Appraising Systematic Reviews and Meta-analyses

Michael Bigby, MD; Hywel Williams, MSc, PhD, FRCP

A systematic review is an overview that answers a specific clinical question and contains a thorough, unbiased search of the relevant literature, explicit criteria for assessing studies, and a structured presentation of the results. Many systematic reviews incorporate a meta-analysis, i.e., a quantitative pooling of several similar studies to produce an overall summary of treatment effect.\(^1\)\(^2\) Meta-analysis provides an objective and quantitative summary of evidence that is amenable to statistical analysis,\(^3\) and it allows recognition of important treatment effects by combining the results of small trials that individually might have lacked the power to consistently demonstrate differences among treatments. Meta-analysis has been criticized for the discrepancies between its findings and those of large clinical trials.\(^3\)\(^6\) The frequency of discrepancies ranges from 10% to 23%\(^3\) and can often be explained by differences in treatment protocols or study populations or changes that occur over time.\(^3\)

In conducting a meta-analysis, the authors should recognize the importance of having clear objectives, explicit criteria for study selection, an assessment of the quality of included studies, and prior consideration of which studies to combine. These items are the essential features of a systematic review. Meta-analyses that are not conducted within the context of a systematic review should be viewed with great caution.\(^7\)

A systematic review can be viewed as a scientific and systematic examination of the available evidence. A good systematic review will have explicitly stated objectives (the focused clinical question), materials (the relevant medical literature), and methods (the way studies are assessed and summarized). The steps taken during a systematic review are detailed in Table 1.

Not all systematic reviews and meta-analyses are equal. A systematic review should be conducted in a manner that will include all of the relevant trials, minimize the introduction of bias, and synthesize the results to be as truthful and useful to clinicians as possible. A systematic review can only be as good as the clinical trials that it contains. The criteria used to critically appraise systematic reviews and meta-analyses\(^8\) are listed in Table 2. In general, these criteria are similar to the criteria used to appraise the individual studies that make up the systematic review. Detailed explanations of each criterion are available.\(^1\)

**ASKING A CLEAR QUESTION**

The validity criteria are designed to ensure that the systematic review is conducted in a manner that minimizes the introduction of bias. Like the "well-built clinical question"\(^9\) for an individual study, a focused...
Clinical question for a systematic review should clearly articulate the following 4 elements of the material under review: (1) the patient, group of patients, or problem being evaluated; (2) the intervention; (3) comparison interventions; and (4) specific outcomes. The patient populations in the reviewed studies should be similar to the actual population most likely to benefit from the review results. The interventions studied should be those commonly available in practice. Outcomes reported should be those that are most relevant to physicians and patients.

**SOURCES OF EVIDENCE WITHIN A SYSTEMATIC REVIEW**

The overwhelming majority of systematic reviews involve therapy. Therefore, randomized, controlled, clinical trials should be used when available for systematic reviews of therapy because they are generally less susceptible to selection and information bias than other study designs. The quality of included trials is assessed using the criteria that are used to evaluate individual, randomized, controlled clinical trials. The quality criteria commonly used include concealed, random allocation; groups with similar known prognostic factors; equal treatment of groups; and inclusion of all trial patients in the results analysis (intent-to-treat design).

Randomized controlled trials are rarely a reliable source of identification of adverse reactions, unless those reactions are very common. Other evidence sources such as case-control studies, case reports, and postmarketing surveillance studies should therefore be examined. Systematic reviews of treatment efficacy should always include an assessment of common and serious adverse events to reach an informed and balanced decision about the utility of a treatment.

**THE HAZARDS OF “QUICK” SEARCHES**

A sound systematic review can be performed only if most or all of the available data are examined. Simply performing a quick MEDLINE search using “clinical trial” as publication type is rarely adequate because complex and sensitive search strategies are needed to identify all potential trials and because some clinical trials will be missed if they are published in a journal not listed by MEDLINE. Potential sources for finding studies about treatment include the Cochrane Controlled Trials Registry, which is part of the Cochrane Library; MEDLINE; EMBASE; bibliographies of studies; review articles and textbooks; symposia proceedings; pharmaceutical companies; and direct communication with experts in the field.

