Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews

Agency for Healthcare Research and Quality. *Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews*, Version 1.0 [Draft posted Oct. 2007]. Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf

Effectiveness and Comparative Effectiveness Reviews, systematic reviews of existing research on the effectiveness, comparative effectiveness, and comparative harms of different health care interventions, are intended to provide relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. In an effort to improve the transparency, consistency, and scientific rigor of the work of the Effective Health Care Program, through a collaborative effort, the Agency for Healthcare Research and Quality (AHRQ), the Scientific Resource Center, and the Evidence-based Practice Centers (EPCs) have developed a Methods Guide for the conduct of Comparative Effectiveness Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting Comparative Effectiveness Reviews.

The first draft of the Methods Guide was posted for public comment for 8 weeks in late 2007. In response to requests from investigators and others interested in Comparative Effectiveness Review methods, we have reposted the original chapters of the draft manual below. As these chapters are revised in response to public and peer review comment, they will replace the previous draft chapter and be posted below. It is anticipated that these papers will also be published as a series in the *Journal of Clinical Epidemiology* in 2008. As further empiric evidence develops and our understanding of better methods improves, we anticipate that there will be subsequent updates and additional chapters to this Methods Guide and that it will continue to be a living document. Comments and suggestions on the Methods Guide and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

Preface

As part of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003, Congress directed the Agency for Healthcare Research and Quality (AHRQ) to conduct and support research on the evidence of outcomes, clinical effectiveness, and appropriateness of pharmaceuticals, devices, and health care services to meet the needs of Medicare, Medicaid, and the State Children's Health Insurance Program (SCHIP). Section 1013 of the Act requires AHRQ to conduct activities pertinent to evaluating, generating, and disseminating evidence about the comparative effectiveness of medications, devices, and other interventions. These activities include, but are not limited to, the following:

- identify priorities for research related to health care items and services, including prescription drugs;
- evaluate and synthesize evidence about comparative clinical effectiveness related to these priorities;
- identify key information gaps for future research; and
- disseminate the results of comparative effectiveness reviews (CERs) to the public, to Medicare Advantage plans, and to other health plans.

These and related activities constitute AHRQ's Effective Health Care (EHC) program, which is described in full at http://effectivehealthcare.ahrq.gov/aboutUs.cfm?abouttype=program.

AHRQ has an established network of Evidence-based Practice Centers (EPCs) that produce Evidence Reports/Technology Assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care. The EPCs lend their expertise to the Effective Health Care program by conducting CERs of medications, devices, and other relevant interventions, including strategies for how these items and services can best be organized, managed, and delivered.

Systematic and comparative effectiveness reviews are the building blocks underlying evidence-based practice; they focus attention on the strengths and limitations of evidence from research studies about the effectiveness and safety of a clinical intervention. In the context of developing recommendations for practice, these reviews are useful because they define the strong and weak points of the evidence, and they clarify whether assertions about the value of the intervention are based on robust evidence from clinical studies.

AHRQ expects that CERs will be helpful to patients, clinicians, health plans, purchasers, government programs, researchers, and the health care system as a whole. CERs are not intended to set national standards of clinical practice or criteria for health care quality standards. Instead, as Congress stated in the MMA, research conducted in the EHC program "should reflect the principle that clinicians and patients should have the best available evidence upon which to make choices…recognizing that patient subpopulations and patient and physician preferences may vary." AHRQ is committed to presenting information in different formats so that consumers who make decisions about their own and their family's health can benefit from the evidence.

TABLE OF CONTENTS

1. OVERVIEW

This guide provides methodological guidance to Evidence-based Practice Centers (EPCs) conducting comparative effectiveness reviews (CERs). It describes recommended approaches for addressing difficult, frequently encountered methodological issues. It informs the public of standards for conducting CERs. Finally, it identifies areas of methodological controversy for which, at present, no standard can be recommended; these will be addressed in future work by EPCs, in updates to this guide, or by others.

Purpose of This Guide

AHRQ and the participating EPCs are fully committed to improving the consistency and quality of CERs. The science of systematic reviews is evolving and dynamic, and thus the recommendations in this guide should be viewed as a work in progress. At the same time, we recognize that excessive variation in methods among systematic reviews gives the appearance of arbitrariness and idiosyncrasy, which undercuts the goals of transparency and scientific impartiality intended for all this work.

The guide is organized around key issues at each step involved in researching and writing a CER, including the following:

- developing key questions for a CER (Chapter 2),
- selecting different types of evidence (Chapters 3 and 4),
- searching for relevant trials and observational studies (Chapter 5),
- assessing the risk of bias (quality) and applicability of studies (Chapter 6),
- when and how to pool studies (Chapter 9), and
- rating the strength of a body of evidence (Chapter 11).

Some topics, such as eligibility criteria, extracting evidence from studies, and constructing evidence tables, are not discussed in this version of the guide.

This guide is not aimed at beginners, and it is not a comprehensive source of guidance for conducting systematic reviews in all circumstances. Nevertheless, we hope that a wide range of users of CERs and those with a broad interest in evidence-based practice will find it a useful reference and sourcebook.

Comparative Effectiveness Reviews

Comparative Effectiveness Reviews (CERs) are a key component of the EHC program. They provide building blocks to support evidence-based practice and decision making. They seek to answer important questions about treatments or diagnostic tests to help clinicians and patients choose the best treatments and tests and to help healthcare policy makers make informed decisions about health care services and quality improvement.

CERs are a type of systematic review, which synthesizes the available scientific evidence on a specific topic. CERs expand the scope of a typical systematic review, which focuses on the

effectiveness of a single intervention, by comparing the relative benefits and harms among a range of available treatments or interventions for a given condition. In doing so, CERs more closely parallel the decisions facing clinicians, patients and policymakers, who must choose among a variety of alternatives in making diagnostic, treatment, and health care delivery decisions.

Comparative Effectiveness Reviews follow the explicit principals of systematic reviews. The first essential step is to carefully formulate the problem, selecting questions that are important to patients and other health care decision makers and examining how well the scientific literature answers them. Studies that measure health outcomes (events or conditions that the patient can feel, such as disability, quality of life or death) are given more weight than studies of intermediate outcomes, such as a change in a laboratory measure. Studies that measure benefits and harms over extended periods of time are usually more relevant than studies that examine outcomes over short periods.

Second, CERs explicitly define what types of research studies provide useful evidence and apply empirically tested search strategies to find all relevant studies. To assess effectiveness of other interventions, such as the efficacy of a drug, reviews may focus on the results of randomized controlled trials. For other questions, or to compare results of trials with those from everyday practice, observational studies may play a key role. The hallmark of the systematic review process is the careful assessment of the quality of the collected evidence, with greater weight given to studies following methods that have been shown to reduce the likelihood of biased results. Although well-done randomized trials generally provide the highest quality evidence, well-done observational studies may provide better evidence when trials are too small, too short, or have important methodological flaws.

A third critical step is to consider whether studies performed in carefully controlled research settings (efficacy studies) are applicable to the patients, clinicians and settings for whom the review is intended. A number of factors may limit the generalizability of results from efficacy studies. Patients are often carefully selected, excluding patients who are sicker or older and those who have trouble adhering to treatment. Racial and ethnic minorities may also be underrepresented. Efficacy studies also often use regimens and follow-up protocols that maximize benefits and limit harms but which may be impractical in usual practice. Effectiveness studies, which are conducted in practice-based settings, use less stringent eligibility criteria and assess longer-term health outcomes, are intended to provide results that are more applicable to "average" patients. They remain much less common than efficacy studies, however. A comparative effectiveness review examines the efficacy data thoroughly to ensure that decision makers can assess the scope, quality, and relevance of the available data and points out areas of clinical uncertainty. Clinicians can judge the relevance of the study results to their practice and should note where there are gaps in the available scientific information. Identified gaps in the available scientific evidence can provide important insight to organizations that fund research.

Finally, CERs aim to present benefits and harms for different treatments and tests in a consistent way so that decision makers can fairly assess the important tradeoffs involved for different treatment or diagnostic strategies. Expressing benefits in absolute terms (for example, a treatment prevents one event for every 100 treated patients) is more meaningful than presenting

results in relative terms (for example, a treatment reduces events by 50%). These reviews also highlight where evidence indicates that benefits, harms, and tradeoffs are different for distinct patient groups, high- vs. low-risk patients, for example. Reviews do not attempt to set a standard for how results of research studies should be applied to patients or settings that were not represented in the studies. With or without a comparative effectiveness review, these are decisions that must be informed by clinical judgment.

In the context of developing recommendations for practice, comparative effectiveness reviews are useful because they define the strengths and limits of the evidence and clarify which interventions are supported by strong evidence from clinical studies and which issues are less certain. Comparative effectiveness reviews do not contain recommendations and they do not tell readers what to do: judgment, reasoning, and considerations of the values of the relevant parties (patients, clinicians, decision makers, and society) must also play a role in decision making. Users of a comparative effectiveness review must also keep in mind that "not proven" does not mean an intervention is proven not effective; that is, if the evidence supporting a specific intervention is weak (i.e., strength of evidence is judged to be low or insufficient), it does not mean that the intervention is ineffective. The quality of the evidence on effectiveness is a key component, but not the only component, in making decisions about clinical policies. Additional factors to consider include acceptability to physicians or patients, the potential for unrecognized harms, the consequences of deferring decisions until better evidence becomes available, applicability of the evidence to practice, and consideration of equity and justice.

CERs are written for an audience of clinical decision makers. The text should be simple, clear, and as free as possible of the jargon of systematic reviews. Although CERs may be used in a variety of settings, the primary users are likely to be clinicians appointed by organizations or public agencies to make recommendations for the use of treatments, diagnostic tests, or other interventions. Payers and insurers may use them to make clinical and group policy decisions on benefits and coverage, and professional groups may base their clinical practice guidelines on them. Experts in informed consumer decision making can use the reports to develop decision aids and other tools that consumers can use to choose among alternative diagnostic and therapeutic strategies.

Review Team

EPC Directors are responsible for ensuring that a qualified team of investigators is available to conduct CERs. At a minimum, the EPC review team must have:

- expertise in conducting systematic reviews, including clinical epidemiology and statistical expertise;
- knowledge of specific issues that arise in conducting CERs; and
- relevant clinical expertise and, when indicated, access to specialists who have expertise in the interventions under review.

EPC directors are also responsible for ensuring that members of the CER teams are familiar with the information in this guide and that they take advantage of opportunities for training and other support provided by AHRQ and the Scientific Resource Center (SRC). EPC investigators

participate in regularly scheduled conference calls with AHRQ and SRC personnel, and they should use these calls to discuss how to apply the guidance.

To maintain public confidence in the scientific integrity and credibility of work produced by an Evidence-based Practice Center (EPC), it is essential that all aspects of the process and methodological approach on which the EPC evidence reports rest are clear and credible. The need for maintaining the scientific integrity of EPC products extends to and includes disclosure of participants' financial, business, and professional interests that are related to the subject matter of an EPC evidence report or other product or that could be affected by the findings of the EPC work. With respect to the types of financial interests to be disclosed, AHRQ is guided by the Department of Health and Human Services Regulations 45 CFR Part 94. Disclosure is required of EPC staff, consultants, subcontractors, and other technical experts. EPC Directors must ensure that all members of the review team comply with AHRQ policy.

Related financial, business and professional interests of EPC staff, consultants, and subcontractors do not, of themselves, disqualify one from substantive participation in development of an EPC evidence report or other product. AHRQ will consider such interests along with other technical attributes of the EPC and potential scientific contribution of available experts and options. Lead authors on the reports are barred from having any significant competing interests. Disclosure of financial, business, and professional interests assists AHRQ in carrying out its stewardship responsibilities in use of public resources to obtain evidence-based products in which the health care community, providers, purchasers, payers, and consumers will have confidence.

How the Guide was Developed

The material in this guide draws on published research and the experience of the investigators at the EPCs and of staff at the Scientific Resource Center and AHRQ in producing systematic reviews. Some of these issues and principles were discussed in a 2005 supplement to the *Annals of Internal Medicine* produced by the EPC program. In 2006, AHRQ commissioned the SRC to work with EPC scientists and AHRQ staff to develop more complete and explicit guidance for comparative effectiveness reviews. We initiated this process by establishing five workgroups made up of EPC investigators, AHRQ staff, and SRC staff. The five workgroups developed chapters on observational studies, applicability, harms and adverse effects, quantitative synthesis, and methods for rating a body of evidence. The workgroups met regularly, identified key issues and relevant methods papers, and reviewed the published guidance from major bodies producing systematic reviews—most importantly the Cochrane Collaboration Handbook (Cochrane Collaboration 2006) and the Centre for Reviews and Dissemination manual on conducting systematic reviews (CRD 2001; CRD 2007). Chapters for this guide were developed collaboratively by members of each workgroup. Individual workgroup leads were responsible for helping to produce the final draft and representing their workgroup decisions throughout the editorial process.

The goal for this draft of the guide was to improve the overall quality of CERs and increase consistency and transparency by providing guidance on a limited number of key issues. A list of

the most important points can be found at the end of this chapter. These points outline principles we are committed to following consistently in all CERs. A chapter on assessing evidence on diagnostic tests will be developed over the coming year and additional topics are likely to be added. The guide also provides discussion, key points, and preliminary guidance on a wider range of issues. For many of these issues, particularly those concerning statistical methods, some variation in practice may persist due to differing opinions about the relative advantages of different approaches and a lack of sufficiently strong empiric evidence to dictate a single method. As further information accumulates, we expect to define more specific requirements related to these issues. Finally, we will continue to assess the ability to implement our recommendations—both primary recommendations and secondary concepts introduced in this guide—as we undertake comparative reviews on a wide assortment of topics. We anticipate this guide will continue to evolve as we identify new issues and accumulate experience with new topic areas

Key Recommendations of this Guide

1) Authors of a review should begin by understanding the clinical and policy decisions that the review is intended to inform. Public input on key questions and consultation with stakeholders and content experts can help define critical issues involving details of the intervention, specific sub-populations of interest, and key outcomes of importance.

2) Decisions to include or exclude studies should be explicit and based on prespecified criteria. High-quality observational studies should generally be included where they can address important gaps in the evidence available from trials. Consulting experts and examining recent reviews and selected major trials provide insight into whether existing trials have sufficient power and are applicable to current practice.

3) All important harms should be assessed, using multiple sources of information. Observational studies should be included to assess long-term harms, uncommon events, and harms in more representative populations.

4) Comprehensive literature searches should include at least two electronic databases and supplemental measures to find relevant studies. Reviews should state whether they excluded studies based on language or publication status.

5) Included and excluded studies should be reported including reasons for exclusion.

6) Characteristics of included studies should be reported in summary tables, including aspects relevant to applicability of studies. Quality (i.e., risk of bias) of individual studies should be assessed and reported using predetermined criteria.

7) Quantitative synthesis should be performed to address pre-specified questions and following consistent approaches outlined in this guide. Clinical and methodological diversity as well as statistical heterogeneity should be considered before pooling studies to calculate a summary effect.

8) Heterogeneity should be explored with subgroup analysis or meta-regression techniques. The relationship between effect size and control rate should be examined if there is sufficient variation in both parameters. Sensitivity analysis is encouraged to explore the robustness of quantitative estimates to specific decisions in the review.

9) The strength of evidence should be assessed and reported for the major conclusions of the review using explicit criteria. Any factors that may limit the applicability of evidence for major conclusions should be summarized.

