen and Mr Moher), and Epidemiology and Community Health (Dr Klassen and Mr Moher), University of Ottawa, Thomas C. Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario's Research Institute (Dr Klassen and Mr Moher) Ottawa; and Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton (Dr Jadad) Ontario.
some important advantages, they also have a few limitations and, if not conducted, reported, and interpreted appropriately, may even be potentially harmful.

**ADVANTAGES**

Systematic quantitative reviews that are conducted and reported appropriately can help resolve controversies between conflicting studies, guide clinical research by providing new hypotheses, identify areas in which insufficient research has been performed or in which additional research may be unethical, and identify beneficial and harmful therapies years before other types of reviews can.13

**LIMITATIONS**

It is difficult to identify all potentially relevant studies to include in a systematic review. In general, electronic searches of the literature identify only 50% of all relevant articles.14 Hand-searching journals may be a useful adjunct to electronic searching but is far more labor intensive and is likely not feasible for many reviewers. Identifying “gray literature” (ie, studies that are unpublished, have limited distribution, and are not included in bibliographic databases) is also troublesome, and there is disagreement concerning its usefulness.15

Even after all relevant studies are identified, their quality is likely to be variable. Variability in the quality of studies included in systematic quantitative reviews can lead to confusing results.16 For example, systematic quantitative reviews that include all relevant studies may indicate an intervention to be effective; however, when the review is limited to those studies of higher quality, the effectiveness disappears. Similar results to these, but in the opposite direction, have also been reported.17

**POTENTIAL HARMs**

As meta-analysis becomes more popular, so does the number of user-friendly computer packages to assist in their conduct. The danger here is that it only takes 2 studies to “conduct” a meta-analysis, and many reviewers may harbor a “combination” urge (ie, a desire to statistically combine studies regardless of its appropriateness).18 Results from such meta-analyses may lead to erroneous conclusions, as they will have an aura of increased precision but may be as prone to bias as narrative reviews. There is little evidence as to the most appropriate conditions under which a systematic review might become a systematic quantitative review.19

To overcome these limitations and maximize the benefits of systematic reviews to health care decisions, an international group of individuals is preparing, maintaining, and disseminating up-to-date systematic reviews of research evidence in all areas of health care. This group is known as the Cochrane Collaboration.20,21

One message of this series is that just because an article is described as a systematic review or meta-analysis, it does not mean that it has been conducted and reported well. A 1987 survey22 of 86 meta-analyses, published in English, assessed each publication on 14 items from 6 content areas thought to be important in the conduct and reporting of meta-analyses of randomized controlled trials: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. The results showed that only 24 (28%) of the 86 meta-analyses addressed all 6 content areas. This survey was updated in 1992, with little change in the results.23 This highlights that not only authors but also some editors and peer reviewers do not fully appreciate the elements that should be taken into account during the design, conduct, and reporting of reviews. It is also likely that some readers do not know how to read and interpret systematic reviews, making it difficult to judge their validity. This is our main justification for this series.

**DEVELOPING THE QUESTION**

Central to all high-quality systematic reviews and meta-analyses is a clearly formulated primary question. A fuzzy question leads to a fuzzy answer or perhaps no answer at all. The start of all scientific query must begin with a question or hypothesis. The question, usually found in the title, abstract, or introduction, will help the reader discern the potential relevance of a systematic review to his or her area of clinical practice. Sacks et al22 found that only 6 (7%) of 86 meta-analyses they reviewed had clear evidence of a written protocol that included a question. If the reader can find no focused question, it is probably best not to spend the time reading that review article but rather to move on to the next one.

The importance of the question is supported from recent work by Cook and colleagues24 demonstrating that the nature of the question can in part explain the different conclusions reached by discordant meta-analyses. Two meta-analyses had been published (1 by Cook et al25 and 1 by Tryba26) that examined stress ulcer prophylaxis in critically ill patients. The conclusions of the meta-analyses, however, differed. These 2 groups of authors then collaborated and performed an updated meta-analysis and explored reasons for discrepancies that had arisen in their initial meta-analyses. In examining stress ulcer prophylaxis in critically ill patients, Cook et al25 had considered only 3 of the most commonly used stress ulcer prophylactic agents, whereas Tryba26 had examined the effect of all stress ulcer prophylaxis agents, including pirenzipine hydrochloride and prostaglandins.

