

SUMMARIZING THE EVIDENCE

Andrew Oxman, Gordon Guyatt, Deborah Cook,
and Victor Montori

The following EBM Working Group members also made substantive contributions to this section: Rose Hatala, Ann McKibbin, Trisha Greenhalgh, Jonathan Craig, and Roman Jaeschke

IN THIS SECTION

Finding the Evidence

Are the Results Valid?

Did the Review Explicitly Address a Sensible Clinical Question?

Was the Search for Relevant Studies Detailed and Exhaustive?

Were the Primary Studies of High Methodologic Quality?

Were Assessments of Studies Reproducible?

What Are the Results?

Were the Results Similar From Study to Study?

What Are the Overall Results of the Review?

How Precise Were the Results?

How Can I Apply the Results to Patient Care?

How Can I Best Interpret the Results to Apply Them to the Care of Patients in My Practice?

Were All Clinically Important Outcomes Considered?

Are the Benefits Worth the Costs and Potential Risks?

Clinical Resolution



CLINICAL SCENARIO

Should We Offer Thrombolytic Drugs to Patients Presenting With Acute Thrombotic Stroke?

You are one of a group of neurologists working at an academic medical center. Your institution is not currently administering thrombolytic therapy to patients who present with acute thrombotic stroke. Some of your colleagues, convinced that thrombolysis will reduce ultimate mortality and morbidity in patients with acute thrombotic stroke, are enthusiastic about offering these patients tissue plasminogen activator (tPA) if they present within a few hours of symptom onset. Other members of your group are much more reluctant to initiate a policy of offering thrombolysis. You are undecided.

Your group has decided to address the issue formally. You join a subcommittee charged with collecting the evidence and generating an initial summary. The subcommittee decides to begin by looking for a systematic review.

FINDING THE EVIDENCE

You start by looking in the Cochrane Library 2000, Issue 1. You enter the terms “stroke” and “tissue plasminogen activator,” locate a relevant review in the Cochrane Database of Systematic Reviews, and find that the latest update was in July 1999.¹

For most of their questions, clinicians can find more than one relevant study. In the same way that it is important to use rigorous methods in primary research to protect against bias and random error, it is also important to use rigorous methods when summarizing the results of several studies. Traditional literature reviews, commonly found in journals and textbooks, typically provide an overview of a disease or condition. This overview may include a discussion of one or more aspects of disease etiology, diagnosis, prognosis, or management and will address a number of clinical, background, and theoretical questions.

For example, a review article or a chapter from a textbook on asthma might include sections on etiology, diagnosis, and prognosis and examine a wide variety of options for the treatment and prevention of asthma. Typically, authors of traditional reviews make little or no attempt to be systematic in the formulation of the questions they are addressing, the search for relevant evidence, or the summary of the evidence they consider. Medical students and clinicians looking for background information nevertheless often find these reviews very useful in obtaining a broad picture of a clinical condition or area of inquiry (see Part 1A1, “Finding the Evidence”).

Unfortunately, expert reviewers often make conflicting recommendations and their advice frequently lags behind or is inconsistent with the best available evidence.² One important reason for this phenomenon is the use of unsystematic

approaches to collecting and summarizing the evidence. Indeed, in one study, self-rated expertise was inversely related to the methodologic rigor of the review.³

In this section of the book, we focus on reviews that address specific clinical questions (eg, foreground information). Clinicians seeking to address focused management issues in providing patient care will find such reviews particularly useful (see Part 1A1, “Finding the Evidence”).

Authors sometimes use the terms overview, systematic review, and meta-analysis interchangeably. We use the term *overview* for any summary that attempts to address a focused clinical question, *systematic review* for any summary that attempts to address a focused clinical question using methods designed to reduce the likelihood of bias; and *meta-analysis* describes reviews that use quantitative methods to summarize the results. Investigators must make a host of decisions in preparing a systematic review, including determining the focus; identifying, selecting, and critically appraising the relevant studies (which we will call the *primary studies*); collecting and synthesizing (either quantitatively or nonquantitatively) the relevant information; and drawing conclusions. To avoid errors in systematic reviews requires an organized approach; enabling users to assess the validity of the results requires explicit reporting of the methods.

During the past decade, rapid expansion has occurred in the literature in terms of describing the methods used in systematic reviews, including studies that provide an empiric basis for guiding decisions about the methods used in summarizing evidence.⁴⁻⁶ Here, we emphasize key points from the perspective of a clinician needing to make a decision about patient care.

In applying the Users' Guides, you will find it useful to have a clear understanding of the process of conducting a systematic review. Figure 1E-1 demonstrates how the process begins with the definition of the question, which is synonymous with specifying selection criteria for deciding which studies to include in a review. These criteria define the population, the exposures or interventions, and the outcomes of interest (see Part 1A1, “Finding the Evidence”). A systematic review will also restrict the included studies to those that meet minimal methodologic standards. For example, systematic reviews that address a question of therapy will often include only randomized controlled trials.