2. TOPIC DEVELOPMENT

Since the inception of the EPC program in 1997, AHRQ has emphasized the importance of input from technical experts and patients to elucidate the clinical logic or reasoning underlying questions for systematic reviews (Woolf, DiGuiseppi et al. 1996). For the Effective Health Care program, topic development extends this approach by soliciting and incorporating public commentary to develop the scope of a review.

Topic Nomination

ARHQ invites the public to nominate topics for CERs on a public website at http://effectivehealthcare.ahrq.gov/getInvolved.cfm. On a quarterly basis, the SRC collates and categorizes nominated topics from a variety of sources, including the public Web site, letters, key stakeholder groups, and program partners. This initial process involves combining duplicate topics and eliminating topics that previously have been evaluated. Once a list of topic nominations is compiled and approved as meeting initial priority criteria by AHRQ, the Scientific Resource Center (SRC) prepares a summary of relevant information for each proposed topic to assist AHRQ in further selecting priorities for research development. AHRQ considers several criteria in considering each nominated topic, include the following.

- The burden, prevalence, incidence, and impact of the condition or disease.
- The type of evidence supporting the efficacy and safety of the interventions (e.g., whether randomized trials have been conducted) and whether there are reasonably well-defined patient populations, interventions, and outcome measures.
- Current controversy about this topic or important uncertainties for decision makers.
- A topic's potential impact, with higher ratings for those nominations that address issues that impose high direct or indirect costs on patients or society; may be under- or over-utilized; or that may significantly improve the prevention, treatment, or cure of diseases and conditions.
- The potential value of a comparative effectiveness review relative to existing sources of systematic information.

Formulation and Refinement of Key Questions

A fully formulated CER topic consists of a set of questions, denoted "key questions," that specify the patient populations, interventions, comparators, outcome measures of interest, timing, and settings (PICOTS) to be addressed in the review (Counsell 1997). The elements of the PICOTS constructs are outlined below.

- Population: Condition(s), disease severity and stage, comorbidities, patient demographics.
- Intervention: Dosage, frequency, and method of administration.
- Comparator: Placebo, usual care, or active control.
- Outcome: Health outcomes: morbidity, mortality, quality of life.
- Timing: Duration of follow-up.
- Setting: Primary, specialty, in-patient; co-interventions.

The research questions largely dictate criteria for determining study eligibility for the systematic review. Therefore, clear, unambiguous, and precise questions are of paramount importance in selecting studies that address the same problem and are clinically and methodologically cohesive.

The formulation stage of a CER has three objectives: (1) developing key questions, (2) constructing definitions of the key concepts that distinguish relevant from irrelevant studies, and (3) establishing inclusion and exclusion criteria for the review. CERs should address important, specific questions that reflect the uncertainty that decision makers, patients, clinicians, and others may have about the topic. A perfunctory set of questions or an incomplete problem formulation that describes only the general outline of comparisons but not the specific circumstances that are of most interest to decision makers would reduce the usability of such a review (Woolf 1996; Counsell 1997; Atkins, Fink et al. 2005; Bravata, McDonald et al. 2005; Matchar, Westermann-Clark et al. 2005).

For each CER, a technical expert group (TEG)—consisting of a wide range of experts, including patients and stakeholders—participates in refining the questions before they are put into final form and assigned to an EPC. SRC personnel interview these experts, either individually or in small groups. Interviews aim to clarify and refine key questions early in the CER process by obtaining information about:

- populations and clinical subgroups of interest,
- patient characteristics that may affect outcomes,
- what interventions should be compared,
- the therapeutic aims of treatment, and
- what outcomes (intended and unintended effects) are relevant, including timing.

The key question process should identify the overarching, long-range goals of interventions. Focusing only on what is actually studied in the literature is insufficient. Sometimes very important questions concern assumptions about long-term effects on quality of life, morbidity, and mortality.

Beliefs about the advantages or disadvantages of various alternative treatments are an important target for exploration in interviews with experts. Some beliefs about the advantages and disadvantages of a treatment are based on direct evidence about health outcomes from long-term comparative trials. More often, though, beliefs about comparative effectiveness are based on clinical theories that invoke understanding of the pathophysiology of a disease, assumptions about its course, or expectations about the health benefits associated with improvements in a surrogate measure of outcome. Often, experts can bring attention to issues that underlie uncertainty about the comparative effectiveness of alternative tests or therapies (**Box 2-1**).

BOX 2-1. Topic development and clinical theories

"…every review, just like every intervention, is based on a theory. Systematic reviews gather evidence to assess whether the expected effect of an intervention does indeed occur." (*Cochrane Manual*)

Understanding the clinical logic underlying claims about comparative effectiveness is an important goal of topic development. Interviews with technical experts aim to answer questions such as:

Why do proponents of one or another treatment believe it is better?
When and for whom?
What characteristics of the alternative treatments are likely to drive choices?

The following examples illustrate how beliefs are linked to clinical theories:

Belief: Newer antisecretory drugs are likely to be better for glycemic control of diabetes than are sulfonylureas.

Theory: Sulfonylureas have been disappointing, and their use has not brought about a meaningful reduction in the risk of macrovascular complications. They may, in fact, be implicated in progression of diabetes, and they make it difficult to lose weight. Newer classes of drugs may result in better long-term outcomes because they have a better metabolic profile.

Context: Proponents of the new drugs do not base their claim of superiority on evidence about short-term glycemic control. The belief that the new drug will have an advantage is based on the understanding on how diabetes progresses; how the new drug works; and evidence from short-term efficacy trials about effects on lipid levels, weight gain, and other metabolic markers.

Belief: A new long-acting opioid drug for relief of pain is likely to play an important role in chronic pain treatment.

Theory: Because of tolerance and individual differences in response, chronic pain patients may have more consistent and prolonged symptom relief when several long-acting opioid medications are used in rotation.

Context: The belief that the new drug has an advantage is based on the fact that it has a long half-life, rather than on how the likelihood and degree of pain relief and the frequency and severity of side effects compare with alternatives. The review may want to focus on evidence about how this drug performs as a part of an opioid rotation regimen rather than as the sole or initial treatment for chronic pain.

Stakeholders and other technical experts can provide important insight to direct the search for evidence that is most relevant to current practice. First, they can clarify specific populations or interventions of greatest clinical or policy interest. For example, if a topic is important to Medicare policy decisions, finding evidence relevant to older Americans becomes critical. Clinical and research experts can offer insights about the extent to which studies in one population are relevant to other populations. For example, studies of oral cancer screening in India may not be applicable in a North American context because of the very high rates of oral

cancer in India. Second, knowledge of current practice can identify areas in which studies differ in ways that may reduce their applicability. For example, important changes in standards of care for acute myocardial infarction over the past 2 decades limit the relevance of studies of interventions for acute MI from the 1980s. Some other features of studies that affect applicability to current practice are dosing of drugs, modifications or special features of devices, and surgical settings.

Key questions—and systematic reviews—must also be patient-centered. Interviews with patients, as well as studies of patients' preferences when they are available, are essential to identify pertinent clinical concerns that even expert health professionals may overlook (Santaguida, Helfand et al. 2005).

Before key questions are put into final form, AHRQ invites the public to comment on them. AHRQ staff use the public comments to develop the final set of questions to be assigned to an EPC. Once AHRQ makes an assignment, the Agency provides a document describing the background and context for the CER and the key questions. The EPC analyzes these materials and prepares questions about its context and content. On a "kickoff" call, the EPC investigators, the AHRQ Task Order Officer, and the SRC investigator who worked on topic development discuss these questions, providing additional information about the input from experts and stakeholders that led to the specific wording and organization of the questions.

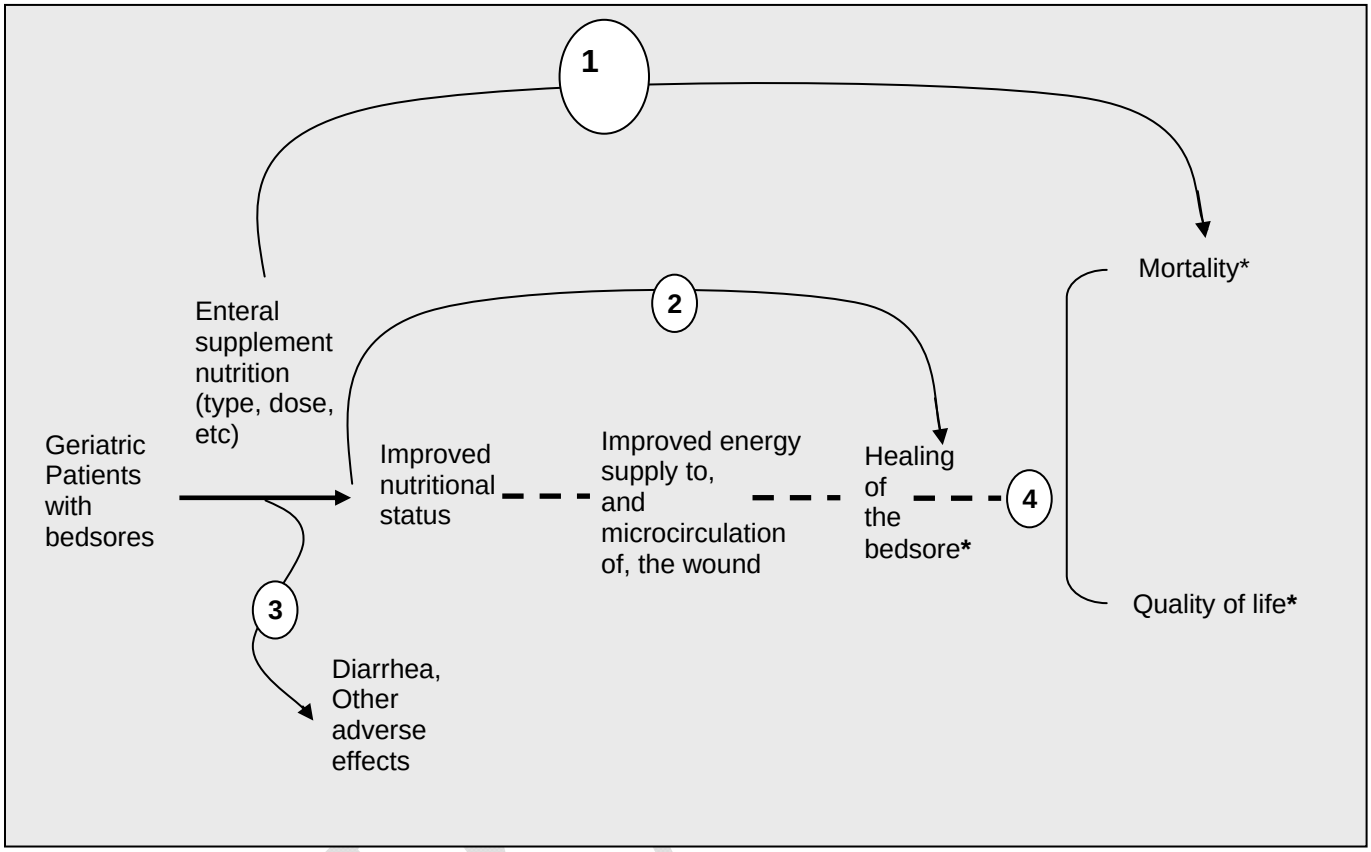
Analytic Frameworks

An evidence model, variously also referred to as an analytic framework or causal pathway, portrays relevant clinical concepts and the clinical logic underlying beliefs about the mechanism by which interventions may improve health outcomes (Woolf, DiGuiseppi et al. 1996). In particular, the evidence model describes the relationship between surrogate or intermediate outcome measures (such as cholesterol levels) and health outcomes (such as myocardial infarctions or strokes).

Several graphical and analytical approaches can be used to build an evidence model. In the EPC program, the most commonly used approach is called an "analytic framework" (Harris, Helfand et al. 2001; Whitlock, Orleans et al. 2002). The main function of the framework is to define the populations, interventions, outcomes, and adverse effects for the literature search and synthesis.

Analytic frameworks are good for illustrating the relationship between intermediate measures and health outcomes, to define bodies of evidence (Mulrow, Langhorne et al. 1997), and to depict clinical controversies and uncertainties (**Figure 2-1**). They reflect insights into the relationship between intermediate or surrogate outcomes and health outcomes and specify the populations, interventions, comparisons, and outcomes of greatest interest.

Figure 2-1. Analytic framework for a new enteral supplement to heal bedsores.



Key questions are associated with the linkages (arrows) in the analytic frameworks. In the figure, Arrow 1 corresponds to a question directly linking enteral supplementation to the two most important outcomes: mortality and quality of life. In the absence of evidence directly linking enteral supplementation with these outcomes, the case for using the nutritional supplement depends on a series of questions representing several bodies of evidence:

- Key question 2: Does enteral supplementation improve wound healing?
- Key question 3: How frequent and severe are side effects such as diarrhea?
- Key question 4: Is wound healing associated with improved survival and quality of life?

Note that in the absence of controlled studies demonstrating that using enteral supplement improves healing (link #2), EPCs may need to evaluate additional bodies of evidence. Specifically this would include evidence linking enteral supplementation to improved nutritional status and other evidence linking nutritional status to wound healing.

A more rigorous alternative to depicting the clinical logic underlying a service is to use a decision analysis or influence diagram to model key parameters of a decision and their relationships (Bravata, McDonald et al. 2005). In either case, the evidence model reflects insights into the relationship between intermediate or surrogate outcomes and health outcomes and specifies the populations, interventions, comparisons, and outcomes of greatest interest.

Modifying Key Questions

Review teams may identify problems with key questions that warrant discussion at the outset. These may include excessively vague terms that would broaden the scope beyond anything manageable. Such problems may also include inappropriate or nonspecific clinical terms when more specific diagnostic or interventional terms are needed. These matters should be clarified at the earliest possible moment.

Occasionally, in the process of conducting a review, investigators may realize that the questions need to be refined (CRD 1996). The most common reason to consider modifying the key questions is that a new patient group or outcome of emerging importance was not specified in the original questions. Occasionally, a new compound or device will receive approval for use in the United States after the work is under way.

The Cochrane Handbook (http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.doc) notes that, when proposing a change in the key questions, investigators should answer the following questions.

- What is the motivation for the refinement?
- When was the refinement made?
- Are the search strategies appropriate for the refined question (especially any that have already been undertaken)?
- Is data collection tailored to the refined question?

Because avoiding bias is critical when changing key questions, EPCs should discuss proposed revisions with AHRQ and the SRC as early as possible. AHRQ will then determine whether the key questions should be modified and, if so, how. Often, incorporating a new patient group, intervention, or outcome into the next update of a report is a reasonable alternative to modifying the key questions in the course of a review.

3. SELECTING EVIDENCE: CONTROLLED TRIALS

This chapter addresses the use of trials in comparative effectiveness reviews (CERs). The term "controlled trials" refers to experimental studies in which there is a comparison between different interventions. The investigator assigns the intervention and patients are allocated to groups either randomly (randomized controlled trials or RCTs) or nonrandomly (other controlled trials). Such studies are also called "experimental" studies to distinguish them from "observational" studies, in which investigators do not assign interventions to subjects. Chapters 4 and 8 discuss the role of observational studies in CERs.

