Systematic Review and Meta-analysis: When One Study Is Just not Enough

Amit X. Garg,* Dan Hackam,† and Marcello Tonelli‡
*Division of Nephrology and Department of Epidemiology and Biostatistics, University of Western Ontario, London, and ‡Division of Clinical Pharmacology and Toxicology, University of Toronto, and Cardiac Rehabilitation and Secondary Prevention Program, Toronto Rehabilitation Institute, Toronto, Ontario, and †Division of Nephrology and Department of Public Health Sciences, University of Alberta, and Institute of Health Economics, Edmonton, Alberta, Canada

W

e live in the information age, and the practice of medicine is becoming increasingly specialized. In the biomedical literature, the number of published studies has dramatically increased: There are now more than 15 million citations in MEDLINE, with 10,000 to 20,000 new citations added each week (1). Multiple relevant studies usually guide most clinical decisions. These studies often vary in their design; methodologic quality; population studied; and the intervention, test, or condition considered. Because even highly cited trials may be challenged or refuted over time (2), clinical decision-making requires ongoing reconciliation of studies that provide different answers to the same question. Both clinicians and researchers can also benefit from a summary of where uncertainty remains. Because it is often impractical for readers to track down and review all of the primary studies (3), review articles are an important source of summarized evidence on a particular topic (4).

Narrative Review, Systematic Review, and Meta-analysis

Review articles have traditionally taken the form of a narrative review, whereby a content expert writes about a particular field, condition, or treatment (5–7). Narrative reviews have many benefits, including a broad overview of relevant information tempered by years of practical knowledge from an experienced author. Indeed, this article itself is in a narrative format, from authors who have published a number of meta-analyses in previous years.

In some circumstances, a reader wants to become very knowledgeable about specific details of a topic and wants some assurance that the information presented is both comprehensive and unbiased. A narrative review typically uses an implicit process to compile evidence to support the statements being made. The reader often cannot tell which recommendations were based on the author’s clinical experience, the breadth to which available literature was identified and compiled, and the reasons that some studies were given more emphasis than others. It is sometimes uncertain whether the author of a narrative review selectively cited reports that reinforced his or her preconceived ideas or promoted specific views of a topic. Also, a quantitative summary of the literature is often absent in a narrative review.

A systematic review uses a process to identify comprehensively all studies for a specific focused question (drawn from research and other sources), appraise the methods of the studies, summarize the results, present key findings, identify reasons for different results across studies, and cite limitations of current knowledge (8,9). In a systematic review, all decisions used to compile information are meant to be explicit, allowing the reader to gauge for him- or herself the quality of the review process and the potential for bias. In this way, systematic reviews tend to be more transparent than their narrative cousins, although they too can be biased if the selection or emphasis of certain primary studies is influenced by the preconceived notions of the authors or funding sources (10).

Depending on the nature of the data, the results of a systematic review can be summarized in text or graphic form. In graphic form, it is common for different trials to be depicted in a plot where the point estimate and 95% confidence interval for each study are presented on an individual line (11). When results are mathematically combined (a process sometimes referred to as pooling), this is referred to as meta-analysis. Graphically, the pooled result is often presented as a diamond at the bottom of the plot.

When performing a meta-analysis, a review team usually combines aggregate-level data reported in each primary study (point and variance estimate of the summary measure). On occasion, a review team will obtain all of the individual patient data from each of the primary studies (12,13). Although challenging to conduct (14), individual patient meta-analyses may have certain advantages over aggregate-level analyses. As highlighted in a review of angiotensin-converting enzyme (ACE) inhibitors for nondiabetic kidney disease, this includes the use of common definitions, coding and cutoff points be-
tween studies, addressing questions not examined in the original publication, and a better sense of the impact of individual patient (versus study level) characteristics (12,15).