FIGURE 1E-1

The Process of Conducting a Systematic Review

Define the Question

- Specify inclusion and exclusion criteria
 - Population
 - Intervention or exposure
 - Outcome
 - Methodology
- Establish a priori hypotheses to explain heterogeneity

Conduct Literature Search

- Decide on information sources: databases, experts, funding agencies, pharmaceutical companies, hand-searching, personal files, registries, citation lists of retrieved articles
- Determine restrictions: time frame, unpublished data, language
- Identify titles and abstracts

Apply Inclusion and Exclusion Criteria

- Apply inclusion and exclusion criteria to titles and abstracts
- Obtain full articles for eligible titles and abstracts
- Apply inclusion and exclusion criteria to full articles
- Select final eligible articles
- Assess agreement on study selection

Create Data Abstraction

- Data abstraction: participants, interventions, comparison interventions, study design
- Results
- Methodologic quality
- Assess agreement on validity assessment

Conduct Analysis

- Determine method for pooling of results
- Pool results (if appropriate)
- Decide on handling missing data
- Explore heterogeneity
 - Sensitivity and subgroup analysis
- Explore possibility of publications bias

Having specified their selection criteria, reviewers must conduct a comprehensive search that yields a large number of potentially relevant titles and abstracts. They then apply the selection criteria to the titles and abstracts, arriving at a smaller number of articles that they can retrieve. Once again, the reviewers apply the selection criteria, this time to the complete reports. Having completed the culling process, they assess the methodologic quality of the articles and abstract data from each study. Finally, they summarize the data, including, if appropriate, a quantitative synthesis or meta-analysis. The analysis includes an examination of differences among the included studies, an attempt to explain differences in results (exploring heterogeneity), a summary of the overall results, and an assessment of their precision and validity. Guidelines for assessing the validity of reviews and using the results correspond to this process (Table 1E-1).

TABLE 1E-1

Users' Guides for How to Use Review Articles

Are the results valid?

- Did the review explicitly address a sensible clinical question?
- Was the search for relevant studies detailed and exhaustive?
- Were the primary studies of high methodologic quality?
- Were assessments of studies reproducible?

What are the results?

- Were the results similar from study to study?
- What are the overall results of the review?
- How precise were the results?

How can I apply the results to patient care?

- How can I best interpret the results to apply them to the care of patients in my practice?
- Were all clinically important outcomes considered?
- Are the benefits worth the costs and potential risks?

ARE THE RESULTS VALID?

Did the Review Explicitly Address a Sensible Clinical Question?

Consider a systematic review that pooled results from all cancer therapeutic modalities for all types of cancer to generate a single estimate of the impact on mortality. Next, consider a review that pooled results of the effects in patients suffering from clinically manifest atherosclerosis (whether in the heart, head, or lower extremities) of all doses of all antiplatelet agents (including aspirin, sulfipyrazone, and dipyridamole) on major thrombotic events (including myocardial



infarctions, strokes, and acute arterial insufficiency in the leg) and mortality. Finally, reflect on a review that addressed the impact of a wide range of aspirin doses to prevent thrombotic stroke in patients who had experienced a transient ischemic attack (TIA) in the carotid circulation.

Clinicians would not find the first of these reviews useful; they would conclude it is too broad. Most clinicians are uncomfortable with the second question, still considering it excessively broad. For this second question, however, a highly credible and experienced group of investigators found the question reasonable and published the results of their meta-analysis in a leading journal.⁷⁻⁹ Most clinicians are comfortable with the third question, although some express concerns about pooling across a wide range of aspirin doses.

What makes a systematic review too broad or too narrow? Elsewhere in this book, we have argued that identifying the population, the interventions or exposures, and the outcomes of interest is a useful way of structuring a clinical question (see Part 1A1, "Finding the Evidence"). When deciding if the question posed in the review is sensible, clinicians need to ask themselves whether the underlying biology is such that they would expect; that is, the same treatment effect across the range of patients. They should ask the parallel question about the other components of the study question. For example, is the underlying biology such that, across the range of interventions and outcomes included, they expect more or less the same treatment effect? Clinicians can also construct a similar set of questions for other areas of clinical inquiry. For example, across the range of patients, ways of testing, and criterion or gold standard for diagnosis, does one expect more or less the same likelihood ratios associated with studies examining a diagnostic test (see Part 1C2, "Diagnostic Tests")?¹⁰

The reason that clinicians reject a systematic review that pools across all modes of cancer therapy for all types of cancer is that they know that some cancer treatments are effective in certain cancers, whereas others are harmful. Combining the results of these studies would yield a meaningless estimate of effect that would not be applicable to any of the interventions.