Evidence-based medicine has been associated with a hierarchy of evidence that ranks RCTs higher than other types of evidence in all possible situations (Bigby 2001; Devereaux and Yusuf 2003). The reason is the superiority of RCTs over other types of studies in reducing the risk of bias (i.e., poor internal validity). In recent years, broader use of systematic reviews in policymaking has brought attention to the danger of over-reliance on RCTs:

> …the simplifications involved in creating and applying hierarchies have also led to misconceptions and abuses. In particular, criteria designed to guide inferences about the main effects of treatment have been uncritically applied to questions about aetiology, diagnosis, prognosis, or adverse effects. (Glasziou, Vandenbroucke et al. 2004)

Box 3-1. Selecting Trials

Published studies vary widely in their quality and relevance. The value of any individual study as evidence depends on the specific question being addressed.

Different types of studies have differential relevance and risk of bias, and their use as evidence depends on the question addressed.

Randomized trials and other controlled clinical trials can (but often don't) address effectiveness and provide evidence that is directly applicable in clinical settings.

Head-to-head effectiveness *trials*—trials that meet the criteria for effectiveness studies—are the best evidence to assess comparative effectiveness Other types of trials usually have characteristics that limit their applicability in practice and their usefulness in comparative effectiveness reviews.

Such characteristics include small or highly selected samples, short duration of follow-up, use of intermediate endpoints, and incomplete ascertainment of benefits and adverse outcomes.

Various experts have made suggestions for changing or expanding the hierarchy of evidence to take better account of evidence about adverse effects and effectiveness in actual practice (Concato, Shah et al. 2000; Tucker and Roth 2006; Walach, Falkenberg et al. 2006).

From the outset, AHRQ's EPC program has taken a broad view of eligible evidence (**Box 3-2**) (Woolf 1996; Atkins, Fink et al. 2005). In contrast to Cochrane reviews, most of which exclude all types of evidence except for RCTs, inclusion of a wider variety of study designs has been the norm rather than the exception in the EPC program (Bravata, McDonald et al. 2005; Chou and Helfand 2005; Norris and Atkins 2005; Pignone, Saha et al. 2005; Shekelle, Morton et al. 2005; Tatsioni, Zarin et al. 2005).

Box 3-2. Types of Evidence: Excerpt from the EPC Manual for Conducting a Systematic Review (1996) (Woolf 1996)

Collecting and reading the literature is one of the most time-consuming tasks in a systematic review. Expanding these resources can be especially wasteful if the reviewers "cast too wide a net" and gather evidence of poor quality or with limited relevance to the questions raised by the evidence model. On the other hand, if the literature review is too narrow, important sources of evidence may be omitted (Slavin 1995).

Published evidence can include a heterogeneous group of data sources of variable quality and relevance. Excluding an entire category of literature is not without risks. Randomized controlled trials are unavailable for many aspects of medicine, due largely to the cost and time requirements to perform them. Limiting a review to such trials might exclude important data from other types of studies (e.g., cohort studies, case-control studies, descriptive epidemiology). For some topics, evidence from animal models or laboratory studies is essential. Even review articles, editorials, and letters-to-the-editor, which are often omitted because they lack primary research data, can provide important insights about published studies. Their reference lists can also help verify the comprehensiveness of the review's bibliographic database.

On the other hand, casting a wide net opens the door to studies of dubious quality and can expand the volume of a search to hundreds or thousands of superfluous articles. Doing so can be especially inefficient. If good evidence from a few major clinical trials is available, there may be no purpose in spending time and money to collect hundreds of retrospective studies and case reports on the same subject. Thus, before they determine the appropriate boundaries for admissible evidence, reviewers should conduct a preliminary literature search to obtain a sense of the type of evidence that is available. They can then perform a "best-evidence" synthesis, limiting the review to the highest quality studies and foregoing the collection and review of other evidence (Slavin 1995).

In the Effective Health Care program, the conceptual model for considering different types of evidence still emphasizes minimizing the risk of bias, but it places highly applicable evidence about *effectiveness* at the top of the hierarchy. The model also emphasizes that simply distinguishing RCTs from observational studies is insufficient because different types of RCTs vary in their usefulness in comparative effectiveness reviews. This chapter describes the roles of different types of controlled trials in conducting CERs. **Chapter 4** describes the role of observational studies in addressing gaps in the evidence, gaps that often reflect the limited applicability of some kinds of RCTs.

While CER investigators usually begin with controlled trials, it's clear that evidence obtained under the carefully controlled setting of RCTs may not accurately reflect the benefits and harms observed under the conditions of everyday practice. Measures to promote rigor in clinical trials, such as careful patient selection criteria and tight control of the intervention may produce results that are less relevant for clinicians and decision makers. Haynes has noted that a useful assessment of a clinical intervention needs to answer not only "Can it work?" but "Will it work?" and "Is it worth it?" (Haynes 1999).

Effectiveness Trials

For CERs, trials that address the questions, "Will it work?" are the best evidence. Such trials are called "effectiveness" or "practical" trials. Effectiveness trials aim to study patients who are likely to be offered the intervention in everyday practice. They also examine clinical strategies that are more representative or likely to be replicated in practice. They may measure a broader set of benefits and harms (whether anticipated or unanticipated), including self-reported measures of quality of life or function and they seek to measure the degree of beneficial effect and harms under "real world" clinical settings. Thus, effectiveness trials are better at answering the second and third questions posed by Haynes: whether an intervention *will* work in typical practice and whether it is *worth it* in terms of the balance of important benefits and harms.

Gartlehner and colleagues developed seven criteria to distinguish effectiveness studies from efficacy (explanatory) studies (Table 3-2) (Gartlehner, Hansen et al. 2006; Gartlehner, Hansen et al. 2006). Effectiveness is context-specific. This list is most useful to judge the applicability of a drug study of a common chronic condition in primary care, but would not work well to define "effectiveness" for studies of treatment of a self-limited, acute episode of illness (e.g, a urinary tract infection) or of a highly specialized procedure (e.g., the MAZE procedure to restore sinus rhythm in patients undergoing open heart surgery who have atrial fibrillation.)

Table 3-2. Criteria for effectiveness studies.

Item 1	Populations in primary care
Item 2	Less stringent eligibility
Item 3	Health outcomes
Item 4	Long study duration; clinically relevant treatment modalities
Item 5	Assessment of adverse events
Item 6	Adequate sample size to assess a minimally important difference from a patient perspective
Item 7	Intention-to-treat analysis

When they are available, head-to-head effectiveness trials—randomized trials that meet the criteria for effectiveness studies—are the best evidence to assess comparative effectiveness. With respect to evaluating results in actual practice, effectiveness trials have the same advantages as observational studies but use better means to minimize the risk of bias from confounding by indication and other threats to internal validity (Mosteller 1996; McAlister, Straus et al. 1999; Medical Research Council 2000; Godwin, Ruhland et al. 2003; Tunis, Stryer et al. 2003; Kotaska 2004; Glasgow, Magid et al. 2005).

Efficacy Trials

For many topics, effectiveness trials are either unavailable or do not address all the key questions about comparative effectiveness and safety. In this common situation, efficacy (also called explanatory) trials may provide the best available evidence to answer a question. Efficacy trials are an efficient, necessary type of clinical research for testing hypotheses about novel

interventions ("Does it work?"). They usually incorporate design elements that provide a high degree of internal validity, such as randomization, concealment of allocation, blinding, and intention-to-treat analysis. They also incorporate features that maximize the chance of finding a difference between the intervention under study and a comparator in the shortest possible time at the lowest possible cost. For example, they may exclude patients who have mild disease or comorbidities, use medications for other conditions, take other effective treatments for the condition under study, or are relatively unlikely to adhere to medication regimens. They may use a placebo for comparison instead of a viable treatment alternative. They usually focus on a selected number of short-term intermediate outcomes. These features reduce the number of subjects needed to enroll and the time required to get the answer to a specific hypothesis about the intervention under study.

Types of efficacy trials used in CERs include:

- long-term head-to-head trials that do not meet all of the criteria for effectiveness studies;
- short-term head-to-head trials that focus on short-term surrogate measures of tolerability and side effects, often in highly selected samples under rigorous study conditions;
- long-term placebo-controlled trials assessing important health outcomes or harms of a particular drug; and
- short-term, placebo-controlled trials that focus on surrogate outcome measures.

The ordering of these categories reflects the policy of looking carefully at head-to-head trials first, then looking at placebo-controlled trials. However, many variations are possible. A particular CER may use any or all of these types of trials, depending on the questions to be answered and the completeness and relevance of evidence from effectiveness trials. Box 3-2 shows an example in which placebo-controlled trials played an important role in a comparative effectiveness review.

Some efficacy trials can also be described as dose-finding, equivalence or noninferiority trials. Dose-finding studies seek to identify the optimal starting dose of a drug by comparing, several different doses to placebo. Equivalence and noninferiority trials compare a new drug with one or more established drugs. An equivalence trial assesses whether the drugs are therapeutically similar. A noninferiority trial seeks to determine whether whether the new treatment has at least as much efficacy as the standard treatment. CER investigators should be familiar with the conceptual underpinnings (Sackett 2005) and methods for reporting equivalence and noninferiority trials (Piaggio, Elbourne et al. 2006).

Box 3-2. Examples Using Different Categories of Trials in a CER of Statins

For lipid-lowering therapies, all-cause mortality, cardiovascular mortality, cardiovascular events, and severe adverse events are the four most important outcome categories. Effects on reaching targets for serum levels of low-density lipoprotein-c (LDL-c) and high-density lipoprotein-c (HDL-c) are important, too, particularly from the viewpoint of everyday clinical management of patients in primary care. No trials met all of the criteria for an effectiveness study. The available trials were categorized as follows:

1. A few long-term, comparative trials reported mortality and cardiovascular events, but they concerned selected patients and compared a high dose of one statin with a low dose of another. Although of some interest, such studies are not effectiveness studies and leave many questions about comparative effectiveness unanswered.

2. Many head-to-head trials report comparative efficacy of statins on serum levels of LDL-c, HDL-c, and triglycerides and on their side effects. However, this body of head-to-head efficacy trials recruits highly selected subjects and may not reflect results in actual practice.

3. Many placebo-controlled trials of statins assessed important outcomes and harms, and, in fact, proved that specific statins can reduce mortality and cardiovascular events. Because they use intention-to-treat analysis, they demonstrate that these benefits accrued after taking into account the effects of early discontinuation and rare but serious adverse reactions. Indirectly, these placebo-controlled trials facilitate comparison among the options, because they show which treatments are proven to improve long-term outcomes in specified populations. However, these trials do not provide a clear picture of the risks of treatment in subgroups underrepresented in these trials, including racial minorities, the elderly, women, and patients with other diseases.

4. Placebo-controlled RCTs that focus on LDL-c, HDL-c, and other intermediate measures and have minimal information about adverse effects. These trials add little to the information available from short-term, head-to-head trials (category 2 above.) Such studies may be useful in a CER if, for example, they address an intervention or outcome (e.g., C-reactive protein levels) not addressed in other studies.

Applicability of Efficacy Trials

The main disadvantage of efficacy trials is that they rarely provide all the necessary information to answer how well a treatment will work in practice or how the benefits compare with adverse effects for a specific patient (Atkins 2007). This has been variously termed *applicability, external validity, generalizability, and relevance*, each with a slightly different connotation. For the purposes of this discussion, we will use the term *applicability*, but we borrow the definition that Shadish and others have put forward for external validity:

> Inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments and outcomes. (Shadish, Cook et al. 2002)

Because efficacy trials are not designed to answer "Will it Work?" questions about their applicability arise in most CERs. It is important that CER authors use a consistent approach to identifying gaps in the trial evidence due to limitations in the applicability of efficacy studies. As described in Chapter 3, identifying these gaps early on can guide selection of observational studies for inclusion in a CER. As described in Chapter 6, it is also important to describe specific study features that may limit applicability in a way that will enable decision makers to judge the relevance of the trials to their patients and settings.

The aspect of applicability that has received the most attention has been whether the study population is representative. A variety of publications have noted that patients enrolled in

clinical trials often differ in important ways from patients seen in practice with the same condition (Zarin, Young et al. 2005; Steg, Lopez-Sendon et al. 2007), usually having better outcomes than patients in the community. Other factors, however, are equally important in assessing applicability. These include the nature of the interventions used, the appropriateness of the comparator chosen, the outcomes and time frame for measuring them, and the setting in which the research was conducted. Rothwell has detailed how specific features of clinical trials may produce results that may not apply to other populations or settings (Rothwell 2005).

Investigators with an interest in behavioral and community-based interventions have highlighted similar concerns about applicability but from a slightly different perspective. Of primary concern is whether interventions in research studies are suited for wider implementation and address needs of policymakers and the community. Green and Glasgow have argued that research publications should pay closer attention to the costs, suitability, population "reach," adaptability, and sustainability of interventions, and they have proposed specific criteria for reporting on such issues (Green and Glasgow 2006).

The availability of evidence from more inclusive trials with longer-term endpoints or better ascertainment of benefits and adverse outcomes may render efficacy studies irrelevant. However, when efficacy studies are included in a CER, it is essential to evaluate and report their applicability.

The primary aim of assessing applicability is to determine whether the results obtained under research conditions are likely to reflect the results that would be expected in broader populations, under "real-world" conditions. To do this, reviewers must begin with basic understanding of the characteristics of treated patients, which interventions are used in practice and how (for example, typical doses for drugs), and the results of interventions in everyday clinical settings. This information can be obtained by consulting clinical experts, examining recent clinical review articles, and retrieving selected studies reporting treatment experience in the community. This approach is more feasible than more systematic attempts to review current practice patterns. Without information about what happens in actual practice, speculation about the applicability of trial results is just speculation.

Applicability is a relative rather than absolute concept; no trial can enroll the exact population or deliver the exact intervention appropriate to all settings and populations of interest. Whether these differences are important enough to render a study's conclusions of limited use is ultimately a judgment that cannot be reduced to simple algorithms or scoring instruments.

Domains of Applicability

The PICOTS (Population, Intervention, Comparator, Outcome, Timing, Setting) format is recommended as a way to consistently frame key questions and present study results, and it provides a useful way of organizing information relevant to applicability (Chapter 6) (Richardson, Wilson et al. 1995). Different domains should be emphasized based on the nature of the intervention: distinguishing among drugs, surgery and invasive procedures, diagnostic tests, and behavioral interventions. The most important issue with respect to applicability is whether the outcomes are different across studies that recruit different populations, use different doses or forms of the intervention or comparator, or differ in duration. That is, important

characteristics are those that affect baseline (control group) rates of events, intervention group rates of events, or both.

Study Characteristics That Affect Applicability

Differences between trial circumstances and everyday circumstances are the rule rather than the exception. Typically, the spectrum of patients who receive an intervention in practice is broader than the spectrum of patients recruited and retained in clinical trials; dosing and adherence are often very different in practice; practitioners use a broad range of alternative treatments (and never use a placebo); and patients are observed for longer periods of time than in most efficacy trials.

Characteristics of efficacy trials that may limit applicability of results in drug and device studies are listed below, using the PICOTS framework. CER investigators should pay attention to these features of individual studies in their decisions about which efficacy trials to include or exclude. Sometimes applicability concerns may become an exclusion criterion – reviewers may elect to exclude studies only reporting intermediate endpoints or other outcomes that are not felt to be sufficiently applicable to clinically important outcomes. Alternatively, reviewers may choose to include a broader set of efficacy trials because it allows them to examine the intervention in a broader population of patients

- **Patient population:** Strategies for recruiting participants in clinical trials can result in homogeneous samples that differ from target population in severity of illness, comorbidities, and demographics (age, sex, and race). Efficacy trials typically seek to maximize the effect of the intervention by selecting for patients with more severe disease or those at highest risk of the outcome of interest, while excluding patients with comorbidities, the frail elderly, and those who may have trouble with adherence or follow-up. Recruitment strategies also usually minimize risk of harm. Run-in periods are a further way to reduce variation in the patient population—run-in periods may be used to select for those responding to treatment, those who are adherent, or those who experience few adverse effects.