One described method for posing a question is to use a format that contains the dimensions of population, intervention, and outcome.27 Consider the following scenario. A mother brings in her 8-year-old boy to your office because she is concerned that sugar in his diet is making him hyperactive and preventing him from concentrating in school. She wants your opinion about this possible relationship. Not immediately knowing the answer, you reschedule an appointment for several weeks later. Before this appointment you search for a systematic review on the topic and find one entitled “The Effect of Sugar on Behavior or Cognition in Children.”28 The question posed by this study is: Does sugar in the diet (interven-
Assessing the Validity of Systematic Reviews*

1. Did the systematic review address a focused, clinical question?
2. Were the criteria used to select articles for inclusion appropriate?
3. Is it likely that important, relevant studies were missed?
4. Was the quality of included studies assessed?
5. Were the assessments of studies reproducible?
6. Were the study results similar across studies?

*Adapted from Oxman et al.32

The development of a clearly defined primary question for the systematic review helps guard against a fishing expedition. Such an expedition would occur if the systematic reviewers did not have a clear primary question and, hence, began fishing among the results of a collection of studies (also referred to as data dredging) until they came up with a “positive result.” When enough statistical comparisons are performed, eventually 1 of them will be positive by chance alone. The primary question, developed a priori, should determine the primary focus of the review. Secondary questions, which should also be posed a priori, form the basis for secondary analyses, such as subgroup analyses. Oxman and Guyatt30 have provided guidelines for the performance of such subgroup analyses.

DEVELOPING SELECTION CRITERIA

With a well-worded question, the development of the selection criteria is straightforward. Clear and concise selection criteria are important to minimize bias and errors when reviewers are selecting studies to be included in the systematic review. Selection criteria should describe which studies would be eligible for inclusion on the basis of a description of the population, intervention, outcome measures, and study design. In addition, the selection criteria should clearly state from what period studies were selected, what languages of publication were eligible, and whether published and unpublished studies were eligible for inclusion. A rationale for the selection criteria should be provided. For example, if only English-language studies are considered, this must be justified because the quality of reporting studies does not differ significantly between English- and non–English-language studies.

Although not previously well described or documented, inclusion criteria bias was described by Felson.30 This bias occurs when the investigator creates a set of selection criteria based on a preliminary review of the literature. These criteria could exclude some important studies that the investigator had been aware of, and this in turn could produce bias. It is difficult to guard against this, since some knowledge of the existing literature is important to guide the formation of the question. It is important for the systematic reviewers to define and justify their selection criteria on the basis of the need to answer an important clinical question. Inclusion of the existing evidence is probably a safer approach, with the reviewer then dealing with study differences within the review itself.

VALIDITY AND SYSTEMATIC REVIEWS

Before even looking at the conclusions, it is important that the reader decide whether they will be credible. The only validated assessment tool for measuring the quality of systematic reviews is one by Oxman and Guyatt.31 Oxman et al32 simplified this tool for use by clinicians to assess the validity of reviews (Table). Many of these issues relate to the validity (ie, how close it is to truth or freedom from bias) of the systematic review. The more valid the systematic review, the more confidence the reader can have in its results for use in clinical practice. One approach is to place the reviews in 3 categories. The best systematic reviews, or the most valid, would be those that meet all the criteria, with only minor omissions. The worst systematic reviews, or those most likely to contain bias, would meet almost none of the criteria. The rest of the reviews would be in the middle, meeting only some of the criteria.

SETTING THE STAGE: AN EXAMPLE

The parents of a 2-year-old girl come to your office because the child has had 6 ear infections in the previous 6 months. They are wondering whether anything can be done for their child to prevent these frequent ear infections. You review the article “Use of Antibiotics in Preventing Recurrent Acute Otitis Media and in Treating Otitis Media With Effusion” to help guide you on this issue.33 You use the guide in the Table to evaluate it.