Clinicians who reject the second review might also argue that the biologic variation in antiplatelet agents is likely to lead to important differences in treatment effect. Further, they may contend that there are important differences in the biology of atherosclerosis in the vessels of the heart, head, and legs. Moreover, because clinicians need to make specific decisions about specific patients, they may be inclined to seek a summary of the evidence for the intervention they are considering in patients who most resemble the patient before them.

Those who would endorse the second review would argue the similar underlying biology of antiplatelet agents—and atherosclerosis in different parts of the body—and thus anticipate a similar magnitude of treatment effects. Moreover, they would point out that the best estimate of effect for an individual patient will often come from a broader review rather than a narrower one. There are three reasons for this.

First, focusing on a narrow group of patients (eg, the most severe, or least severe), interventions (such as a single aspirin dose in our cerebrovascular disease

example), or studies (eg, those only in the English language)—in each case, a subgroup of those one might have chosen—increases the risk of chance producing a spurious result.^{11,12} Second, focusing on a subgroup introduces a risk of false conclusions owing to bias, if the criterion used to select the subgroup is confounded with another determinant of treatment effect. For example, a reviewer may select studies based on the type of patient even though the quality of studies of those patients is methodologically weaker than other studies, resulting in a spurious overestimate of treatment effect. Third, review of all potentially relevant data facilitates exploration of the possible explanations for variability in study results—the patients, the interventions, and the ways of measuring outcome. Thus, a broadly focused review provides a better basis for estimating the effect of a specific agent for a specific manifestation; it also provides a better basis for determining whether to believe a subgroup analysis, rather than a narrowly focused review that risks an inappropriate subgroup analysis (see Part 2E, “Summarizing the Evidence, When to Believe a Subgroup Analysis”).

Turning to the third question, most clinicians would accept that the biology of aspirin action is likely to be similar in patients whose TIA reflected right-sided or left-sided brain ischemia, in patients older than 75 years and in younger patients; in men and women, across doses, over periods of follow-up ranging from 1 to 5 years, and in patients with stroke who have been identified by the attending physician and those identified by a team of expert reviewers. The similar biology is likely to result in a similar magnitude of treatment effect. Nonetheless, even within this more narrowly focused question, there is still variation in the types of patients and the types of interventions, as well as possible differences in the types of outcome measures and methods of the included studies. Thus, there will still be a need to examine possible sources of variation in the results (see Part 2E, “Summarizing the Evidence, Evaluating Differences in Study Results”). As a result, the question about whether it is sensible to pool across studies cannot, in general, be resolved until one has looked at the results. If the effect was similar across studies, the results support pooling; if not, they raise questions about any inferences one can make from the pooled results.

The task of the clinician, then, is to decide whether, across the range of patients, interventions or exposures, and outcomes, it is plausible that the intervention will have a similar impact. Doing so requires a precise statement of what range of patients, exposures, and outcomes the reviewer has decided to consider; in other words, explicit selection criteria for studies included in the review are necessary. In addition, criteria are necessary that specify what types of studies were considered relevant. Generally these should be similar to the primary validity criteria we have described for original reports of research in other parts of this book (see Table 1E-2). Explicit eligibility criteria not only facilitate the user’s decision regarding whether the question was sensible, but also make it less likely that the authors will preferentially include studies that support their own prior conclusions.



TABLE 1E-2

Guides for Selecting Articles That Are Most Likely to Provide Valid Results³

Therapy	<ul style="list-style-type: none"> • Were patients randomized? • Was follow-up complete?
Diagnosis	<ul style="list-style-type: none"> • Was the patient sample representative of those with the disorder? • Was the diagnosis verified using credible criteria that were independent of the clinical manifestations under study?
Harm	<ul style="list-style-type: none"> • Did the investigators demonstrate similarity in all known determinants of outcome, or adjust for differences in the analysis? • Was follow-up sufficiently complete?
Prognosis	<ul style="list-style-type: none"> • Was there a representative and well-defined sample of patients at a similar point in the course of disease? • Was follow-up sufficiently complete?

Bias in choosing articles to cite is a problem for both systematic reviews and original reports of research (in which the discussion section often includes comparisons with the results of other studies). Gøtzsche, for example, reviewed citations in reports of trials of new nonsteroidal anti-inflammatory drugs in rheumatoid arthritis.¹³ Among 77 articles in which the authors could have referenced other trials with and without outcomes favoring the new drug, nearly 60% (44) cited a higher proportion of the trials with favorable outcomes. In 22 reports of controlled trials of cholesterol lowering, Ravnskov found a similar bias toward citing positive studies.¹⁴ In 26 reports of RCTs in general medical journals, Clarke and Chalmers found only two articles in which the results were discussed in the context of an updated systematic review.¹⁵ Users should exercise caution when interpreting the results of a study outside of the context of a systematic review.