- **Intervention - Intensity (dose, duration, and co-interventions):** Prescribing behavior and cointerventions are major reasons why treatment patterns in actual practice may differ from trial protocols. For example, physicians in practice may use lower than recommended doses, especially when prescribing for patients who they believe to be more vulnerable to adverse effects than those included in trials. Trials may include co-therapies, supportive care, and more frequent follow-up than is typical in practice, all of which may affect applicability. Behavioral interventions may include visit duration and frequencies that are impractical in most clinical practice settings. Drug trials frequently report short-term results even when drugs are typically used for long-term therapy.

- **Intervention adherence:** As used in the pharmacoepidemiologic literature, adherence encompasses whether a patient takes a prescribed drug at all (acceptance), whether they take it as prescribed (execution), and how long they take it (persistence). Because trials often recruit motivated patients and may use follow-up techniques (including pill counts)

- **Intervention – training and expertise:** Therapies that involve technical or other special skills depend on the level of training and experience of the operator, whether it is a surgeon, a physician performing an invasive procedure, a clinician performing or interpreting a diagnostic test, or a nurse delivering a lifestyle intervention. Expertise in trials usually exceeds that in the community, due to the settings involved (e.g., specialty centers), strategies for recruiting and training participating physicians or clinical sites, and training and monitoring performed as part of the trial. This higher expertise may exaggerate benefits and underestimate harms of the intervention in more representative settings.

- **Comparator - choice and dosing:** The most appropriate comparator for a new intervention is usually the best alternative care. For example, surgery trials should compare surgery with best medical care and new drug therapies should be compared with the current best alternative. Use of an inferior comparison treatment, such as an inadequate dose of the comparison drug, can exaggerate comparative benefits of the new therapy.

- **Outcomes:** Outcomes should include the most important clinical benefits and harms. Surrogate (or intermediate) outcomes should be viewed with caution unless the link between the surrogate and clinical outcomes has been previously validated in an intervention trial. Similarly, composite outcomes should not be relied upon unless the components are of equal importance and the intervention effects on each component are comparable (Ferreira-Gonzalez, Permanyer-Miralda et al. 2007).

- **Timing of outcome measurement:** Follow-up duration should be long enough to detect likely adverse effects, examine the persistence of benefits, and assess the sustainability of the intervention. This is true for surgery or invasive interventions, as well as for drug therapies where adverse effects may depend on accumulated dose or time and where benefits may decline as adherence falls off over time. Although inadequate follow-up may underestimate benefits (where beneficial effects require a longer duration of treatment), it more often will exaggerate benefits by overlooking any adverse effects that emerge more slowly and by overestimating long-term adherence to chronic therapies.

- **Setting:** Differences in setting—including country of study, rural vs. urban, or primary vs. specialty care—can influence numerous aspects of the population and intervention. Geographic differences can influence population characteristics, the intensity and quality of the intervention and cointerventions (e.g., available infrastructure or training for an invasive procedures), available comparators (e.g., what constitutes usual care), and the measurement of outcomes, among other factors. The clinical setting, such as primary versus specialty care, can influence population characteristics (e.g., severity of illness) and the intensity of the intervention (e.g., training and experience with a surgical procedure). Setting is often more important for interventions dependent on skill or technology and less important for drug interventions.

These features relevant to applicability should be abstracted and reported in evidence tables for all included studies. Chapter 6 discusses how these features should be assessed in summarizing applicability for a body of evidence. Chapter 11 discusses the role of applicability in weighing the strength of evidence.

4. SELECTING EVIDENCE: OBSERVATIONAL STUDIES OF BENEFICIAL EFFECTS.

Box 4-1. Key points (Selecting Evidence)

Because it is unusual to find sufficient evidence from trials to answer all key questions, the default approach for CERs is that the EPC will consider observational studies for inclusion in reviews assessing benefits for drugs, procedures, or devices.

The decision and rationale to include or exclude observational studies, of various types, must be thoughtfully presented in the methods or results section, as appropriate.

Whether trial data alone will provide a sufficient body of evidence for the CER rests in part on whether the trials answer all aspects of the key questions—i.e., all PICOTS (population, intervention, comparator, outcome, and setting) characteristics mentioned in Chapters 2 and 3.

Whether observational studies will provide useful information rests on several considerations: (1) whether they are biased or confounded, so that they will not provide meaningful data; (2) the estimated potential magnitude of benefits compared with the estimated potential magnitude of harms; and (3) the estimated potential random variation in outcomes. These factors may differ by type of observational study.

Recommended approach

Clearly define the key review questions with respect to PICOTS.

Perform a preliminary search for relevant trials and consult experts in the field as to the potential number of trials that address the review questions. Focus carefully on all aspects of the review questions, e.g., ensure that subgroups of interest have been specifically examined in trials.

Examine well-known or large trials. If these trials address all important aspects of the review questions, then observational studies may not need to be included. Since this rarely occurs, EPCs would need to justify any decision to exclude observational studies.

If data from trials do not appear to be sufficient to answer the review questions, then assess the potential for bias or imprecision in observational studies and the magnitude of the effect size in relationship to the decision threshold and potential harms. If the bias and uncertainty among observational studies likely outweigh the effect size, observational studies should not be included. The rationale for excluding observational studies should be made explicit.

If observational studies are likely to provide valid data on important outcomes, proceed with a search to identify these studies and proceed with their analysis and synthesis.

Systematic reviewers disagree about the role of observational studies to answer questions about the benefits or intended effects of interventions. In contrast, there is wide agreement that observational studies, particularly those derived from large clinical and administrative databases, should be used routinely to identify and quantify potential harms (Chapter 8).

A previous review of the Evidence-based Practice Centers' work found wide variation in the inclusion of observational studies (Moja, Telaro et al. 2005; Norris and Atkins 2005). It noted that no established guidelines address situations in which nonrandomized studies can or should be considered for inclusion in a systematic review or what study designs to consider. It also noted that there was also a lack of consensus on how to assess the internal validity of such studies. The authors recommended that systematic reviewers assess the strength of the RCTs before determining final inclusion criteria, then consider the pros and cons of nonrandomized study designs.

The Effective Health Care program endorses this recommendation. As described above, in Chapter 3, evidence from RCTs is often insufficient to answer the key questions convincingly, most often because the evidence they provide may not be widely applicable in practice. It is unusual to find sufficient evidence from trials to answer all key questions. Therefore, the initial assumption or default approach for CERs is that the EPC will consider observational studies for inclusion in reviews assessing benefits for drugs, procedures, or devices.

Decision Framework

Including data from observational studies involves a significant increase in time and resources required to complete a CER. Therefore, EPCs should use a step-wise process, whereby reviewers first examine trial data to see if they are sufficient before considering various types of observational studies.

In considering whether to use observational studies in CERs for addressing beneficial effects, EPCs should answer two questions.

1. Are there gaps in the RCT evidence regarding the key questions?
2. Considering the potential for bias and other factors, will observational studies provide valid and useful information to address the review questions?

1. Are there gaps in the RCT evidence regarding the key questions?

Identifying gaps in the RCT evidence available to answer the key questions can occur at almost any point in the review. For example, gaps may have been identified by the Coordinating Center in the initial scoping exercise, and be included in the RFTO or key questions themselves. Existing reviews on related topics or clinical experts may have already identified important gaps in the RCT evidence. Another common point at which gaps are identified is on the initial screening of RCTs, where the review team determines that all the RCTs involve short term outcomes, or lack a key outcome of interest. A third common point at which gaps are identified occurs after detailed review of the RCTs (that covers the items listed in **Table 4-1**.).

Table 4-1. Criteria for assessing whether a body of trial data is sufficient

Criteria	Definition	Considerations
Risk of bias (Internal validity)	Minimize bias and adjust for confounding, so that conclusions are valid.	Serious flaws in study design or execution should be considered within and across studies; these flaws potentially invalidate the results (e.g., lead to a conclusion of benefit when there is none).

Consistency	The extent to which effect size and direction vary within and across study designs.	This may be due to heterogeneity across PICOTS or the etiology may not be apparent.
Directness	Outcomes that are important to users of the CER (whether patients, clinicians, or policymakers).	These are often health outcomes and not surrogate, intermediate, or physiologic outcomes.
Precision	Precision encompasses sample size, number of studies, and heterogeneity within or among studies.	Greater levels of precision may be needed if the sizes of benefits and/or harms are closely balanced or if either is near a threshold that decision makers might use to make a recommendation.
Magnitude of benefit compared with harms	The size of the beneficial effect compared with the size of potential adverse effects and their relationship to the threshold for decision making.	Estimates of benefits that are much greater than those of harms and that exceed the likely threshold for decision making may contribute to sufficient trial data.
Reporting bias	Trial authors appear to have reported all outcomes examined and there is no strong evidence for publication bias (at the study level)	
Applicability	The extent to which the trial data are likely to be applicable to populations, interventions, and settings of interest to the user.	The review questions should reflect the PICOTS characteristics of interest.

The most compelling situation for using observational studies occurs when all trials can be classified as efficacy studies (Chapter 3). Efficacy trials often recruit selected populations and do not adequately examine longer-term, patient-centered outcomes. When all trials have these characteristics, gaps in applicability may be apparent at the outset, and observational studies may be useful to answer applicability questions. For example, in a review of antipsychotic medications (McDonagh, Carson et al. 2006), short-term trials evaluated a relatively narrow spectrum of patients with schizophrenia; raising the following applicability questions.

- Is the effect size observed in the RCTs similar to that observed in practice?
- Do groups of patients excluded from the trials respond as frequently and as well as those included in the trials?
- Are long-term outcomes similar to short-term outcomes?
- For a broad spectrum of patients with schizophrenia initiating treatment with an atypical antipsychotic medication, which drug(s) have better persistency and sustained effectiveness for 6 months to 2 years?

Well-done observational studies can address these questions if they include more representative patient populations, have relevant comparators, and report more meaningful clinical outcomes over longer time periods. Sometimes concerns about applicability can be identified at the outset—for example, many drug trials in schizophrenia are relatively short and exclude patients

with co-morbidities. In other cases, stakeholders may raise concerns about whether certain trial results are applicable to typical patients (Box 4.2). Lastly, examining features of available trials will reveal whether the interventions or patient populations are representative of current practice.

Box 4-2. Expert Input raise questions about applicability

A review of percutaneous coronary intervention (PCI) vs. coronary artery bypass (CABG) for coronary disease identified 23 RCTs conducted from 1987 to 2002. At the beginning of the review, cardiothoracic surgical experts raised concerns that the studies enrolled patients with a relatively narrow spectrum of disease (generally single or two-vessel disease) relative to those getting the procedures in current practice. The review also included 96 articles reporting findings from 10 large cardiovascular registries. The registry data confirmed that the choice between the two procedures in the community varied substantially with extent of coronary disease. For patients similar to those enrolled in the trials, mortality results in the registries reinforced the findings from trials (i.e., no difference in mortality between PCI and CABG). At the same time, the registries reported that the relative mortality benefits of CABG vs. PCI varied markedly with extent of disease, raising caution about extending trial conclusions to patients with greater or lesser disease than those in the trial population.

Identifying gaps early in the review process may lead the team to perform their initial searches very broadly, to identify both RCT and observational study evidence in the same search. Or, EPCs may do these searches sequentially, and search for observational studies only after reviewing in detail all the identified RCTs. The important point is that there is an explicit assessment of whether or not there are gaps in the RCT evidence, and if so, an explicit consideration of the potential usefulness of observational studies to help fill these gaps. If trial data are sufficient to answer the key questions, EPCs do not need to consider other study designs. In the example provided in Box 4-3, reviewers found conclusive trial data, and did not go on to assess observational studies of antioxidant supplementation. It is expected that in most CER reviews, gaps will be present, and observational studies will be considered.

Box 4-3. Trial data are sufficient: antioxidant supplementation to prevent heart disease mortality.

This clinical question has been studied in numerous large clinical trials, including among 20,536 elevated-risk subjects participating in the Heart Protection Study.(Heart Protection Study Collaborative 2002) No beneficial effects were seen in numerous cardiovascular endpoints including mortality. The size of the trial, the rigor of its execution, the broad spectrum of adults who were enrolled, and the consistency of the findings across multiple outcomes all support the internal validity and applicability of the findings of the Heart Protection Study to most adults with an elevated risk of cardiovascular events.

Box 4-4 illustrates a more common scenario that EPCs may face in assessing the sufficiency of trial data. In this example, although a large number of head-to-head efficacy trials were available, they provided insufficient evidence to assess two important longer-term outcomes. In another review (Box 4-5), few or no trials were identified, so the authors planned to consider including observational studies early in the review process. This scenario is also common,

particularly for reviews of certain surgical procedures, diagnostic procedures, and therapeutic devices.

Box 4-4. Important outcomes are not captured in trials

More than 50 RCTs of triptans focused on the speed and degree of migraine pain relief related to a few isolated episodes of headache. These trials provided no evidence about two outcomes important to patients: the reliability of migraine relief from episode to episode over a long period of time, and the overall effect of use of the triptan on work productivity. The best evidence for these outcomes came from a time-series study based on employment records merged with prescription records comparing work days lost before and after a triptan became available. Although the study did not compare one triptan with another, the study provided assurance that a particular triptan improved work productivity—information that was not available for other triptans.

Box 4-5. Paucity of trial data and inadequacy of available evidence

In a recently completed EPC report (AHRQ Report #148) on heparin to treat burn injury (Oremus, Hanson et al. 2006), the McMaster EPC determined very early in its process that observational data should be included in the report to address effectiveness key questions. Based on preliminary, cursory reviews of the literature and input from experts, the authors determined that there were few (if any) RCTs on the use of heparin for this indication. Therefore, they decided to include all types of studies that included a comparison group before running the main literature searches. The major limitation of the included studies (both trials and observational studies) was poor methodologic quality (e.g., inadequate randomization, no control of confounding), and the observational data added little to the review. However, the review was comprehensive and discussed the extent and limitations of all available evidence.

2. Will observational studies provide valid and useful information to address key questions?

To decide whether including observational studies will add useful information, reviewers need to:

- refocus the study questions (including PICOTS characteristics) on gaps in the trial evidence;
- assess the suitability of observational studies to address these questions;
- assess the potential biases, magnitude of benefits and harms, heterogeneity of effects, and random variation that may influence the results of observational studies and determine whether the magnitude of effect is such that it cannot be explained by these factors .

Refocus the study questions (including PICOTS characteristics) on gaps in the trial evidence. Specifying the PICOTS for gaps in the trial evidence guides subsequent steps in assessing whether observational studies will be helpful.

Even when trial data are insufficient, observational studies will be suitable for filling in the gaps only if they provide stronger or more applicable evidence than do available trials. For example, reviewers are commonly asked to evaluate observational studies that focus on intermediate

outcome measures such as persistency, adherence, and compliance. If the questions are clearly defined, reviewers who are familiar with the sources, designs, advantages, and limitations of observational studies can make a reasonable determination of the suitability of these study designs.