As first highlighted a decade ago (16), the number of systematic reviews in nephrology and other fields has increased dramatically with time, paralleling the rapid growth of biomedical literature during the past half century. Initiatives such as the Cochrane Collaboration have further increased the profile and rigor of the systematic review process (details of the structured process of Cochrane systematic reviews are available through their Web site) (17,18). From 1990 to 2005, there were more than 400 systematic reviews and meta-analyses published in the discipline of nephrology (Figure 1). Of these reviews, 40% pertained to chronic kidney disease or glomerulonephritis and 20, 16, 15, and 7% pertained to kidney transplantation, dialysis, acute kidney injury, and pediatric nephrology, respectively. As a publication type, however, systematic reviews have not been without controversy: Some authors consider a meta-analysis the best possible use of all available data, whereas others question whether they add anything meaningful to scientific knowledge (19). The strengths and weaknesses of this publication type are described next.

**Strengths of Systematic Review and Meta-analysis**

Physicians make better clinical decisions when they understand the circumstances and preferences of their patients and combine their personal experience with clinical evidence underlying the available options (20). The public also expects that their physicians will integrate research findings into practice in a timely way (21). Thus, sound clinical or health policy decisions are facilitated by reviewing the available evidence (and its limitations), understanding reasons why some studies differ in their results (a finding sometimes referred to as heterogeneity among the primary studies), coming up with an assessment of the expected effect of an intervention or exposure (for questions of therapy or etiology), and then integrating the new information with other relevant treatment, patient, and health care system factors.

In this respect, reading a properly conducted systematic review is an efficient way to become familiar with the best available research evidence for a focused clinical question. The review team may also have obtained information from the primary authors which was not available in the original reports. The presented summary allows the reader to take into account a whole range of relevant findings from research on a particular topic. The process can also establish whether the scientific findings are consistent and generalizable across populations, settings, and treatment variations and whether findings vary significantly by particular subgroups. Again, the potential strength of a systematic review lies in the transparency of each phase of the synthesis process, allowing the reader to focus on the merits of each decision made in compiling the information, rather than a simple contrast of one study to another as sometimes occurs in other types of reviews.

For example, studies demonstrating a significant effect of treatment are more likely to be published than studies with negative findings, are more likely to be published in English, and more likely to be cited by others (22–27). A well-conducted systematic review attempts to reduce the possibility of bias in the method of identifying and selecting studies for review, by using a comprehensive search strategy and specifying inclusion criteria that ideally have not been influenced by a priori knowledge of the primary studies.

Mathematically combining data from a series of well-conducted primary studies may provide a more precise estimate of the underlying “true effect” than any individual study (28). In other words, by combining the samples of the individual studies, the size of the “overall sample” is increased, enhancing the statistical power of the analysis and reducing the size of the confidence interval for the point estimate of the effect. It is also more efficient to communicate a pooled summary than to describe the results for each of the individual studies. Sometimes, if the treatment effect in small trials shows a nonsignificant trend toward efficacy, then pooling the results may establish

**Figure 1.** There have been more than 400 systematic reviews and meta-analyses published in the discipline of nephrology since 1990, with the annual number increasing with time. Frequencies were estimated from a MEDLINE and EMBASE search performed by an experienced renal librarian in December 2006. Citations were reviewed by a nephrologist for relevance. Duplicate publications from the same group of authors were counted only once.
the benefits of therapy (16). For example, 10 trials examined whether ACE inhibitors were more effective than other antihypertensive agents for the prevention of nondiabetic kidney failure (29). Many of the 95% confidence intervals for the estimate provided by each study overlapped with a finding of no effect; however, the overall pooled estimate established a benefit of ACE inhibitors.

For these reasons, a meta-analysis of similar, well-conducted, randomized, controlled trials has been considered one of the highest levels of evidence (30–32). It is important to stress that the primary trials all have to be conducted with high methodologic rigor for the meta-analysis to be definitive. Alternatively, when the existing studies have important scientific and methodologic limitations, including smaller sized samples (which is more often the case), the systematic review may identify where gaps exist in the available literature. In this case, an exploratory meta-analysis can provide a plausible estimate of effect that can be tested in subsequent studies (33,34).