Was the Search for Relevant Studies Detailed and Exhaustive?

Authors of a systematic review should conduct a thorough search for studies that meet their inclusion criteria. Their search should include the use of bibliographic databases, such as MEDLINE and EMBASE, the Cochrane Controlled Trials Register (containing more than 250,000 RCTs), and databases of current research.¹⁶ They should check the reference lists of the articles they retrieve, and they should seek personal contact with experts in the area. It may also be important to examine recently published abstracts presented at scientific meetings and to look at less frequently used databases, including those that summarize doctoral theses and databases of ongoing trials held by pharmaceutical companies. Listing these sources, it becomes evident that a MEDLINE search alone will not be satisfactory. Unless the authors tell us what they did to locate relevant studies, it is difficult to know how likely it is that relevant studies were missed.

There are two reasons that reviewers should contact experts in the area under consideration. The first is to identify published studies that may have been missed

(including studies that are labeled “in press” and those that have not yet been indexed or referenced). The second is to identify unpublished studies and to include them to avoid publication bias.

Publication bias occurs when the publication of research depends on the direction of the study results and whether they are statistically significant. Studies in which an intervention is not found to be effective sometimes are not published. Because of this, systematic reviews that fail to include unpublished studies may overestimate the true effect of an intervention.¹⁷⁻²¹ (See Part 2E, “Summarizing the Evidence, Publication Bias.”)

If investigators include unpublished studies in a review, they should obtain full written reports and they should appraise the validity of both published and unpublished studies. Reviewers may also use statistical techniques to explore the possibility of publication bias and other reporting biases, although the power of these techniques to detect bias is limited.²² Systematic reviews based on a small number of studies with small sample sizes are the most susceptible to *publication bias*, and users should be cautious about drawing conclusions in such cases. Results that seem too good to be true may well not be true.

Reviewers may go even farther than simply contacting the authors of primary studies. They may recruit these investigators as collaborators in their review, and in the process they may obtain individual patient records. Access to individual patient records facilitates powerful analysis and strengthens the inferences from a systematic review.

Were the Primary Studies of High Methodologic Quality?

Even if a review article includes only RCTs, knowing whether they were of good quality is important. Unfortunately, peer review does not guarantee the validity of published research (see Part 1B1, “Therapy”).²³ For exactly the same reason that the guides for using original reports of research begin by asking if the results are valid, it is essential to consider the validity of primary articles in systematic reviews.

Differences in study methods might explain important differences among the results.²⁴⁻²⁶ For example, less rigorous studies tend to overestimate the effectiveness of therapeutic and preventive interventions.²⁷ Even if the results of different studies are consistent, determining their validity still is important. Consistent results are less compelling if they come from weak studies than if they come from strong studies.

Consistent results from observational studies are particularly suspect. Physicians may systematically select patients with a good prognosis to receive therapy, and this pattern of practice may be consistent over time and geographic setting. Observational studies summarized in a systematic review,²⁸ for instance, have consistently shown average relative risk reductions in major cardiovascular events of about 50% with hormone replacement therapy. The only large RCT addressing this issue found no effect of hormone replacement therapy on cardiovascular risk.²⁹

There is no one correct way to assess the quality of studies, although in the context of a systematic review the focus should be on validity and users should be cautious about the use of scales to assess the quality of studies.^{30,31} Some investigators



use long checklists to evaluate methodologic quality, whereas others focus on three or four key aspects of the study. When considering whether to trust the results of a review, check to see whether the authors examined criteria similar to those we have presented in other sections of this book (see Part 1B1, “Therapy”; Part 1C, “The Process of Diagnosis”; Part 1B2, “Harm”; and Part 1D, “Prognosis”). Reviewers should apply these criteria both in selecting studies for inclusion and in assessing the validity of the included studies (see Figure 1E-1 and Table 1E-2).

Were Assessments of Studies Reproducible?

As we have seen, authors of systematic review articles must decide which studies to include, how valid they are, and what data to extract. These decisions require judgment by the reviewers and are subject to both mistakes (ie, random errors) and bias (ie, systematic errors). Having two or more people participate in each decision guards against errors; if there is good agreement beyond chance between the reviewers, the clinician can have more confidence in the results of the systematic review (see Part 2C, “Diagnosis, Measuring Agreement Beyond Chance”).