Assess the suitability of observational studies to address these comparative effectiveness questions. Consideration of the clinical context, population, natural history of the condition will help to determine the suitability of observational studies. Glasziou and colleagues considered various clinical examples to identify patient populations in which observational studies were likely to provide valid and meaningful answers to questions about efficacy (Glasziou, Chalmers et al. 2007). They found that conditions that are fluctuating or intermittent are much more difficult to assess with observational studies, particularly those without a comparison group. For example, individuals afflicted with acute low back pain often recover spontaneously; hence, a cohort study of treatments for acute low back pain cannot establish, with any degree of certainty, whether the treatments affected patient outcomes. Uncontrolled studies of interventions for diseases with stable or steadily progressing courses, however, may be useful. For example, individuals afflicted with amyotrophic lateral sclerosis (ALS) steadily decline in function, and spontaneous recovery is virtually unknown. An uncontrolled cohort study of a treatment for ALS, and a cohort study that compared treatments, may well be able to demonstrate meaningful effects.

Knowledge of the sources and designs of studies used in pharmacoepidemiology and in device and procedure registries can help inform judgments about the likelihood that observational studies would add useful information. Sources and types of observational studies used in pharmacoepidemiology are described in detail in Chapter 5.

Procedure registries are among the strongest source of data for observational studies—for example, results of observational comparison studies based on cardiac procedure registry data have been validated against trial results among patients recruited at the same time. (Holloway and Schocken 1988) In a CER comparing coronary stents with coronary artery bypass surgery, the reviewers knew that high-quality registry data would be available and used them to address gaps in the trial evidence.

Many data sources for observational studies are suited to long-term follow-up but are limited in the type of outcomes that can be measured. For example, databases that combine data from claims and laboratory, pharmacy, and clinical records usually can ascertain deaths accurately. Outcomes such as exacerbations or relapses of chronic diseases, serious adverse events, or major changes in function may be determined from proxy outcomes such as diagnoses, procedures, and health services utilization (e.g., emergency room visits, hospital admissions, discontinuation of a drug, initiation of a drug associated with treatment of a side effect, or a surgical procedure). With few exceptions, however, administrative and clinical databases lack data on quality of life, severity of symptoms, and function.

Knowledge of the sources of pharmacoepidemiologic studies can inform decisions about whether observational studies are likely to answer questions about patient populations inadequately evaluated in trials. For example, many observational studies of antipsychotic medications are

open-label extensions of clinical trials, in which participants continue to be followed for a period of time after the blinded intervention phase ended. A potential advantage of this type of study is that longer-term harms and tolerability can be evaluated. An important disadvantage is that participants followed during the extension phase are even more highly selected than participants originally enrolled in the trial. Such subjects, who tolerated and responded to a particular drug for 6 weeks, have much lower rates of discontinuation for lack of efficacy or side effects thereafter than the broader population specified in the key question.

Assess how potential biases, magnitude of benefit and harm, heterogeneity of effects, and random variation may influence the results. To decide whether observational data will provide valid and useful information, the review team should carefully consider potential biases in observational studies, the anticipated effect size for important outcomes in relationship to harms, and potential random variation in the effect size. EPCs should explicitly state their decisions on inclusion and exclusion of observational studies and carefully describe the rationale for those decisions.

Table 4-2. Determining whether observational studies can provide valid and useful information when trial data are insufficient.

Criteria	Definition	Considerations
Potential biases	Systematic error or deviation of the results from the true effect	Includes selection, detection, performance, and attrition bias
Magnitude of benefit and harms	The size of the beneficial effect compared with the size of potential adverse effects, and their relationship to the threshold for decision making	If anticipated benefits from the intervention are much greater than anticipated harms, and benefits exceed the likely threshold for decision making, observational studies may provide useful data
Heterogeneity	Variation in populations, interventions, study methods, and outcome measures	Includes known sources of variation (PICOTS characteristics) as well as sources that can't be proven or quantified
Random Variation	Random variation in effect size	

Reviewers should assess the potential biases in the observational studies in relation to the magnitude of benefits and harms, heterogeneity, and random variation in effect size (Table 4-2). If the anticipated effect size is expected to be much larger than potential biases, heterogeneity, or random variation might explain, then observational studies may provide strong evidence about effectiveness (Box 4-6). Conversely, if the anticipated effect size is very small, then bias, heterogeneity, and random variation may render observational studies useless, particularly if the anticipated effect size is near the threshold likely to be used by decision makers.

Box 4-6. Including observational studies when potential biases are unlikely to explain a large effect size

In a recent review of bariatric surgery (Shekelle PG, Morton SC et al. 2004), one of the included studies was the Swedish Obese Subjects study (SOS) (Sjostrom, Lissner et al. 1999), a matched, cohort study. Weight loss at 1 year was about 10 times greater than the weight loss reflected in the pooled estimates of pharmaceutical therapies. This large difference was judged unlikely to be entirely due to any of the possible biases, and the authors concluded, based on this study, that bariatric surgery promotes greater weight loss than diet, exercise, and pharmacotherapy. Additionally, although no direct comparative data were available (e.g., surgery compared with diet) and various types of bias may have affected results (e.g., attrition or cointerventions), this study provided real-world information on the potential benefits of this surgery.

The four main biases are selection bias, detection bias, performance bias, and attrition bias. (Higgins 2006) These biases may arise in trials that have methodologic flaws, and, conversely, they may be prevented in observational studies that take appropriate precautions against bias. The likelihood of serious bias, and its potential impact on the results, can be assessed only by a careful analysis of each study.

Selection bias. Selection bias refers to systematic differences among the groups being compared that arise from self-selection of treatments, physician-directed selection of treatments, or association of treatment assignments with social characteristics such as income, education, race, age, access to health care, social support, or literacy. The result of selection bias is that the differences among the compared groups in prognosis, likelihood of compliance, responsiveness to treatment, susceptibility to adverse effects, and the use of other interventions can distort or overwhelm the attempt to compare the effects of different interventions (Higgins 2006).

When different diagnoses, severity of illness, or comorbid conditions are important reasons that physicians assign different treatments, selection bias is called "confounding by indication" (Box 4-7). Confounding by indication is a common problem in pharmacoepidemiologic studies comparing beneficial effects of interventions because physicians often assign treatment based on their expectations of beneficial effects (Vandenbroucke 2004).

One important source of selection bias in CERs of pharmaceutical agents is the fact that new users may differ from established or prior users in treatment response. In trials, investigators know when patients started the study drug, and all benefits should be captured during follow-up. Moreover, the control group will be followed from a meaningful point in the natural history of patients' disease, facilitating interpretation of comparative benefits of a drug with respect to duration of therapy. Investigators who conduct observational studies can approximate that methodologic rigor by excluding established users of the drug and following only patients with new drug use (Ray 2003), although determining who is a 'new user' from administrative claims data can be challenging.

Systematic reviewers should look carefully for how investigators defined new use. Most investigators who conduct observational studies require a 6-month period in which a patient had no record of using the cohort-defining drug (e.g., no prescription fills in an insurance database), although briefer periods may suffice, especially for prospective cohort studies and registries. Longer periods without evidence that the patient used the cohort-defining drug probably reduce the potential for selection bias because longer periods would make it unlikely that apparent new users are actually former users returned from an extended drug holiday.

Also useful is determining whether the study authors required patients to be new users of the specific cohort-defining drug or new users of the entire class of drugs. For example, comparative cohort studies can still suffer bias when patients who fail one drug in a class switch to a newer (or different) drug in the same class. Although the patients who switched drugs appear to be new users for one of the comparative cohorts, they are not new to the entire class and the investigators may not know why they switched (e.g., insufficient biochemical response with first-line therapy). The least biased observational studies will require all patients in the cohort to be new users of the entire class of drugs related to the key question.

Box 4-7. Confounding by Indication

Carvedilol is an expensive, proprietary beta-blocker proven to reduce mortality in moderate-to-severe heart failure. A retrospective analysis of a clinical administrative database sought to compare the outcomes of heart failure patients taking carvedilol with those of patients taking atenolol, an inexpensive, generic beta blocker. However, in some health systems, carvedilol is restricted to patients who meet symptomatic and echocardiographic or angiographic criteria for moderate or severe chronic heart failure, usually requiring consultation with a prescribing cardiologist. For example, nearly all patients waiting for a heart transplant take carvedilol. Atenolol is usually prescribed by primary care physicians and its use is unrestricted. At baseline, then, the patients in the carvedilol group are more likely to have severe, chronic symptomatic heart failure and have a worse prognosis than are those taking atenolol.

Detection bias. This refers to systematic differences among the comparison groups in outcome assessment (Higgins 2006). This bias is important in cohort studies in which comparison groups may be assessed at different time points and by different and nonblinded assessors. It is particularly important in case-control studies, where subjects are entered into studies based on the measured outcome, although these study designs are less commonly encountered in CERs.

Potential sources of detection bias in observational studies used in CERs include:

- the comparison of data across different sources (e.g., different databases),
- the use of different measurement techniques or assessors across study groups,
- the use of different outcomes measures (e.g., different definitions of outcomes),
- variation in record quality across sources,
- variation in payment incentives or disincentives across study groups or assessors,
- differences in the timing of outcome measures (e.g., dating from discharge versus treatment onset),
- use of open-ended versus closed response options,

- outcomes assessors were not blinded to treatment, and
- the *a priori* delineation of outcomes and how they will be measured versus *post hoc* delineation.

Performance bias. Performance bias refers to systematic differences in the care other than the intervention under investigation provided to participants in the comparison groups (Higgins 2006). Because retrospective observational studies are virtually never double-blinded, treatment groups may differ in the expectations, information, and enthusiasm that providers and patients bring to treatment. These differences can influence behaviors, such as adherence or health practices such as diet and exercise, which can affect the outcomes of interest. Contamination (provision of the intervention to the comparison group) and cointervention (provision of unintended additional care to either comparison group) (Higgins 2006) occur more often in observational studies and are much more likely to go undetected than in trials.

Potential sources of performance bias in nonblinded observational studies in CERs include:

- comparison group participants are aware of the potential effects of the study drug;
- providers (who are not blinded to treatment groups) provide care to both treatment and comparison groups, resulting in contamination;
- other treatments are given in conjunction with the drug or device (cointerventions);
- different health care providers among study groups (e.g., specialists vs. generalists); and
- exposure was measured or determined in a different way in the groups being compared.

Attrition bias. Finally, attrition bias refers to systematic differences among the comparison groups in the loss of participants from the study and how they were accounted for in the results (Higgins 2006). The issues here are similar to those in trials.

Potential sources of attrition bias in the use of observational studies in CERs include:

- subjects who know they are taking a drug that they view as less beneficial may be more likely to drop out;
- providers who know treatment assignment could influence dropout rates; and
- providers may selectively exclude subjects from the study after allocation, based on treatment group.

5. FINDING EVIDENCE

This chapter suggests various resources for locating studies and evidence for CERs. The focus is on methods for finding clinical trials, with some information provided on observational studies. Experience suggests that searching multiple sources is necessary to avoid bias in identifying relevant studies (Crumley and Wiebe 2005). These resources may include:

- previously published systematic reviews,
- bibliographic databases,
- other Web sites and databases,
- Scientific Information Packets, and
- miscellaneous resources.

Box 5-1 Key Points (Finding Evidence)

Searches for primary studies should be extensive, otherwise reviews risk producing biased and/or imprecise estimates of effects.

To develop a thorough search strategy, reviewers and librarians should work together to identify search terms and resources to be searched.

Thorough searching can be achieved only by using a variety of search methods (both computerized and manual) and searching multiple, possibly overlapping, sources of studies.

Although the majority of searching will be undertaken at the beginning of the review, a series of updating searches may need to be scheduled to take place near the end of the project.

The search should be well documented and search results should be saved and retained for future potential reanalysis.

Previously Published Systematic Reviews

Conducting a comprehensive, *de novo* literature search may be unnecessary if other organizations have recently published a review of the topic. This review can be accepted with some degree of confidence if its methodology is documented and meets appropriate criteria for judging the quality of the review and if it has used a similar (or at least relevant) evidence model.

Several strategies for identifying systematic reviews in MEDLINE® are available (see http://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_sources.html). In addition to MEDLINE®, several databases of systematic reviews are available from the Cochrane Library (http://www.thecochranelibrary.com), including the Cochrane Database of Systematic Reviews (CDSR), the Database of Abstracts of Reviews of Effects (DARE), The Health Technology Assessment (HTA) Database, and the National Institute for Health and Clinical Excellence (NICE). Subject-relevant Cochrane Groups may be contacted for additional trials they may have accrued since their last update, and completed and in-progress AHRQ reports may also be helpful. Several reports in the literature provide guidance for optimal methods of searching for systematic reviews (Shojania and Bero 2001; Montori, Wilczynski et al. 2005).

Search filters

Another strategy in locating systematic reviews is the use of search filters, or "hedges," which are pre-tested, widely available search strategies for use with bibliographic databases. Besides systematic reviews, there are filters available for randomized controlled trials, observational studies, diagnostic studies, economic studies, etc. These filters are designed for specific

databases, such as MEDLINE, EMBASE, CINAHL, etc. They have been rigorously tested for specificity and sensitivity and can save time in the development of search strategies for CERs. The PubMed Clinical Queries database is a good example of a filter for systematic reviews. Other filters are available from the Cochrane Collaboration at http://www.cochrane.dk/cochrane/handbook/appendices/appendix_5c_example_of_a_search_strategy_for_electronic_databases.htm. The NHS Centre for Reviews and Dissemination also lists several filters at http://www.york.ac.uk/inst/crd/revs.htm. In addition there are several researchers who have published work on search strategies (Haynes 2005; Wong 2006). The Scottish Intercollegiate Guidelines Network (SIGN) at http://www.sign.ac.uk/methodology/filters.html has developed a filter for observational studies.

Bibliographic Databases

Major Databases

Historically, literature searches for randomized controlled trials (RCTs) have relied heavily on the following databases:

- MEDLINE®, EMBASE, and Cochrane CENTRAL (although Cochrane CENTRAL is the best single source for published RCTs, searching all three of these databases improves the yield of eligible trials for systematic reviews) (Royle and Milne 2003);
- other databases pertinent to specific subjects, for example, AIDSLINE, PsycInfo, and Cinahl; and
- citation tracking databases such as Web of Science or Scopus.

EPCs have considerable expertise and are familiar with recent advances (Glanville, Lefebvre et al. 2006; Sampson, Zhang et al. 2006; Wilczynski and Haynes 2006; Zhang, Ajiferuke et al. 2006) in identifying relevant studies from the above databases focusing on benefits. However, identifying relevant studies of harms in bibliographic databases can be challenging. Broad search strategies that are not restricted by terms for adverse events can be very inefficient, but they may be the best approach when the overall size of the literature for an intervention is relatively small. More restrictive search strategies based on general terms for adverse events can miss relevant citations, because about one-quarter of relevant studies do not include adverse events indexing terms or text words in the title or abstract (Derry, Kong Loke et al. 2001). Searches that include terms for specific adverse events can be useful if the adverse events are known beforehand, but this tactic can also be problematic because the same adverse event is often described using a wide range of terms (e.g., fatigue, somnolence, weakness, lethargy, or central nervous system effects).