**Limitations of Systematic Review and Meta-analysis**

This type of publication type has many potential limitations that should be appreciated by all readers. First, the summary provided in a systematic review and meta-analysis of the literature is only as reliable as the methods used to estimate the effect in each of the primary studies. In other words, conducting a meta-analysis does not overcome problems that were inherent in the design and execution of the primary studies. It also does not correct biases as a result of selective publication, whereby studies that report dramatic effects are more likely to be identified, summarized, and subsequently pooled in meta-analysis than studies that report smaller effect sizes (an issue referred to as publication bias). Because more than three quarters of meta-analyses did not report any empirical assessment of publication bias (35), the true frequency of this form of bias is unknown.

Controversies also arise around the interpretation of summarized results, particularly when the results of discordant studies are pooled in meta-analysis (36). The review process inevitably identifies studies that are diverse in their design, methodologic quality, specific interventions used, and types of patients studied. There is often some subjectivity when deciding how similar studies must be before pooling is appropriate. Combining studies of poor quality with those that were more rigorously conducted may not be useful and can lead to worse estimates of the underlying truth or a false sense of precision around the truth (36). A false sense of precision may also arise when various subgroups of patients defined by characteristics such as their age or gender differ in their observed response. In such cases, reporting an aggregate pooled effect might be misleading if there are important reasons to explain variable treatment effects across different types of patients (36–40).

Finally, simply labeling a manuscript as a “systematic review” or “meta-analysis” does not guarantee that the review was conducted or reported with due rigor (41). To reduce the chance of arriving at misleading conclusions, guidelines on the conduct and reporting of systematic reviews were recently published (42,43); however, important methodologic flaws of systematic reviews published in peer-reviewed journals have been well described (44–54). For example, of the 86 renal systematic reviews published in 2005, the majority (58%) had important methodologic flaws (Mrkobrada M, Thiessen-Philbrook H, Haynes RB, Iansavichus AV, Rehman F, and Garg AX, submitted). The most common flaws among these renal reviews were failure to assess the methodologic quality of included primary studies and failure to avoid bias in study inclusion (Mrkobrada M, Thiessen-Philbrook H, Haynes RB, Iansavichus AV, Rehman F, and Garg AX, submitted). In some cases, industry-supported reviews of drugs have had fewer reservations about methodologic limitations of the included trials than rigorously conducted Cochrane reviews on the same topic (10); however, the hypothesis that less rigorous reviews more often report positive conclusions than good-quality reviews of the same topic has not been borne out in empirical assessment (48,53,55). Nonetheless, like all good consumers, users of systematic reviews should carefully consider the quality of the product and adhere to the dictum “caveat emptor”: Let the buyer beware. The limitations described in this section may explain differences in the results of meta-analyses as compared with subsequent large, randomized, controlled trials, which have occurred in approximately one third of cases (56).

**How to Appraise Critically a Systematic Review and Meta-analysis**

Users of systematic reviews need to assure themselves that the underlying methods used to assemble relevant information were sound. Before considering the results or how the information could be appropriately applied in patient care (9), there are a few questions that the reader can ask him- or herself when

<table>
<thead>
<tr>
<th>Table 1. Questions to ask when assessing the quality of a systematic review*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the review conducted according to a prespecified protocol?</td>
</tr>
<tr>
<td>2. Was the question focused and well formulated?</td>
</tr>
<tr>
<td>3. Were the right types of studies eligible for the review?</td>
</tr>
<tr>
<td>4. Was the method of identifying all relevant information comprehensive?</td>
</tr>
<tr>
<td>a. Is it likely that relevant studies were missed?</td>
</tr>
<tr>
<td>b. Was publication bias considered?</td>
</tr>
<tr>
<td>5. Was the data abstraction from each study appropriate?</td>
</tr>
<tr>
<td>a. Were the methods used in each primary study appraised?</td>
</tr>
<tr>
<td>6. Was the information synthesized and summarized appropriately?</td>
</tr>
<tr>
<td>a. If the results were mathematically combined in meta-analysis, then were the methods described in sufficient detail, and was it reasonable to do so?</td>
</tr>
</tbody>
</table>

*Adapted from reference (9).
assessing the methodologic quality of a systematic review (Table 1).