The Cochrane Controlled Trials Registry, the largest and most complete database of clinical trials in the world, includes more than 300,000 controlled clinical trials. It is compiled through several complex searches of the MEDLINE and EMBASE databases and by hand searching many journals, a process that is quality controlled and monitored by the Cochrane Collaboration in Oxford, England. Hand searching journals to identify controlled clinical trials and randomized, controlled clinical trials was undertaken because members of the Cochrane Collaboration noticed that many trials were incorrectly classified in the MEDLINE database. As an example, Adetugbo and Williams hand searched the Archives of Dermatology from 1990 through 1998 and identified 99 controlled clinical trials. Nineteen of these trials were not classified as controlled clinical trials in MEDLINE, and 11 trials that were not controlled clinical trials were misclassified as controlled clinical trials in MEDLINE.

MEDLINE is the National Library of Medicine’s bibliographic database covering medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. The MEDLINE file contains bibliographic citations and author abstracts from approximately 3900 current biomedical journals published in the United States and 70 foreign countries. The file contains approximately 9 million records dating back to 1966.

MEDLINE searches have inherent limitations that make their reliability less than ideal. For example, Spuls et al conducted a systematic review of systemic treatments of psoriasis. Treatments analyzed included UV-B, psoralen plus UV-A, methotrexate, cyclosporin A, and retinoids. To find relevant references, the authors used an exhaustive strategy that included searching MEDLINE, contacting pharmaceutical companies, polling leading authorities, reviewing abstract books of symposia and congresses, and reviewing textbooks, reviews, editorials, guideline articles, and the reference lists of all articles identified. Of 665 studies found, 356 (54%) were identified by MEDLINE search (30%-70% for different treatment modalities).

EMBASE is Excerpta Medica’s database covering drugs, pharmacology, and biomedical specialties. EMBASE has better coverage of European and non-English language
sources than MEDLINE and may be more up to date. The overlap in journals covered by MEDLINE and EMBASE is about 34% (10%-75%, depending on the subject).14,15

PUBLICATION BIAS

Publication bias (ie, the tendency of easy-to-locate studies to show more “positive” effects) is an important concern for systematic reviews, and a useful analysis of this issue can be found elsewhere.18 Publication bias results when issues other than the quality of the study are allowed to influence the decision to publish. Several studies have shown that factors such as sample size, direction and statistical significance of findings, and investigators’ perceptions of whether the findings are “interesting” are related to the likelihood of publication.17,18

Language bias may also be a problem; studies with positive findings are more likely to be published in an English-language journal and also more quickly than those with inconclusive or negative findings.19 A thorough systematic review should therefore include a search for high-quality unpublished trials and not restrict itself to journals written in English.

Studies with small samples are less likely to be published than those with larger samples, especially if they have negative results.17,18 This type of publication bias jeopardizes one of the main goals of meta-analysis: to increase power by pooling the results of numerous small studies. Creation of study registers and advance publication of research designs have been proposed as ways to prevent publication bias.20,21

Publication bias can be detected by using a simple graphic test (funnel plot) or by calculating the “fail-safe N.”22,23 These techniques are of limited value when fewer than 10 randomized controlled trials are included.

For many diseases, the studies published are dominated by drug company–sponsored trials of new expensive treatments. This bias in publication can result in data-driven systematic reviews that draw more attention to those medicines. In contrast, question-driven systematic reviews answer the sorts of clinical questions of most concern to practitioners. In many cases, studies that are of most relevance to doctors and patients have not been done in the field of dermatology owing to inadequate sources of independent funding.

Systematic reviews that have been sponsored directly or indirectly by industry are also prone to bias by overinclusion of unpublished studies with positive findings that are kept “on file” by that industry. Until it becomes mandatory to register all clinical trials conducted on human beings in a central registry and to make all of the results available in the public domain, all sorts of distortions due to selective withholding or release of data may occur.

Generally, reviews that have been conducted by volunteers in the Cochrane Collaboration are of better quality than non-Cochrane reviews.24 However, potentially serious errors have been noted in up to a third of Cochrane reviews.25

DATA ABSTRACTION

In general, the studies included in systematic reviews are reviewed by at least 2 reviewers. Data such as numbers of people entered into studies, numbers lost to follow-up, effects sizes, and quality criteria are recorded on predesigned data abstraction forms by at least 2 reviewers. Differences among reviewers are usually settled by consensus or by a third-person arbitrator. A systematic review in which there are large areas of disagreement among reviewers should lead the reader to question the validity of the review.

POOLING RESULTS

Results in the individual clinical trials that make up a systematic review may be similar in magnitude and direction (eg, they may all indicate that treatment A is superior to treatment B by a similar magnitude). Assuming that publication bias can be excluded, systematic reviews of studies with findings that are similar in magnitude and direction provide results that are most likely to be true and useful. It may be impossible to draw firm conclusions from systematic reviews of studies that have results of widely different magnitude and direction.