Box 5-2. Harms subheadings

For MEDLINE:
- /adverse effects
- /poisoning
- /toxicity
- /chemically induced
- /contraindications
- /complications

For EMBASE:
- /side effect
- /adverse drug reaction
- /drug toxicity
- /complication

Development of efficient methods for identifying studies of harms is an ongoing area of research (Golder, McIntosh et al. 2006). Strategies based on a combination of approaches—including using terms for specific adverse events, "floating" subheadings for adverse events (searched without being attached to an indexing term), and text words for adverse events—appear to be

highly sensitive (> 97 percent), but are inefficient (< 3 percent specificity) (Golder, McIntosh et al. 2006). The Cochrane Handbook suggests some subheadings (see **Box 5.2**) that may be useful for identifying studies reporting adverse events. (Loke, Price et al. 2007) Specific recommendations for sensitive but more efficient search strategies are not currently available.

Online repositories of full-text articles.

While MEDLINE is continuously updated, lag time between publication and indexing is a major limitation, and indexing time varies among journals. Some versions of MEDLINE (e.g., Ovid) also experience a lag between the time a citation is indexed and the time it is added to the database. In Ovid, these references can be accessed in the database "Ovid MEDLINE(R) Daily Update." The National Library of Medicine (NLM) temporarily halts indexing of all journals in November and December. During that time, searching in-process citations in MEDLINE is essential using text words that appear in the title and abstract.

The value of searching other bibliographic repositories of articles such as Google Scholar (http://scholar.google.com/), Highwire Press (http://highwire.stanford.edu/lists/freeart.dtl) and Journals@OVID has not been assessed formally. The potential advantages of these databases over in-process MEDLINE citations is the ability to search the full-text of the article and to link to other articles of interest, particularly references.

Open Access Journals.

Open access sites such as the Public Library of Science (PloS at http://www.plos.org/) and Biomed Central (http://www.biomedcentral.com/) are indexed in MEDLINE and elsewhere, but may provide full-text not available elsewhere. In addition, PLoS has just launched its "PLoS Hub for Clinical Trials," which will gather clinical trials published in all PLoS publications. Other open-access sites (though not strictly healthcare subject matter) such as the University of Michigan-based OISter (http://www.oaister.org/), offer "freely available, previously difficult-to-access, academically-oriented digital materials." The Directory of Open-access Journals (DOAJ) (http://www.doaj.org/) currently has over 800 open-access journals listed.

Other Web Sites and Databases

CER review teams need to consider sources that may not be used routinely in other types of systematic reviews. These may include proprietary trial registries (Song, Fry-Smith et al. 2004; Crumley and Wiebe 2005); regulatory sites, such as the FDA; and pharmacoepidemiologic information.

Several sources have been identified that provide information on unpublished trials or unpublished data from published trials. These are:

- clinical trial results databases;
- government regulatory sites, such as the FDA;
- pharmacoepidemiologic databases; and
- other sites.

Clinical trial results databases

Online trial registries may include results of completed but unpublished clinical trials, although the focus has been on early registration of proposed or in-progress trials. An early evaluation of broad search strategies found trial registries to be useful in identifying studies eligible for inclusion in systematic reviews (Savoie, Helmer et al. 2003). In 2004, Song and colleagues provided a detailed description of six online clinical trial registries and assessed their usefulness in identifying unpublished results (Song, Fry-Smith et al. 2004). Since 2004, the number of registration and results databases has increased. These resources can be helpful in identifying otherwise unreachable trials and in providing additional details of trials that have been published (Crumley and Wiebe 2005).

ClinicalTrials.gov (http://clinicaltrials.gov), the U.S. Government's data bank of clinical trials, contains records for more than 9,800 completed trials and 9,350 studies that are no longer recruiting. At present, ClinicalTrials.gov does not publish study results, but some completed trials do list citations of trial results. Published results from registered trials may be identified by searching the phrase **"**ClinicalTrials.gov number**"** in the abstract field. It is also possible to search 'ClinicalTrials.gov' in the secondary source ID field in PubMed and combine it with topic statements.

The Food and Drug Administration Amendments Act of 2007, enacted in September, 2007, mandates expansion of ClinicalTrials.gov to include results of completed trials of approved drugs and devices. This expansion will probably not be complete until 2008 or 2009.

Current Controlled Trials (http://www.controlled-trials.com/) was established to promote the exchange of information worldwide, and allows searching across multiple clinical trial registers, including the National Health Services (NHS) in England, ClinicalTrials.gov, and direct access to BioMedCentral. In addition, PubMed can now be configured to automatically link published reports of clinical trials to the study protocol in the *Current Controlled Trials* database using the "LinkOut" utility (http://www.controlled-trials.com/news/linkout.asp).

At **Clinicalstudyresults.org** (http://www.clinicalstudyresults.org) individual drug manufacturers contribute trial information on selected drugs; this Web site is sponsored by the pharmaceutical industry. Records indicate whether the study has been indexed in MEDLINE® and provides citations. Some unpublished study results are provided full-text. Many individual drug companies have posted trial registries on their Web sites, however they are not standardized and the quality can vary.

Government regulatory sites

Center for Drug Evaluation and Research (CDER), run by the FDA, is an important site for CER research (http://www.fda.gov/cder/). Several different sections of the site provide potential access to published and unpublished trials.

As required under the Freedom of Information Act, the FDA provides detailed information about the trials submitted in support of a new drug approval application on Drugs@FDA (http://www.accessdata.fda.gov/scripts/cder/drugsatfda/). The site may be searched by drug name or active ingredient. If available, approval documents include internal clinical and statistical reviews submitted by a manufacturer in support of its product, including results of completed but

unpublished trials. These documents help identify publication bias even when complete methodological details of unpublished trials are not available (Bennett and Jull 2003). Unfortunately, this information is not accessible prior to drug approval and is often unavailable even at the time the drug is approved. And, when the reviews are provided, some may be heavily redacted for proprietary reasons. When they are available, however, the documents enable reviewers to compare results of published and unpublished trials, and to compare the material published in journals with the material submitted to the FDA. The labeling information on the Web site might also provide information about the number of unpublished trials.

Dockets for FDA advisory panel meetings at http://www.fda.gov/ohrms/dockets/default.htm can also provide relevant data. There are usually PowerPoint presentations and backgrounders on the clinical trials. Transcripts may include discussion of unpublished trial data that was presented during the hearings.

The CDER site also lists any post-marketing study commitments (http://www.fda.gov/cder/pmc/default.htm) that are conducted after the FDA has approved a product for marketing (e.g., studies requiring the sponsor to demonstrate clinical benefit of a product following accelerated approval). Unfortunately, the studies are not available on the site, but the commitments are listed and noted if they are in process or completed.

For FDA information about devices, pre-marketing approval documents (http://www.fda.gov/cdrh/pmapage.html#search) including an "Approval Letter and Summary of Safety and Effectiveness" are available from 1995-present, and include a summary of clinical trials.

European Medicines Agency (EMEA) is another potential source for unpublished trials is the European Medicines Agency (http://www.emea.europa.eu/htms/human/epar/a.htm). The Scientific Discussion section for approved drugs provides summaries of clinical trials performed worldwide.

Pharmacoepidemiologic Databases

A working knowledge of the sources and designs of studies used in pharmacoepidemiology is essential to make judgments about the likelihood that observational studies add useful information to a CER. Pharmacoepidemiologic studies examine the use and effects of drugs in large numbers of people (Strom 2005). Data sources for such studies include: large administrative datasets; registries based on a particular diagnosis or on exposure to a specific procedure or drug; prospectively designed safety monitoring systems; and clinical trials. Examples of the uses for pharmacoepidemiologic studies include: exploratory analysis of large administrative databases to identify unsuspected adverse effects, rates of discontinuation or treatment failure, and longer-term observational follow-up of patients who had been enrolled in an RCT (Strom 2005).

Pharmacoepidemiologic studies use a variety of observational study designs: cohort studies, case-control studies, nested case-control studies, case-crossover studies, and case-time-control studies (Etminan 2004; Etminan and Samii 2004). Advantages of pharmacoepidemiologic research are that analyses of large, representative populations can be available at relatively low cost and with little delay. Databases used in pharmacoepidemiologic research, however, often

lack information on important clinical factors and may be more prone to confounders and misclassification of exposures (Schneeweiss and Avorn 2005). By contrast, traditional observational studies such as the Framingham Study or the Nurses' Health Study, obtain information directly from patients. Such data have the capacity to be much more detailed than information from administrative databases, although cohort studies may be more subject to recall or other biases.

Practice networks, particularly in primary care, are an important source for pharmacoepidemiologic studies (Santaguida, Helfand et al. 2005). Data collected by practice-based research networks in community settings provide information about benefits and harms of health care interventions that is more applicable to everyday clinical practice than "traditional" research based in academic centers (Lindbloom, Ewigman et al. 2004). Practice-based research data sets are often richer in clinical detail than are large administrative databases, making it possible to identify and measure exposures, outcomes, and likely confounders with more confidence (Jollis, Ancukiewicz et al. 1993).

One large, well-known example of a practice network is the U.K.-based General Practice Research Database (GPRD) (Wood and Martine 2004). A recently published study of suicide risk associated with selective serotonin reuptake inhibitors based on GPRD data reported no clear association with increased risk in adults (Martinez, Rietbrock et al. 2005). Although this finding is similar to those of meta-analyses of RCTs (Fergusson, Doucette et al. 2005; Gunnell, Saperia et al. 2005), additional implications from the GPRD analysis are that the risks of suicide are not significantly higher in clinical practice, where patients may not be as closely monitored or be as highly selected as in clinical trials.

How to identify pharmacoepidemiologic studies efficiently while avoiding bias is a subject of intense research (Lemeshow, Blum et al. 2005; Fraser, Murray et al. 2006; Furlan, Irvin et al. 2006; Kuper, Nicholson et al. 2006). Searching for specific drug names and classes, along with the names of specific adverse effects, can be effective but it can also be inefficient. Using such a strategy to search for large observational studies of 66 adverse effects, Papanikolaou and colleagues had to examine 18,198 abstracts to find 15 eligible studies (Papanikolaou, Christidi et al. 2006). Even though a large number of abstracts were identified, there is no way to evaluate the sensitivity or specificity of such a search.

Another approach is to search citation abstracts, full-text online repositories, or selected Web sites using terms for the large databases and practice networks most commonly used for pharmacoepidemiologic research. Examples of search terms include "The Health Improvement Network" (THIN), "GPRD" (for the UK General Practice Research Database), TennCare for "Tennessee Medicaid," "Integrated Primary Care Information Project," and "Pharmaco-morbiditeitskoppeling" (Lewis, Schinnar et al. 2006) (Stricker and Psaty 2004).

In addition to the sources listed above, Chou and Helfand (Chou and Helfand 2005) recommend including pharmacokinetic and pharmacodynamic data and case reports, while Bennett et al. recommend postmarketing surveillance databases (Bennett, Nebeker et al. 2005).

Other sites

The National Guidelines Clearinghouse (http://www.ngc.gov) may link to documents used in the development of the guideline. Conference databases, such as OCLC's ProccedingsFirst and PapersFirst, CSA's Conference Papers Index, or BIOSIS Previews may mention clinical trials, as well as conference abstracts cited in bibliographic databases. Various databases for theses and dissertations may also be useful. A search engine based in France—Exalead (http://www.exalead.com/search) pulls together a wide variety of resources for each search.

Scientific Information Packets

AHRQ is interested in identifying as many studies as possible that are relevant to the questions for each of its CERs. When a list of drugs or devices to be included in a CER is put into final form, AHRQ invites pharmaceutical companies and device manufacturers to submit Scientific Information Packets (SIPs), or scientific information about their product. Manufacturers are specifically invited to submit a list of all known, completed RCTs of their product. Guidelines for submission of such material from stakeholders are available on the Effective Health Care Web site (http://effectivehealthcare.ahrq.gov/submitData.cfm?submittype=submit).

Packets vary with each company, but usually include information on studies submitted to the FDA for approval (with updates), summaries of observational studies, and full text of selected studies. They often include a bibliography of potentially relevant articles.

SIPs submitted to the Effective Health Care program (through the EHC Web site or any other means) are considered to be public. The program does not solicit confidential information from manufacturers because the public must be able to scrutinize all evidence that is included in a CER.

Miscellaneous Resources

EPC review teams routinely supplement bibliographic database searches by reviewing reference lists and soliciting experts for additional citations. Identifying topic-specific databases and specific citations is a major role of the technical expert groups convened for each topic (see Chapter 2).

Patent information, press releases from pharmaceutical companies, and other product information can also be helpful, since they often publicize the progress or completion of their clinical trials. This information is searchable on proprietary aggregators such as Lexis, PRN Newswire, and Scopus, and on on pharmaceutical company web sites. EPCs have made use of grey literature, particularly in reviews of therapeutic devices and surgical interventions (Hartling, McAlister et al. 2005). EPCs are not required to search these sources routinely, but may find them helpful in specific instances.

6. ASSESSING THE QUALITY AND APPLICABILITY OF INCLUDED STUDIES

The purpose of assessing the quality of individual studies is to inform a judgment about the validity of a study's results. Assessing the quality of studies is a critical element of making judgments about the overall strength of a body of evidence about a given key question (Chapter 11).

Box 6-1. Key points (Assessing Quality and Applicability)

Quality rating is part of assessing the risk that a study is biased.

Applicability and quality should both be evaluated.

To assess quality, use predefined criteria and apply them thoughtfully.

Use 3 levels (good, fair, and poor) to rate the overall quality of individual studies.

Trials meeting criteria for effectiveness trials should be identified and highlighted.

Features that may limit applicability of individual efficacy studies should be noted in evidence tables. Formal rating of applicability of individual studies is not required.

For key outcomes or comparisons in CERs, important features limiting applicability of a body of studies should be summarized in a table.

Quality can be defined as "the extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error" (Lohr 2004). Thus, assessing the quality of a study can be thought of as assessing the risk that the results reflect bias in study design or execution rather than the true effect of the interventions under study.

We use the term *applicability* to describe a separate set of concerns that have been variously referred to as external validity, relevance, or generalizability (Shadish, Cook et al. 2002). Although some systems include issues of applicability (i.e., "external validity") as a component of quality, we recommend keeping these considerations separate when assessing individual studies. In part, this is because the importance of applicability depends on the context in which the evidence is being used. A study may have very limited applicability for answering a broad policy question but be very applicable for answering a specific clinical question relating to patients similar to those enrolled in the study.

This chapter describes general procedures for assessing quality and applicability. Considerations for assessing the quality of studies of harms and quality issues related to quantitative synthesis of results are described in more detail in Chapters 8 and 9, respectively.

Stages in Rating Quality of Studies

Rating the quality of individual studies takes place in a series of stages:

Stage 1. Classify the study design by answering the following questions:

1. Is the study a synthesis of several individual studies? If not:

2. Is the study comparative?
3. Did investigators assign the exposure?
    a. If so, was the intervention allocated randomly? Was randomization done at the individual level?
    b. If not, was more than one group of subjects studied? Were exposure and outcome assigned at the same time? Were groups assigned by exposure or by outcome?

Based on the answers to these questions, most studies can be classified as:

- review, systematic review, or meta-analysis;
- trials—RCT or other controlled clinical trial;
- observational, comparative studies—cohort study, case-control study, cross-sectional study;
- before-after or interrupted time series; or
- noncomparative study.

Stage 2. Apply predefined criteria for quality and critical appraisal. Rating systems can be categorized as scales, checklists, or checklists with a summary judgment (Sanderson, 2007). Some systems, such as that of the US Preventive Services Task Force, are designed to be used with more than one type of study (Downs & Black, 1998; Harris, Helfand et al, 2001). Systems for rating the quality of observational studies may use different criteria for cohort studies, case-control studies, and other specific study designs. They may also distinguish between studies of prognosis or risk factors and studies of causal relationships.