**Was the Review Conducted According to a Prespecified Protocol?**

It is reassuring if a review was guided by a written protocol (prepared in advance) that describes the research question(s), hypotheses, review method, and plan for how the data will be extracted and compiled. Such an approach minimizes the likelihood that the results or the expectations of the reviewing team influenced study inclusion or synthesis. Although most systematic reviews are conducted in a retrospective manner, reviews and meta-analyses can in theory be defined at the time several similar trials are being planned or under way. This allows a set of specific hypotheses, data collection procedures, and analytic strategies to be specified in advance before any of the results from the primary studies are known. Such a prospective effort may provide more reliable answers to medically relevant questions than the traditional retrospective approach (41).

**Was the Question Focused?**

Clinical questions often deal with issues of treatment, etiology, prognosis, and diagnosis. A well-formulated question usually specifies the patient’s problem or diagnosis, the intervention or exposure of interest, any comparison group (if relevant), and the primary and secondary outcomes of interest (57).

**Were the “Right” Types of Studies Eligible for the Review?**

Different study designs can be used to answer different clinical questions. Randomized, controlled trials; observational studies; and cross-sectional diagnostic studies may each be appropriate depending on the primary question posed in the review. When examining the eligible criteria for study inclusion, the reader should feel confident that a potential bias in the selection of studies was avoided. Specifically, the reader should ask her- or himself whether the eligibility criteria for study inclusion were appropriate for the question asked. Whether the right types of studies were selected for the review also depends on the depth and breadth of the underlying literature search.

For example, some review teams will consider only studies that were published in English. There is evidence that journals from certain countries publish a higher proportion of positive trials than others (58). Excluding non-English studies seemed to change the results of some reviews (59,60) but not others (61,62).

Some review teams use broad criteria for their inclusion of primary studies (e.g., effects of agents that block the renin-angiotensin system on renal outcomes [63]), whereas other teams use more narrow inclusion criteria (e.g., restricting the analysis only to patients who have diabetes without evidence of nephropathy [64]). There is often no single correct approach; however, the conclusions of any meta-analysis that is highly sensitive to altering the entry criteria of included studies should be interpreted with some caution (25). For example, two different review teams considered whether synthetic dialysis membranes resulted in better clinical outcomes compared with cellulose-based membranes in patients with acute renal failure. In one meta-analysis (65) but not the other (66), synthetic membranes reduced the chance for death. The discordant results were due to the inclusion of a study that did not meet eligibility for the second review (67).

**Was the Method of Identifying All Relevant Information Comprehensive?**

Identifying relevant studies for a given clinical question among the many potential sources of information is usually a laborious process (68). Biomedical journals are the most common source of information, and bibliographic databases are frequently used to search for relevant articles. MEDLINE currently indexes approximately 4800 medical journals and contains 13 million citations (69). Similarly, EMBASE indexes approximately 5000 medical journals and contains more than 11 million records. There are some key differences between EMBASE and MEDLINE, and the review team should have searched both databases (70–72). For example, EMBASE provides the best coverage of European research as well as pharmaceutical research including renal adverse events (73). Positive studies may be more often published in journals that are indexed in MEDLINE, compared with nonindexed journals (25).