USING THE GUIDE

Returning to our opening scenario, the Cochrane review you located included trials enrolling patients with acute ischemic stroke in whom CT excluded hemorrhage.¹ These patients were randomized to receive or not receive thrombolytic therapy, and an intention-to-treat analysis had been or could be conducted (see Part 2B1, “Therapy and Validity, The Principle of Intention-to-Treat”). (An intention-to-treat analysis examines outcomes for study participants based on the treatment arm to which they were originally randomized rather than the treatment they actually received.) You are concerned that the impact of treatment might differ substantially in patients who present early or late, in those with major or minor deficits, in those who received different thrombolytic agents, and in studies with different ways of measuring functional status or different durations of follow-up. Nevertheless, you are uncertain about the extent to which these variables might affect outcome, and you suspect that combining results across all patients, interventions, and outcomes might prove informative.

The reviewers searched the Cochrane Registry of Controlled Trials and EMBASE. In addition, they hand-searched a number of Japanese-language journals; contacted 321 pharmaceutical companies; contacted principal investigators in Europe, the United States, Japan, and China; attended a number of international stroke treatment symposia; and searched references quoted in the articles they found. It is likely they obtained all the relevant trials.

Of the 17 trials included, seven used centralized randomization of patients to treatment or control groups to ensure concealment. In 13 studies, participants

and health care personnel were then blinded to allocation by using sealed, prepacked, and identical-looking thrombolytic and placebo infusions. Because of bleeding complications of thrombolytic therapy, blinding of participants and health care personnel may be difficult to ensure, underscoring the importance of blinding the outcome assessors; long-term outcome assessors were blinded to allocation in only four of the studies. The reviewers do not report on the proportion of patients lost to follow-up in any trial.

One of the review's authors decided whether potentially eligible trials met inclusion criteria. A different author extracted the data but then verified them with the principal investigators and corrected any errors. In 10 trials, the authors of the systematic review were able to obtain scores on a measure of functional status, the Rankin instrument, on individual patients. Scores of up to two out of five on this functional status measurement instrument indicate that patients are still able to look after themselves,³² so the investigators classified scores of three to five on this instrument as characterizing a poor outcome. In another two trials for which they could not obtain individual data, scores of two or greater represented a poor outcome.

Overall, the methods of the systematic review—and the methodologic quality of the trials included in the systematic review—were strong.

WHAT ARE THE RESULTS?

Were the Results Similar From Study to Study?

Most systematic reviews document important differences in patients, exposures, outcome measures, and research methods from study to study. As a result, the most common answer to the initial question about whether we can expect similar results across the range of patients, interventions, and outcomes is “perhaps.”

Fortunately, one can resolve this unsatisfactory situation. Having completed the review, investigators should present the results in a way that allows clinicians to check the validity of the initial assumption. That is, did results prove similar from study to study?

There are two things to consider when deciding whether the results are sufficiently similar to warrant making a single estimate of treatment effects that applies across the populations, interventions, and outcomes studied (see Part 2E, “Summarizing the Evidence, Evaluating Differences in Study Results”). First, how similar are the best estimates of the treatment effect (that is, the *point estimates*) from the individual studies? The more different they are, the more clinicians should question the decision to pool results across studies.

Second, to what extent are differences among the results of individual studies greater than you would expect by chance? Users can make an initial assessment by



examining the extent to which the confidence intervals overlap. The greater the overlap, the more comfortable one is with pooling results. Widely separated confidence intervals flag the presence of important variability in results that requires explanation (see Part 2E, “Summarizing the Evidence, Evaluating Differences in Study Results”).

Clinicians can also look to formal statistical analyses called tests of heterogeneity, which assess the degree of difference or variance among samples, groups, or populations. When the *P* value associated with the test of heterogeneity is small (eg, $< .05$), chance becomes an unlikely explanation for the observed differences in the size of the effect (see Part 2B2, “Therapy, Hypothesis Testing”). Unfortunately, a higher *P* value (.1, or even .3) does not necessarily rule out important heterogeneity. The reason is that, when the number of studies and their sample sizes are both small, the test of heterogeneity is not very powerful. Hence, large differences between the apparent magnitude of the treatment effect between studies—that is, the point estimates—dictates caution in interpreting the overall findings, even in the face of a nonsignificant test of homogeneity. Conversely, if the differences in results across studies are not clinically important, then heterogeneity is of little concern, even if it is statistically significant (see Part 2E, “Summarizing the Evidence, Evaluating Differences in Study Results”).

Reviewers should try to explain between-study variability in findings. Possible explanations include differences between patients (eg, thrombolytic therapy in acute myocardial infarction may be much more effective in patients who present shortly after the onset of chest pain than those who present much later), between interventions (eg, tPA may have a larger treatment effect than streptokinase), between outcome measurement (eg, the effect may differ if the outcome is measured at 30 days rather than at 1 year after myocardial infarction), or methodology (eg, the effect may be smaller in blinded trials or in those with more complete follow-up). Although appropriate and, indeed, necessary, this search for explanations of heterogeneity in study results may be misleading (see Part 2E, “Summarizing the Evidence, When to Believe a Subgroup Analysis”). Furthermore, how is the clinician to deal with residual heterogeneity in study results that remains unexplained? We will deal with this issue in our discussion of the applicability of the study results.