The magnitude of the difference between the treatment groups in achieving meaningful outcomes is the most useful summary result of a systematic review. The most easily understood measures of the magnitude of the treatment effect are the difference in response rate and its reciprocal, the number needed to treat (NNT).1,8,12 The NNT represents the number of patients one would need to treat to achieve 1 additional cure. Whereas the interpretation of NNT might be straightforward within a single trial, interpretation of NNT requires some caution within a systematic review because this statistic is highly sensitive to baseline event rates. For example, if treatment A is 30% more effective than treatment B for clearing psoriasis, and 50% of people who undergo treatment B are cleared with therapy, then 65% will clear with treatment A. These results correspond to a rate difference of 15% (65−50) and an NNT of 7 (1.00/0.15). This difference sounds quite worthwhile clinically. But if the baseline clearance rate for treatment B in another trial or setting is only 30%, the rate difference will be only 9%, and the NNT now becomes 11, and if the baseline clearance rate is 10%, then the NNT for treatment A will be 33, which is perhaps less worthwhile. In other words, it rarely makes sense to provide 1 NNT summary measure within a systematic review because “control” or baseline events rates usually differ considerably between studies owing to differences in study populations, interventions, and trial conditions.26 Instead, a range of NNTs for a range of plausible control event rates that occur in different clinical settings should be given, along with their 95% confidence intervals.

The precision of the estimate of the differences among treatments should be estimated. The confidence interval provides a useful measure of the precision of the treatment effect.1,8,12,27,28 The calculation and interpretation of confidence intervals has been extensively described.29 In simple terms, the reported result (known as the point estimate) provides the best estimate of the treatment effect. The population or “true” response to treatment will most likely lie near the middle of the confidence interval and will rarely be found at or near the ends of the interval. The population or true response to treatment has only a 1 in 20 chance of being outside the 95% confidence interval.
Certain conditions must be met when meta-analysis is performed to synthesize results from different trials. The trials should have conceptual homogeneity. They must involve similar patient populations, have used similar treatments, and have measured results in a similar fashion at a similar point in time. There are 2 main statistical models used to combine results: random effects models and fixed effects models. Random effects models assume that the different studies’ results may come from different populations with varying responses to treatment. Fixed effects models assume that each trial represents a random sample of a single population with a single response to treatment. In general, random effects models are more conservative (ie, less likely to show statistically significant results) than fixed effects models. When the combined studies have statistical homogeneity (ie, when the studies are reasonably similar), random effects and fixed effects models give similar results.

The key principle to keep in mind when considering combining results from several studies is that conceptual homogeneity precedes statistical homogeneity. In other words, results of several different studies should not be combined if it does not make sense to combine them (eg, if the patient groups or interventions studied are not sufficiently similar to each other). Although what constitutes "sufficiently similar" is a matter of judgment, the important thing is to explicitly articulate the decision to combine or not combine different studies. Tests for statistical heterogeneity are typically of very low power, so that statistical homogeneity does not mean clinical homogeneity. When there is evidence of heterogeneity, reasons for heterogeneity between studies such as different disease subgroups, intervention dosage, or study quality should be sought.

Sometimes, the robustness of an overall meta-analysis is tested further by means of a sensitivity analysis. In a sensitivity analysis the data are reanalyzed excluding those studies that are suspect because of quality or patient factors, to see whether their exclusion makes a substantial difference in the direction or magnitude of the main original results. In some systematic reviews in which a large number of trials have been performed, it is possible to evaluate whether certain subgroups (eg, children vs adults) are more likely to benefit than others. Subgroup analysis is rarely possible in dermatology because few trials are available.

The conclusions in the discussion section of a systematic review should closely reflect the data that have been presented within that review. The authors should make it clear which of the treatment recommendations are based on the review data and which reflect their own judgments. In addition to making clinical recommendations of therapies when evidence exists, many reviews in dermatology find little evidence to address the questions posed. This lack of conclusive evidence does not mean that the review is a waste of time, especially if the question addressed appears to be an important one. For example, the systematic review of antistreptococcal therapy for guttate psoriasis by Owen et al30 provided the authors with an opportunity to call for primary research in this area and to make recommendations on study design and outcomes that might help future researchers.

Applying evidence summarized in a systematic review to specific patients requires the same processes used to apply the results of individual controlled clinical trials to patients.

REFERENCES