Did the systematic review address a focused, clinical question? While this review addresses a number of different questions, the question relevant to your clinical problem is clearly stated in the introduction: “What is the magnitude of treatment effect, if any, of prophylactic antibiotics in suppression of recurrent AOM [acute otitis media]?”

©1998 American Medical Association. All rights reserved.
Were the criteria used to select articles for inclusion appropriate? The inclusion criteria are appropriately linked to the research question; in the “Methods” section, the authors state that they included “only published, randomized, controlled trials of the use of antibiotics in recurrent AOM . . . .” They do not state the age of the population of interest or whether they restricted their focus to only studies published in English.34,35

Is it likely that important, relevant studies were missed? The authors’ search strategy included a MEDLINE search from 1966 to 1993, a review of reference lists of identified studies, and a hand search of textbooks, monographs, and Current Contents from 1990 to 1992. Their MEDLINE search strategy is listed in the “Methods” section. Their adopted search strategy, although not exhaustive, is probably of reasonable effort for their question. They address the issue of possible missing evidence by calculating how many negative studies would be required to negate the beneficial treatment effect seen in their meta-analysis. According to Rosenthal’s formula,36 324 unpublished studies with no treatment effect would be needed to negate their positive finding, inferring that their conclusions are probably robust.

Was the quality of included studies assessed? The authors assess the quality of the primary studies by means of a scale developed by Chalmers et al.37 On review of this scale, it becomes obvious that many of the included items do not directly relate to the internal validity of the study. In addition, the authors do not use the score in relation to the study results. Therefore, it is not clear that the authors are measuring study quality or how this study quality is supposed to affect the results of the meta-analysis.

Were the assessments of studies reproducible? It is not clear whether the authors used 2 reviewers to select studies for inclusion in the systematic review. For assessing study quality, 2 blind reviewers were used, and when differences arose, a third blind reviewer was used to resolve differences.

Were the study results similar across studies? While there was no statistical measurement of heterogeneity, it is obvious from Figure 1 in the study by Williams et al.33 that there are differences between the different studies. The authors then try to explain some of these study differences by examining the type of antibiotic used in the study. Studies that used sulfisoxazole tended to have a better outcome than those that used other antibiotics. Another possible source of heterogeneity was that there tended to be a larger treatment effect in studies that treated patients for less than 6 months than in studies in which patients were treated for a longer period. The number of episodes of otitis media in the preceding 6 months did not affect the degree of benefit conferred by prophylactic antibiotic treatment.

Overall, this systematic review appears to be of reasonable quality and to have reliable conclusions. The risk difference in favor of preventing otitis media by the use of antibiotics is 0.11 episode of acute otitis media per patient-month. Given this result, 10 patients would need to be treated with antibiotics to prevent 1 episode of acute otitis media per month of treatment. In the era of increasing antibiotic-resistant bacteria, this degree of benefit may not be sufficient to embark on treatment with oral sulfisoxazole.

CONCLUSIONS

We have focused this series on systematic reviews of randomized controlled trials, as these have been the focus of most methodological research. Throughout the series, the clinical examples used are focused on reviews and evaluations of interventions. The discipline of systematic reviews is evolving and is likely to change as evidence accumulates. Therefore, this series should not be considered as the “last word” (ie, written in stone). We intend to update the series accordingly.

We think all systematic reviews need to have a minimum amount of information for readers, and ultimately patients, to be of any benefit. We recognize that there may be practical reasons why some items, or a more complex assessment of items, are not included in the report of a systematic review. However, readers need to have the tools necessary to balance between the relevance of missing information and its potential influence on the validity of what they are reviewing. In each article in the series we have tried to reflect this compromise.

This is the first of a 3-part series. The other parts to follow in the August and September issues.

Accepted for publication December 29, 1997.

Reprints: David Moher, MSc, Thomas C. Chalmers Center for Systematic Reviews, Children’s Hospital of Eastern Ontario’s Research Institute, 401 Smyth Rd, Ottawa, Ontario, Canada K1H 8L1 (e-mail: 107656.3375@compuserve.com).

REFERENCES