What Are the Overall Results of the Review?

In clinical research, investigators collect data from individual patients. Because of the limited capacity of the human mind to handle large amounts of data, investigators use statistical methods to summarize and analyze them. In systematic reviews, investigators collect data from individual studies. Investigators must also summarize these data and, increasingly, they are relying on quantitative methods to do so.

Simply comparing the number of positive studies to the number of negative studies is not an adequate way to summarize the results. With this sort of “vote counting,” large and small studies are given equal weight, and (unlikely as it may seem) one investigator may interpret a study as positive, whereas another investigator may interpret the same study as negative.³³ For example, a clinically

important effect that is not statistically significant could be interpreted as positive in light of clinical importance and negative in light of statistical significance. There is a tendency to overlook small but important effects if studies with statistically nonsignificant (but potentially clinically important) results are counted as negative.³⁴ Moreover, a reader cannot tell anything about the magnitude of an effect from a vote count even when studies are appropriately classified using additional categories for studies with a positive or negative trend.

Typically, meta-analysts weight studies according to their size, with larger studies receiving more weight. Thus, the overall results represent a weighted average of the results of the individual studies (see Part 2E, “Summarizing the Evidence, Fixed-Effects and Random-Effects Models”). Occasionally studies are also given more or less weight depending on their quality, or poorer-quality studies might be given a weight of zero (excluded) either in the primary analysis or in a secondary analysis that tests the extent to which different assumptions lead to different results (a sensitivity analysis).

You should look to the overall results of a systematic review the same way you look to the results of primary studies. In a systematic review of a therapeutic question, you should look for the relative risk and relative risk reduction or the odds ratio (see Part 2B2, “Therapy and Understanding the Results, Measures of Association”). In systematic reviews regarding diagnosis, you should look for summary estimates of the likelihood ratios (see Part 1C2, “Diagnostic Tests”).

Sometimes the outcome measures that investigators have used in different studies are similar but not identical. For example, different trials might measure functional status using different instruments. If the patients and the interventions are reasonably similar, estimating the average effect of the intervention on functional status still might be worthwhile. One way of doing this is to summarize the results of each study as an effect size.³⁵ The *effect size* is the difference in outcomes between the intervention and control groups divided by the standard deviation. The effect size summarizes the results of each study in terms of the number of standard deviations of difference between the intervention and control groups. Investigators can then calculate a weighted average of effect sizes from studies that measured a given outcome in different ways.

You may find it difficult to interpret the clinical importance of an effect size. For example, if the weighted average effect is one half of a standard deviation, is this effect clinically trivial or is it large? Once again, you should look for a presentation of the results that conveys their practical importance (eg, by translating the summary effect size back into natural units³⁶). For instance, clinicians may have become familiar with the significance of differences in walk test scores in patients with chronic lung disease. Investigators can then convert the effect size of a treatment on a number of measures of functional status (eg, the walk test and stair climbing) back into differences in walk test scores.³⁷

Although it is generally desirable to have a quantitative summary of the results of a review, it is not always appropriate. When quantitative summaries are inappropriate, investigators should still present tables or graphs that summarize the results of the primary studies.



How Precise Were the Results?

In the same way that it is possible to estimate the average effect across studies, it is possible to estimate a confidence interval around that estimate, that is, a range of values with a specified probability (typically 95%) of including the true effect (see Part 2B2, “Therapy and Understanding the Results, Confidence Intervals”).

USING THE GUIDE

Returning to our opening scenario, four trials used streptokinase, three trials used urokinase, two used Pro-Urokinase, and eight used tPA. Data from six trials for death during the first 7 to 10 days showed that 16.6% of those receiving thrombolytic agents and 9.8% of the control patients died (OR, 1.85; 95% CI, 1.48-2.32). The *P* value for the test of heterogeneity showed borderline significance with the value for the tPA trials being lower and non-significant (OR, 1.24; 95% CI, 0.85-1.81). Considering data from 11 trials, investigators found that thrombolytic therapy increased fatal intracranial hemorrhage from 1.0% to 5.4% (OR, 4.15; 95% CI, 2.96-5.84), and the results were consistent across studies.

The final assessment of outcome (at 1 month in six trials, 3 months in nine trials, and 6 months in two trials) showed an increase in deaths from 15.9% to 19% (OR, 1.31; 95% CI, 1.13- 1.52). The results showed considerable heterogeneity ($P < .01$).

Thrombolysis reduced the combined endpoint of death and dependency (55.2% in patients receiving thrombolysis and 59.7% in those allocated to the control group (OR, 0.83; 95% CI, 0.73-0.94). The results were consistent across the trials.