A wide variety of tools to assess methodologic quality are in use (Bhandari, Devereaux et al. 2002; Moja, Telaro et al. 2005; Mallen, Peat et al. 2006). The wide variation indicates the lack of empirical evidence demonstrating the superiority of any one system. Despite this variation, however, there is more agreement than disagreement on the most important, or "core" criteria for each type of study (West, King et al. 2002; Lohr 2004). For example, in a recent inventory, 92 percent of tools to assess the quality of observational studies assessed methods for selecting study participants, 86 percent assessed methods for measuring study variable and design-specific sources of bias, and 78 percent assessed the appropriate use of statistics and methods to control for confounding (Sanderson, Tatt et al. 2007).

The most efficient strategy for CERs is to use a generic system, applying criteria to trials as well as to observational studies as applicable. Several core elements apply to trials as well as observational studies:

- similarity of groups at baseline in terms of baseline characteristics and prognostic factors,
- extent to which valid primary outcomes were described,
- blinding of subjects and providers,
- blinded assessment of the outcome,
- intention-to-treat analysis,
- differential loss to follow-up between the compared groups or overall high loss to follow-up, and
- conflict of interest.

For trials, two additional elements are important:

- methods used for randomization and
- allocation concealment.

For observational studies, yet another set of elements should be considered:

- sample size;
- methods for selecting participants (inception cohort, methods to avoid selection bias);
- methods for measuring exposure variables;
- methods to deal with any design-specific issues such as recall bias, interviewer bias, etc.; and
- analytical methods to control confounding.

Stage 3. Arrive at a summary judgment of the study's quality. CERs should use three categories to indicate the summary judgment of the quality of individual studies (Box 6-2). In the methods section, CERs should refer to Box 6-2 to clarify the definitions of good, fair, and poor. The "fair" category is broad, and the majority of studies will likely receive this rating. For that reason, reviewers may wish to point out important distinctions among individual studies in this category. At present, however, no empirical justification exists for creating additional categories.

Box 6-2. Three Summary Ratings of Quality of Individual Studies

Good (low risk of bias). These studies have the least bias and results are considered valid. A study that adheres mostly to the commonly held concepts of high quality including the following: a formal randomized controlled study; clear description of the population, setting, interventions, and comparison groups; appropriate measurement of outcomes; appropriate statistical and analytic methods and reporting; no reporting errors; low dropout rate; and clear reporting of dropouts.

Fair. These studies are susceptible to some bias, but it is not sufficient to invalidate the results. They do not meet all the criteria required for a rating of good quality because they have some deficiencies, but no flaw is likely to cause major bias. The study may be missing information, making it difficult to assess limitations and potential problems.

Poor (high risk of bias). These studies have significant flaws that imply biases of various types that may invalidate the results. They have serious errors in design, analysis, or reporting; large amounts of missing information; or discrepancies in reporting.

The summary quality rating may differ for different outcomes within the same study. For example, some trials use a blinded assessment of final status for some outcomes, but not for others.

Critical appraisal—applying the principles of clinical epidemiology—is a key component of assessing the validity of a study. Many systematic reviews make the mistake of failing to report adequately or use quality assessments in the analysis and interpretation of results (Moja, Telaro et al. 2005; Norris and Atkins 2005). Especially when the rating is "poor" or "fair," reporting a rating without explanation can confuse or even antagonize readers, and CER authors should avoid this by providing a brief rationale for, at least, any poor rating.

Rating a study as "poor" simply because it did not nominally meet every criterion for quality is not acceptable. Instead, identifying a flaw in the design or execution of a study should prompt an assessment of the potential consequences of that flaw. During the data extraction process, investigators should record comments relating to potential sources of bias and other study limitations. Such comments should be included in the evidence tables or in the text of the report.

Studies rated poor may be included or excluded from the main synthesis (qualitative or quantitative) but authors should examine and report whether results are altered by excluding or including poor studies. One reason to include poor studies is if they provide some specific valuable information not available from higher quality studies. Examples include key questions about subgroups for which no good or fair studies are available or studies that give the only information about a clinically important comparison. Authors should make clear, however, the reason for including a given poor study if they have excluded others.

Rating Applicability

Because applicability depends on the specific review question being addressed and the users' needs, it is difficult to devise a uniform scale for rating applicability of individual studies. Several investigators have outlined series of questions or checklists for assessing applicability. Rothwell delineated 39 factors that may lead to ungeneralizable RCT results (Rothwell, Slattery et al. 1996). Bornhoft and others described a checklist for considering internal validity, external validity and "model validity" (Bornhoft, Maxion-Bergemann et al. 2006). Glasgow and Green have developed a checklist for potential reviewers and journal editors to assess the reporting of characteristics critical to external validity (Glasgow, Green et al. 2006). We found no empiric data validating any scorings system for assessing applicability across a range of studies. We do not, therefore, recommend the use of a scale to rate applicability of either an individual study or a body of evidence.

The factors affecting applicability will differ for different types of interventions (for example, drugs vs. devices) and for different outcomes (for example, benefits vs. harms). The most common factors affecting applicability of drug studies are overly restricted study populations (including restrictions to promote adherence), inappropriate comparison therapies, or insufficient duration of follow-up. Common factors affecting applicability of studies of surgery and invasive devices include restriction to clinical settings with high levels of expertise and exclusion of a high proportion of prospective patients due to age, comorbidities, or other factors.

We recommend assessing potential threats to applicability following the PICOTS format used to summarize study characteristics. Table 6-1 below summarizes specific issues that may limit applicability for trials. Although many of these issues relate to enrolment criteria and are more relevant to trials, they also apply to non-randomized studies where the study populations or interventions may not always reflect typical practice. These features should routinely be abstracted into evidence tables and highlighted when they appear likely to affect applicability.

Table 6-1 Features of Individual Studies that Affect Applicability

	Features that should always be abstracted	Conditions That May Limit Applicability
Population	Eligibility criteria and proportion of screened patients enrolled	Narrow eligibility criteria and high exclusion rate
	Demographic characteristics (range and mean): Age, gender, race and ethnicity	Large differences between demographics of study population and that of patients in the community
	Severity or stage of illness	Narrow or unrepresentative severity or stage of illness
	Run in period (for drugs); if reported, include attrition before randomization and reasons (non-adherence, side-effects, non- response) (Charlson and Horwitz 1984; Davis, Applegate et al. 1995)	Run in period with high-exclusion rate for non-adherence or side effects
	Event rates in treatment and control groups	Event rates much higher or lower than observed in population-based studies
	Prevalence of disease (for diagnostic studies)	Disease prevalence is study population higher than expected for target population for diagnostic test
Intervention	Dose, duration, and cointerventions	Doses or schedules not reflected in current practice Intensity of behavioral interventions that is not likely to be feasible for routine use Co-interventions that are likely to modify effectiveness of therapy
	Adherence (e.g., monitoring, frequent contact)	Monitoring practices or visit frequency not used in typical practice
	Training and expertise -- selection process, training and skill of intervention team (for surgery/ technical interventions/ diagnostics).	Highly selected intervention team or level of training/proficiency not widely available
Comparator	Dose and schedule of comparator, if applicable	Inadequate dose of comparison therapy

	Whether comparator is the best available alternative to the treatment under study	Use of sub-standard alternative therapy
Outcomes	Clinical benefits on relative and absolute scale Individual harms and how defined, on relative and absolute scale	Surrogate rather than clinical outcomes; failure to measure most important outcomes Failure to distinguish minor from serious adverse effects
Timing of outcomes measurement	Timing of follow-up	Follow-up too short to detect important benefits or harms; lack of long-term follow-up for interventions requiring long-term interventions
Setting	Geographic setting Clinical setting (specialty vs. primary care setting)	Settings where standards of care differ markedly from setting of interest Specialty population or level of care that differs importantly from that seen in primary care

No study can perfectly replicate all the conditions of interest to decision makers. A thoughtful systematic review must do more than summarize how the populations, interventions, or outcomes in the available studies may differ from the questions of interest. Considering applicability also requires making judgments about whether the differences between the available evidence and the "ideal" evidence are likely to alter the observed effectiveness or safety of the intervention, and in which direction. Subgroup analysis may help reveal factors that influence effect size for benefits or harms. For example, evidence that effect sizes vary with age or disease severity calls into question the applicability of evidence that comes largely from younger, healthier subjects.

Whether differences between the features of the available studies and current practice are large enough to threaten applicability is in many cases a judgment, and the input of experts and stakeholders can be useful. Clinical experts can provide insight into whether changes in the dosing or timing of a drug or modifications to a medical device are likely to alter benefits or safety. Stakeholders such as a professional society or a health plan medical director may be able to judge whether an intervention employed in a study is feasible and relevant to their members.

Applicability depends heavily on context—studies that are highly applicable for one stakeholder such as a Medicaid director may not be relevant to another, for example a health plan enrolling Medicare beneficiaries. Thus, we do not recommend assigning an overall summary grade for applicability. Table 6-2 illustrates how a review might summarize the evidence and implications for applicability, to help inform different stakeholders about the limitations of the evidence and to assist in rating the strength of evidence. Because many comparative effectiveness reviews

involve numerous interventions, comparisons, and outcomes, the aim of such a table is not to characterize all of the evidence but to highlight the most important areas where applicability is a concern.

Table 6-2 Summary Applicability Table

	Describe Available Evidence	**Describe Implications for Applicability**
Population	Describe general characteristics of enrolled populations. Where possible, describe the proportion with important characteristics (e.g., % over age 65) rather than the range.	Describe how enrolled populations differ from target population and how this might affect risk of benefits or harms
Intensity or quality of treatment	Describe the general characteristics of interventions	Describe how studied interventions compare to those in routine use and how this might affect risk of benefits or harms
Choice of, and dosing of, the comparator	Describe the comparators used	Describe whether comparators reflect best alternative treatment and how this may influence treatment effect size
Outcomes	Describe what outcomes are most frequently reported	Describe whether measured outcomes are known to reflect most important clinical benefits and harms
Timing of follow-up	Describe range of follow-up	Describe whether follow-up is sufficient to detect clinically important benefits and harms

7. ASSESSING DIAGNOSTIC TECHNOLOGIES

This chapter is deliberately omitted.

8. HARMS

To be useful to decision makers and to generate balanced results and conclusions, comparative effectiveness reviews should address both benefits and harms (Ernst and Pittler 2001; Atkins, Best et al. 2004; GRADE Working Group; Loke, Price et al. 2007). CERs should assess harms that are important to decision makers and users of the intervention under consideration. Technical Expert Groups (TEGs) are a valuable resource for helping to set the priorities in evaluating harms in CERs.

Box 8.1. Key Points (Harms)

- Assess all important harms, whenever possible.
- Use multiple sources of information to identify harms.
- Gather evidence on harms from a broad range of sources, including observational studies, particularly when clinical trials are lacking; when generalizability is uncertain; or when investigating rare, long-term, or unexpected harms.
- Do not assume studies adequately assess harms because methods used to assess and report benefits are appropriate; rather, evaluate how well studies identify and analyze harms.
- Avoid inappropriate combining of data of harms, and thoroughly investigate inconsistent results.
- Be cautious about drawing conclusions on harms when events are rare and estimates of risk are imprecise. When describing evidence on harms, avoid using terms implying causality when causality is uncertain.
- Do not draw conclusions about equivalence and non-inferiority for harms unless there is appropriate data justifying such statements.
- Avoid assuming that class effects on harms are present for two or more interventions unless there is a strong pathophysiologic rationale as well as supporting clinical outcomes data (such as similar estimates of risk). Include analyses of inconsistency when combining data on harms from two or more interventions.
- Avoid implicit indirect comparisons when judging comparative risk of adverse events. Rather, evaluate whether different sets of trials meet assumptions for similarity of treatment effects, and if so, perform "formal" indirect comparisons if possible.

High-priority harms should routinely include the most serious adverse events, but they may also include common adverse events or other specific adverse events salient to clinicians or patients. In many cases, evidence on rare but important harms may be unavailable or available only from sources of evidence highly susceptible to bias (Loke, Price et al. 2007). In these situations, CERs should clearly present and critically discuss the limitations or gaps in the evidence

Assessing harms in CERs can be challenging for several reasons. Most clinical trials focus on assessing benefits; measuring or reporting harms is typically a lesser or secondary consideration. In addition, reviewing evidence for all possible harms may not be feasible, as interventions may be associated with dozens of potentially important adverse events. EPCs may often face a major tradeoff between increasing comprehensiveness and decreasing quality of harms data (McIntosh, Woolacott et al. 2004; Loke, Price et al. 2007).

Adequately assessing harms requires EPCs to consider a broad range of data sources; for that reason, they need to deal with other challenges such as choosing

which types of evidence to include, identifying relevant studies of harms, assessing their quality, and summarizing and synthesizing data from different types of evidence.

Terminology

Terminology related to reporting of harms is poorly standardized (Ioannidis, Evans et al. 2004). This can cause confusion or misleading conclusions. EPCs should strive for consistent and precise usage of terminology when reporting data on harms in their CERs.

For example, the term "harms" is generally preferred over the term "safety" because the latter sounds more reassuring and may obscure important concerns. "Harms" is also preferable to the term "unintended effects," which could refer to either beneficial or harmful outcomes. Terms that do not imply causality (such as "adverse events") should be the default term to describe harmful events, unless causality is reasonably certain.

Definitions for commonly used terms for harms reporting are summarized here, along with suggested usage. They are adapted from definitions published by the Cochrane Collaboration, the CONSORT Group, and the World Health Organization Uppsala Monitoring Centre (Edwards and Aronson 2000; Ioannidis, Evans et al. 2004; Loke, Price et al. 2007).

- ***Adverse effect***: A harmful or undesirable outcome that occurs during or after the use of a drug or intervention for which there is at least a reasonable possibility of a causal relation.
- ***Adverse event***: A harmful or undesirable outcome that occurs during or after the use of a drug or intervention but is not necessarily caused by it. When causality is uncertain or establishing causality is the purpose of the CER, it should generally by the default term over "adverse effect" or "adverse reaction/adverse drug reaction."
- ***Adverse reaction/adverse drug reaction (ADR):*** An adverse effect specifically associated with a drug.
- ***Complications:*** A term often used to describe adverse events following surgery or other invasive interventions.
- ***Harms:*** The totality of all possible adverse consequences of an intervention.
- ***Passive surveillance of harms:*** Recorded adverse events are those that study participants spontaneously report on their own initiative. In ***active surveillance of harms***, participants are asked about the occurrence of specific adverse events in structured questionnaires or interviews, or predefined laboratory or other diagnostic tests are performed at prespecified time intervals.
- ***Risk-benefit ratio:*** A common expression for the comparison of overall harms and benefits. However, because benefits and harms of an intervention are usually very different in character and are measured on different scales, a true 'risk-benefit ratio' is rarely calculable. In addition, there may be several distinct benefits and harms. A preferred term is ***balance of benefits and harms***.
- ***Safety:*** Substantive evidence of an absence of harm. The term is often misused when evidence on harms is simply absent or insufficient.