Depending on the question posed, other databases may also have been searched. For example, if a team is summarizing the effects of exercise training in patients who receive maintenance hemodialysis, then searching the Cumulative Index to Nursing and Allied Health Literature (CINAHL) database would be appropriate (74). Alternatively, the ECONOLIT database may be useful for identifying information on the out-of-pocket expenses incurred by living kidney donors (75). As a supplementary method of identifying information, searching databases such as the Science Citation Index (which identifies all articles that cite a relevant article), as well as newer Internet search engines such as Google Scholar and Elsevier’s Scirus, can be useful for identifying articles that are not indexed well in traditional bibliographic databases (76). Searching bibliographies of retrieved articles can also identify relevant articles that were missed.

Whatever bibliographic database was used, the review team should have used a search strategy that maximized the identification of relevant articles (77,78). Because there is some subjectivity in screening databases, citations should be reviewed independently and in duplicate by two members of the reviewing team, with the full-text article retrieved for any citation deemed relevant by any of the reviewers. There is also some subjectivity in assessing the eligibility of each full-text article, and the risk for incorrectly discarding relevant reports is reduced when two reviewers independently perform each assessment in a reliable manner (79).

Important sources of information other than journal articles should not be overlooked. Conference proceedings, abstracts, books, and manufacturers all can be sources of potentially valuable information. Inquiries to experts, including those listed in trial registries, may have also proved useful (28).

A comprehensive search of available literature reduces the possibility of publication bias, which occurs when studies with statistically significant results are more likely to be published.
and cited (80,81). It is interesting that some recent reviews of acetylcysteine for the prevention of contrast nephropathy analyzed as few as five studies, despite being submitted for publication almost 1 yr after publication of a review of 12 studies (82). Although there are many potential reasons for this, one cannot exclude the possibility that some search strategies missed eligible trials. In addition to a comprehensive search method, which makes it unlikely that relevant studies were missed, it is often reassuring if the review team used graphic and statistical methods to confirm that there was little chance that publication bias influenced the results (83).

Was the Data Abstraction from Each Study Appropriate?
In compiling relevant information, the review team should have used a rigorous and reproducible method of abstracting all relevant data from the primary studies. Often two reviewers abstract key information from each primary study, including study and patient characteristics, setting, and details about the intervention, exposure, or diagnostic test as is appropriate. Language translators may be needed. Teams who conduct their review with due rigor will indicate that they contacted the primary authors from each of the primary studies to confirm the accuracy of abstracted data as well as to provide additional relevant information that was not provided in the primary report. Some authors will go through the additional effort of blinding or masking the results from other study characteristics so that data abstraction is as objective as possible (84,85).

One element that should have been abstracted is the methodologic quality of each primary study (recognizing this is not always as straightforward as it may first seem) (86–91). The question to be posed by the reader is whether the reviewing team considered if each of the primary studies was designed, conducted, and analyzed in a way to minimize or avoid biases in the results (92). For randomized, controlled trials, lack of concealment of allocation, inadequate generation of the allocation sequence, and lack of double blinding can exaggerate estimates of the treatment effect (54,90,93). The value of abstracting such data is that it may help to explain important differences in the results among the primary studies (90).

For example, long-term risk estimates can become unreliable when participants are lost to study follow-up; those who participate in follow-up often systematically differ from nonparticipants. For this reason, prognosis studies are vulnerable to bias, unless the loss to follow-up is less than 20% (94). In a systematic review of 49 studies on the renal prognosis of diarrhea associated hemolytic uremic syndrome, on average, 21% of patients were lost to follow-up (range 0 to 59% across studies) (95). It was hypothesized that patients who were lost to follow-up would contribute to worse estimates of long-term prognosis because they are typically healthier than those who continue to be followed by their nephrologists. Indeed, studies with a higher proportion of patients lost to follow-up demonstrated a higher proportion of patients with long-term renal sequelae, explaining 28% of the between-study variability.