The authors explored possible sources of heterogeneity for differences in death rate. Despite large differences in point estimates (urokinase OR, 0.71; streptokinase OR, 1.43; tPA OR, 1.16), differences among drugs failed to reach statistical significance. Death rate was increased when streptokinase and aspirin were given together in comparison to streptokinase alone. The authors failed to find a relationship between control event rate and mortality, though they note that individual data would be required to properly explore the relationship between stroke severity and thrombolytic benefit and harm. Trials in which some patients were randomized within 3 hours and some were randomized after 3 hours showed no difference in deaths between the two groups.

HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

How Can I Best Interpret the Results to Apply Them to the Care of Patients in My Practice?

Even if the true underlying effect is identical in each of a set of studies, chance will ensure that the observed results differ (see Part 2B, “Therapy and Harm, Why Study Results Mislead: Bias and Random Error”). As a result, systematic reviews risk capitalizing on the play of chance. Perhaps the studies with older patients happened, by chance, to be those with the smaller treatment effects. The reviewer may erroneously conclude that the treatment is less effective in elderly patients. The more subgroup analyses the reviewer undertakes, the greater is the risk of a spurious conclusion.

The clinician can apply a number of criteria to distinguish subgroup analyses that are credible from those that are not (see Part 2E, “Summarizing the Evidence, When to Believe a Subgroup Analysis”). Criteria that make a hypothesized difference in subgroups more credible include the following: conclusions drawn on the basis of within-study rather than between-study comparisons; a large difference in treatment effect across subgroups; a highly statistically significant difference in treatment effect (eg, the lower the *P* value on the comparison of the different effect sizes in the subgroups, the more credible the difference); a hypothesis that was made before the study began and that was one of only a few that were tested; consistency across studies; and indirect evidence in support of the difference (eg, “biologic plausibility”). If these criteria are not met, the results of a subgroup analysis are less likely to be trustworthy and you should assume that the overall effect across all patients and all treatments, rather than the subgroup effect, applies to the patient at hand and to the treatment under consideration.

What are clinicians to do if subgroup analyses fail to provide an adequate explanation for unexplained heterogeneity in study results? Although a number of reasonable possibilities exist, including not to pool findings at all, we suggest that, pending further trials that may explain the differences, clinicians should look to a summary measure from all of the best available studies for the best estimate of the impact of the intervention or exposure.³⁸⁻⁴⁰

Were All Clinically Important Outcomes Considered?

Although it is a good idea to look for focused review articles because they are more likely to provide valid results, this does not mean that you should ignore outcomes that are not included in a review. For example, the potential benefits of hormone replacement therapy include a reduced risk of fractures and a reduced risk of coronary heart disease, and potential downsides include an increased risk of breast cancer and endometrial cancer. Focused reviews of the evidence are more likely to provide valid results of the impact of hormone replacement therapy on each one of these four outcomes, but a clinical decision requires considering all of them.



Systematic reviews frequently do not report the adverse effects of therapy. One reason is that the individual studies often measure these adverse effects either in different ways or not at all, making pooling, or even effective summarization, difficult. Costs are an additional outcome that you will often find absent from systematic reviews.

Are the Benefits Worth the Costs and Potential Risks?

Finally, either explicitly or implicitly, the clinician and patient must weigh the expected benefits against the costs and potential risks (see Part 1F, “Moving From Evidence to Action”). Although this is most obvious for deciding whether to use a therapeutic intervention or a preventive one, providing patients with information about causes of disease or prognosis also can have both benefits and risks. For example, informing city dwellers about the health risks of air pollution exposures might result in their reducing their risk of exposure, with potential benefits; however, it might also cause anxiety or make their lives less convenient. Informing an asymptomatic woman with newly detected cancer about her prognosis might help her to plan better, but it might also label her, cause anxiety, or increase the period during which she is “sick.”

A valid review article provides the best possible basis for quantifying the expected outcomes, but these outcomes still must be considered in the context of your patient’s values and concerns about the expected outcomes of a decision. Ultimately, trading off benefits and risks will involve value judgments (see Part 1F, “Moving From Evidence to Action”), and in individual decision making, these values should come from the patient (see Part 2F, “Moving From Evidence to Action, Incorporating Patient Values”).

CLINICAL RESOLUTION

Returning to the opening scenario, the committee decides it can confidently reach two conclusions on the basis of the systematic review. First, thrombolytic therapy increases the odds of intracranial hemorrhage by a factor of between approximately 3 and 6, with the best estimate being approximately 4. In absolute terms, thrombolytic therapy will cause one intracranial hemorrhage for every 23 patients who are treated. Second, thrombolytic therapy reduces the odds of the combined outcome of death and dependency after approximately 3 months by approximately 5% to 30%, the best estimate being an OR of 0.83 (17%). In absolute terms, 22 patients need to be treated to prevent one patient from dying or becoming seriously dependent after 3 months. A third conclusion also seems likely: the concomitant administration of aspirin increases the risk of intracranial hemorrhage.

The committee concludes that many areas of uncertainty remain. They include questions about whether the risk of death during the 3-month period after stroke

is lower for tPA than the combined estimate suggests, as well as the relative effect on both hemorrhage and death and disability, according to the severity and nature of symptoms at initial presentation. Given the extent and nature of the uncertainties, the committee agrees that administration of thrombolytic therapy should be restricted to highly selected patients who are ready to risk an increase in the likelihood of early death to achieve a subsequent reduction in morbidity.

References

1. Wardlaw JM, del Zoppo G, Yamaguchi T. Thrombolysis for acute ischaemic stroke. *Cochrane Database Syst Rev.* 2000;2:CD000213.
2. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA.* 1992;268:240-248.
3. Oxman AD, Guyatt GH. The science of reviewing research. *Ann N Y Acad Sci.* 1993;703:125-133; discussion 133-134.
4. Clarke M, Olsen KL, Oxman AD, eds. The Cochrane Review Methodology Database. In: *The Cochrane Library.* Oxford: Update Software; 2000, issue 1.
5. Clarke M, Oxman AD, eds. Cochrane Reviewers' Handbook 4.0 [updated July 1999]. In: *The Cochrane Library.* Oxford: Update Software; 2000, issue 1.
6. Egger M, Davey Smith G, Altman DG, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context.* 2nd ed. London: BMJ Books; 2000.
7. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ.* 1994;308:81-106.
8. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, II: maintenance of vascular graft or arterial patency by antiplatelet therapy. *BMJ.* 1994;308:159-168.
9. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, III: reduction in venous thrombosis and pulmonary embolism by antiplatelet prophylaxis among surgical and medical patients. *BMJ.* 1994;308:235-246.
10. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med.* 1994;120:667-676.
11. Counsell CE, Clarke MJ, Slattery J, Sandercock PA. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *BMJ.* 1994;309:1677-1681.



12. Clarke MJ, Halsey J. D.I.C.E. 3: the need for cautious interpretation of meta-analyses. Paper presented at: First Symposium on Systematic Reviews: Beyond the Basics; January 1998; Oxford.
13. Gøtzsche PC. Reference bias in reports of drug trials. *Br Med J (Clin Res Ed)*. 1987;295:654-656.
14. Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ*. 1992;305:15-19.
15. Clarke M, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? *JAMA*. 1998;280:280-282.
16. The *meta*Register of Controlled Trials (*mRCT*). Current Controlled Trials. Available at: www.controlled-trials.com/. Accessed January 31, 2001.
17. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990;263:1385-1389.
18. Dickersin K, Min Y, Meinert CL. Factors influencing publication of research results. *JAMA*. 1992;267:374-378.
19. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev*. 1997;9(suppl 1):15-21.
20. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315:640-645.
21. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279:281-286.
22. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315:629-634.
23. Williamson JW, Goldschmidt PG, Colton T. The quality of medical literature: analysis of validation assessments. In: Bailar JC, Mosteller F, eds. *Medical Uses of Statistics*. 2nd ed. Waltham: NEJM Books; 1992:370-391.
24. Horwitz RI. Complexity and contradiction in clinical trial research. *Am J Med*. 1987;82:498-510.
25. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol*. 1992;45:255-265.
26. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352:609-613.
27. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317:1185-1190.

28. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med.* 1991;20:47-63.
29. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA.* 1998;280:605-613.
30. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials.* 1995;16:62-73.
31. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282:1054-1060.
32. de Haan R, Limburg M, Bossuyt P, van der Meulen J, Aaronson N. The clinical meaning of Rankin 'handicap' grades after stroke. *Stroke.* 1995;26:2027-2030.
33. Glass GV, McGaw B, Smith ML. *Meta-analysis in Social Research.* Beverly Hills: Sage Publications; 1981:18-20.
34. Cooper HM, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. *Psychol Bull.* 1980;87:442-449.
35. Rosenthal R. *Meta-analytic Procedures for Social Research.* 2nd ed. Newbury Park: Sage Publications; 1991.
36. Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis.* 1992;145:533-539.
37. Lacasse Y, Wong E, Guyatt GH, King D, Cook DJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronic obstructive pulmonary disease. *Lancet.* 1996;348:1115-1119.
38. Peto R. Why do we need systematic overviews of randomized trials? *Stat Med.* 1987;6:233-244.
39. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med.* 1992;116:78-84.
40. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA.* 1991;266:93-98.