How Was the Information Synthesized and Summarized?
In cases in which the primary studies differ in the design, populations studied, interventions and comparisons used, or outcomes measured, it may have been appropriate for the review team simply to report the results descriptively using text and tables. When the primary studies are similar in these characteristics and the studies provide a similar estimate of a true effect, then meta-analysis may have been used to derive a more precise estimate of this effect (96). In meta-analysis, data from the individual studies are not simply combined as though they were from a single study; rather, greater weights are given to the results from studies that provide more information, because they are likely to be closer to true effect being estimated. Mathematically combining the results from the individual studies can be accomplished under the assumption of “fixed” effects or “random” effects model. Although a thorough description and merits of each approach is described elsewhere (97), it is fair to say that a random-effects model is more conservative than the fixed-effects approach, and a finding that is statistically significant with the latter but not the former should be viewed with skepticism.

Whenever individual studies are pooled in meta-analysis, it is important for the reader to determine whether it was reasonable to do so. One way to assess the similarity of various studies is to inspect the graphic display of the results, looking for similarities in the direction of the estimated effect. Even without considering any combined meta-analytic result, a reader becomes much more confident when a similar effect is being observed across many studies (i.e., the results have replicated across many studies). Some review teams may report a statistical test to determine how different the studies are from one another (as described previously, this is often termed heterogeneity of the study results [98]). This can help to prove or disprove that differences in the results that were observed between the primary studies is no different from what would be expected by chance. The most common statistical test to quantify heterogeneity is something called the Q statistic, which is similar in concept to a y^2 test. Although a nonsignificant result (by convention P > 0.1) is often taken to indicate that there are no substantial differences between the studies, it is important to consider that this test is underpowered, especially when the number of studies being pooled is small. A new statistic that is frequently being reported in meta-analysis these days is something called the I^2 statistic. This statistic describes the percentage variability between the studies that is present beyond what would be expected by chance. When interpreting an I^2 statistic, values of 0 to 30, 31 to 50, and >50% represent mild, moderate, and marked differences between the studies, respectively (99).

Whenever a review team identifies significant differences between the primary studies, they should try to explain possible reasons for these differences. This can be done in an informal way by analyzing certain types of studies separately or by selectively combining studies to determine which are particularly different from the remaining studies. Alternatively, a statistical approach can be taken to explore differences across studies, using a technique similar to linear or logistic regression.
(which at the study level is something called meta-regression) (100). Either way, a careful exploration of why study results differ can yield important information about potential determinants of the effect being observed.

Conclusions

Like all types of research, systematic reviews and meta-analyses have both potential strengths and weaknesses. With the growth of renal clinical studies, an increasing number of these types of summary publications will certainly become available to nephrologists, researchers, administrators, and policy makers who seek to keep abreast of recent developments. To maximize their advantages, it is essential that future reviews be conducted and reported properly, with judicious interpretation by the discriminating reader.

Acknowledgments

A.X.G. was supported by a Clinician Scientist Award from the Canadian Institutes of Health Research (CIHR). D.H. was supported by a CIHR Fellowship Award, the Chisholm Memorial Fellowship, and the Clinician-Scientist Training Program of the University of Toronto. M.T. was supported by a Population Health Investigator Award from the Alberta Heritage Foundation for Medical Research and a New Investigator Award from the CIHR.

We thank Drs. Chi Hsu and Harvey Feldman for help and advice. We thank Arthur Iansavichus, MLIS, who helped compile systematic reviews published in the discipline of nephrology.

Disclosures

None.

References


64. Strippoli GF, Craig MC, Schena FP, Craig JC: Role of blood pressure targets and specific antihypertensive agents used to prevent diabetic nephropathy and delay its progression. *Am Soc Nephrol* 17: S153–S155, 2006


82. Balk EM, Lau J, Bonis PA: Reading and critically appraising systematic reviews and meta-analyses: A short primer with a focus on hepatology. J Hepatol 43: 729–736, 2005