- ***"Serious" adverse events***: Any adverse event with serious medical consequences, including death, hospital admission, prolonged hospitalization, and persistent or significant disability or incapacity.
- ***"Severe" adverse events:*** The intensity of an adverse event (including "nonserious" adverse events). For example, a rash could be "severe" but not "serious" (i.e., not resulting in death, hospital admission, prolonged hospitalization, or persistent or significant disability).
- ***Side effects:*** Unintended drug effects (beneficial or harmful) that occur with doses normally used for therapeutic effects. Use of this term may tend to understate the important of harms because the word "side" may be perceived to suggest secondary importance.
- ***Tolerability:*** This term is often used imprecisely but should be used to refer to a patient's or subject's ability or willingness to tolerate or accept unpleasant drug-related adverse events without serious or permanent sequelae.
- ***Toxicity:*** A general term used to refer to drug-related harms. This term may be most appropriate for describing laboratory-determined abnormalities, although it is also used to describe clinical adverse events. The disadvantage of the term "toxicity" is that is implies causality. When causality is uncertain, the terms "abnormal laboratory measurements" or "laboratory abnormalities" may be more appropriate.

Sources of Evidence on Harms

Randomized Controlled Trials

Published Trials. As noted in previous chapters, properly designed and executed randomized controlled trials (RCTs) are considered the "gold standard" for evaluating efficacy because they minimize potential bias. However, relying solely on published randomized trials to evaluate harms in CERs is problematic for several reasons. First, most randomized trials lack prespecified hypotheses for harms (Ioannidis, Evans et al. 2004). Rather, hypotheses are usually designed to evaluate beneficial effects, with assessment of harms a secondary consideration. As such, the quality and quantity of harms reporting in clinical trials has consistently been found to be inadequate (Edwards, McQuay et al. 1999; Ioannidis and Lau 2001; Loke and Derry 2001; Papanikolaou, Churchill et al. 2004).

Second, because randomized trials can be expensive to carry out, few have large enough sample sizes or are long enough in duration to assess uncommon or long-term harms adequately (Ray 2003; Dieppe, Bartlett et al. 2004; Vandenbroucke 2004). Further, most randomized trials are explanatory, rather than pragmatic, in design—i.e., they assess benefits and harms in ideal, homogeneous populations and settings (Rothwell 2005). Even when harms are appropriately assessed and reported, such "efficacy trials" have limited ability to assess harms in individuals encountered in everyday practice.

Third, few randomized trials directly compare alternative treatment strategies. Although CER authors can indirectly compare the efficacy of two treatment strategies based on trials in which they are compared with a common third treatment (usually placebo), the results of indirect comparisons do not always agree with direct comparisons (Bucher, Guyatt et al. 1997; Song, Altman et al. 2003; Chou, Fu et al. 2006).

Fourth, publication and selective outcomes reporting bias can lead to distorted conclusions about harms when data are unpublished, partially reported, downplayed, or omitted (Easterbrook, Berlin et al. 1991; Sterne, Egger et al. 2001; Chan, Hrobjartsson et al. 2004; Whittington, Kendall et al. 2004; Ridker and Torres 2006).

Fifth, in some cases, relying on randomized trials for information about harms is impossible. For example, surgical procedures and medical devices often become widely disseminated with few or no randomized trial data. The same can be true for older therapeutic devices, such as hyperbaric oxygen chambers (McDonagh, Helfand et al. 2004).

Despite their limitations, RCTs are the basis for regulatory approval and advertising and other claims made on behalf of drugs and other interventions. For this reason, CERs must address them in detail when they are available. Head-to-head RCTs provide the most direct evidence on comparative harms. However, placebo-controlled RCTs are often more plentiful than head-to-head trials and may provide important information on absolute risks as well as the most robust estimates of risk. For example, in a systematic review evaluating myocardial infarction risk associated with celecoxib, 60 percent of the myocardial infarctions occurred in two placebo-controlled trials (Kearney, Baigent et al. 2006). In addition, risks associated with celecoxib were not apparent in head-to-head trials against most other nonsteroidal anti-inflammatory drugs (NSAIDS), which were also associated with increased risk. CERs should generally routinely include placebo-controlled trials for assessment of harms, particularly for rare or uncommon adverse events.

Unpublished Supplemental Data. In addition to evaluating results of published RCTs, CERs should also consider supplementing published results with unpublished ones as well as unpublished data from published trials. Such information has several potentially valuable uses:

1. to assess the number of unpublished trials or outcomes, which can help in evaluating risk for publication or outcomes reporting bias,
2. to evaluate whether conclusions based on unpublished data are qualitatively different than those based on published RCTs, and
3. to conduct formal quantitative meta-analysis including published and unpublished RCTs or outcomes.

Unpublished Trials. Unpublished clinical trials tend to report lower estimates of treatment benefit than do published trials (i.e., stronger intervention effects) (McAuley, Pham et al. 2000; Egger, Juni et al. 2003). The impact of unpublished trials on assessments of harms has not been extensively studied, but a recent systematic review of antidepressants in children found that addition of data from unpublished trials changed the balance of risks and benefits from favorable to unfavorable for several drugs (Whittington, Kendall et al. 2004). In a systematic review of cardiovascular risk associated with rosiglitazone, 27 of 42 included trials were unpublished (Nissen and Wolski 2007). Excluding unpublished trials would have decreased the precision in estimates of increased myocardial risk (relative risk = 1.43; 95% CI, 1.03 to 1.98), possibly resulting in loss of statistical significance.

Two main drawbacks of using data from unpublished trials, assuming they are available, should be considered. One is that, frequently, evidence is insufficient to assess fully the risk of bias. Another is that results and conclusions of trials may change between initial presentation of data and publication in a peer-reviewed journal (Rosmarakis, Soteriades et al. 2005; Toma, McAlister et al. 2006).

Unpublished data from published trials. Journal publications may omit important information because of space limitations or other reasons (Sterne, Egger et al. 2001; Ridker and Torres 2006). Drug approval information—especially the clinical and statistical reviews prepared by staff of the US Food and Drug Administration (FDA)—frequently provides details about harms not included in journal publications.

For example, the Celecoxib Long-term Arthritis Safety Study (CLASS), a major trial of celecoxib, was published in *JAMA* as a 6-month study and reported finding fewer gastrointestinal adverse events for celecoxib than for two nonselective NSAID comparators (diclofenac and ibuprofen) (Silverstein, Faich et al. 2000). The *JAMA* article did not mention that some patients in the trial had been observed for longer than 6 months (Hrachovec and Mora 2001). In contrast, the FDA review reported all the outcomes data, including no difference in gastrointestinal adverse events at the end of follow-up (Witter 2000).

As another example, for a major trial of rofecoxib (Vioxx Gastrointestinal Outcomes Research Study, or VIGOR), an FDA statistical review made available to the public in 2001 has six pages of analysis on the issue of cardiovascular risk (Lee); the *New England Journal of Medicine* publication had three lines (Bombardier, Laine et al. 2000). In fact, before publication of VIGOR, myocardial infarctions were omitted from most published reports of trials evaluating selective or nonselective NSAIDs because an association with cardiovascular events was not suspected. A recent systematic review obtained unpublished myocardial infarction data from sponsoring pharmaceutical companies; it found an increased risk with high doses of all evaluated NSAIDs (selective or nonselective) other than naproxen (Kearney, Baigent et al. 2006). An analysis of myocardial infarction risk based on only published data would be seriously compromised by incomplete data.

Limited evidence suggests an inverse relationship between the proportion of included trials reporting outcomes and the estimates of benefit (Furukawa, Watanabe et al. 2007). How the proportion of included trials reporting outcomes affects estimates of harms is not clear, particularly when pooled estimates are not statistically significant. Nonetheless, when a significant proportion of trials included in a CER fail to report an important or critical adverse event, investigators should consider efforts to obtain unpublished data (e.g., by querying study authors, funding sources, clinical trials registries, or FDA documents).

Observational Studies

Observational studies are almost always necessary to assess harms adequately. Including such studies in evidence reports has long been the standard of practice for the EPCs. Although observational studies are more susceptible to bias than well-conducted clinical trials, they can be particularly useful when sufficient effectiveness, head-to-head, long-term, or sufficiently large randomized trials (for uncommon adverse events) do not exist (Vandenbroucke 2004). Observational studies may also provide the best (or only) evidence for evaluating harms in

minority or vulnerable populations (such as pregnant women, children, or elderly patients with multiple comorbidities) who are underrepresented in clinical trials.

The term "observational studies" refers to a broad range of study designs. These include case reports; retrospective analyses of large claims or practice-based databases; population-based, longitudinal cohort studies; uncontrolled series of patients receiving surgery or other invasive interventions; and others (Kleinbaum, Kupper et al. 1982). All can yield useful information.

The types of observational studies in a CER will vary depending on the type or frequency of adverse events being evaluated. The choice of study designs to examine harms also depends on whether investigators are seeking to test a hypothesis or to generate new ones. Different types of observational studies might be included or rendered irrelevant by availability of data from stronger study types.

Cohort and Case-Control Studies. Investigators should routinely consider including well-designed observational studies, such as case-control and population-based cohort studies. Such studies are well suited for testing hypotheses on whether one intervention is associated with a greater risk for an adverse event than is another and for quantifying the risk. Although they are also subject to confounding and biases that are encountered less commonly in RCTs, they take stronger precautions against bias than do other observational designs, and their strengths and weaknesses are better understood. For unexpected adverse events, for example, confounding by indication may not be as important an issue as when evaluating beneficial or intended effects because they are usually not associated with the reasons for choosing a particular treatment (Psaty, Koepsell et al. 1999; Stricker and Psaty 2004; Vandenbroucke 2004).

A recent report found that large observational studies usually report smaller absolute risks of harm than do large randomized trials (Papanikolaou, Christidi et al. 2006). There was no clear predilection for randomized trials or observational studies to estimate greater relative risks. In more than one-half of the comparisons assessed, estimates of relative or absolute risk varied more than twofold. Discrepancies between randomized trials and observational studies may occur because of differences in populations, settings, or interventions; differences in study design; differential effects of biases; or some combination of these factors.

Observational Studies Based on Analyses of Large Databases. Pharmacoepidemiologic studies using large databases are increasingly common, and they may be very valuable for comparing the risk of uncommon adverse events. Nonetheless, additional empirical research is needed to identify features of pharmacoepidemiologic studies that are associated with valid findings. In some cases, data from large administrative databases may be supplemented or verified by more detailed clinical information. Regardless of how data are collected, all observational studies should employ appropriate methods for minimizing bias and misclassification of data.

Case Reports and Postmarketing Surveillance. About 30 percent of the primary published literature on adverse drug events is in the form of case reports (Aronson, Derry et al. 2002). Case reports can be useful for identifying uncommon, unexpected, or long-term adverse events, particularly for new drugs or other interventions (Stricker and Psaty 2004). The adverse events

identified by case reports often differ from those detected in clinical trials (Loke, Derry et al. 2004). However, case reports are generally considered hypothesis-generating because calculating information from them about the frequency or comparative risk of adverse events is difficult.

The FDA receives about 280,000 reports of postmarketing adverse events annually and collects them into a database (Strom 2004), and issues information about adverse drug events on its MedWatch website (http://www.fda.gov/medwatch/. Although pharmaceutical companies and other investigators may also perform high-quality analyses on postmarketing data, such analyses are not always made public in a timely fashion, as in the case of the withdrawn lipid-lowering drug cerivastatin (Psaty, Furberg et al. 2004). Active, hypothesis-driven postmarketing surveillance systems have also recently been developed for identifying and evaluating serious adverse drug events (Bennett, Nebeker et al. 2005).

Case reports and other hypothesis-generating studies are probably most useful for CERs evaluating new drugs suspected of being associated with serious but uncommon adverse events. For other CERs, investigators may consider their inclusion on a case-by-case basis.

Other Observational Studies. Several other types of observational studies may also report data on harms. However, they are likely to be more prone to bias than are RCTs, and their use needs to be considered cautiously. For example, studies reporting harms from surgical or other invasive interventions often consist of a series of patients who received the procedure. Data are often insufficient to assess the methods used to select participants (Oleson 1999). In addition, because such studies lack control groups, evaluating effects of confounding or selective outcome reporting bias on outcomes is impossible, as is comparing risks of adverse events across interventions.

Other nonrandomized study designs may not evaluate populations more applicable to routine practice than the ones enrolled in randomized trials. Open-label extensions of clinical trials are one example. Although they are designed to follow patients for an extended period of time, they also usually evaluate a more highly selected population (patients who completed the randomized trial, tolerated the medication, and agreed to participate in the extension), are open-label and often lack a comparison arm. Because such studies generally offer few advantages over randomized trials, they usually can be excluded from CERs if more reliable long-term, comparative data are available. If they are included in CERs, their limitations should be described clearly.

Criteria to Select Observational Studies for Inclusion. In general, many more observational studies than randomized trials will be available for nearly all health care interventions. Evaluating a large number of observational studies can be unmanageable when conducting a CER, especially when a significant proportion either do not add useful information or carry a high risk of reporting biased results.

Several criteria have commonly been used in systematic reviews and CERs to screen observational studies of harms for inclusion. Empirical data are lacking on how use of different selection criteria affects estimates of harms. However, CERs should clearly describe any

selection criteria they use along with the rationale for choosing the criteria (e.g., to assess long-term harms or populations not covered well in trials). Commonly used inclusion criteria for observational studies include:

1. studies meeting certain study design definitions,
2. studies not exceeding a certain risk of bias threshold,
3. studies meeting a certain threshold for duration of follow-up,
4. studies meeting a sample size threshold, and
5. studies evaluating a specific population of interest.

Pharmacokinetic, Pharmacodynamic, and Pharmacogenomic Studies

When evaluating harms in CERs of adverse events, EPCs should consider whether specific interventions are more likely to be harmful for specific populations than for other populations. CERs should focus on studies that deliberately look at the risks and benefits of specific drugs in subgroups. In many situations, however, the risks in different populations can be difficult to address systematically because clinical trials and observational studies exclude certain groups of vulnerable patients or do not adequately analyze harmful effects in subgroups.

When clinical data on subpopulations are lacking, systematic reviewers may consider including pharmacodynamic, pharmacokinetic, or pharmacogenomic studies, even though such data do not always correlate with clinical outcomes. In the case of the lipid-lowering agent rosuvastatin, for example, the FDA required labeling indicating that drug levels are higher in Asians and could potentially lead to more adverse events in this population, even though a recently published meta-analysis of trials submitted to the FDA found no differences in clinical adverse events according to ethnicity, sex, or age (Shepherd, Hunninghake et al. 2004). Pharmacokinetic studies may also provide useful information on drug-drug interactions. One role of systematic reviews, however, is to help distinguish concerns based on clinical data from what is based on pharmacologic properties or on other considerations. In this way, CERs can highlight important areas for future clinical research. If included in CERs, pharmacokinetic, pharmacodynamic, and pharmacogenomic studies are likely best considered hypothesis-generating.

Assessing Risk of Bias (Quality) of Harms Reporting

Randomized Trials

Features distinguishing higher-quality randomized trials are similar regardless of whether benefits or harms are being assessed. These include use of adequate randomization sequence generation and allocation concealment techniques; blinding of participants, providers, and outcomes assessors; and analysis according to intention-to-treat principles (Juni, Altman et al. 2001). However, because evaluating harms is often a secondary consideration in randomized trials, quality of harms assessment and reporting can be inadequate even when assessment of the primary (beneficial) outcome is appropriate.

Systematic reviewers should pay particular attention to how withdrawals and dropouts are handled in any analyses of adverse event rates (Wood, White et al. 2004). Dropouts are exposed to the intervention or assessed for a shorter period of time than are persons who completed the trial. However, when calculating rates of adverse events, they are usually analyzed as if they remained in the study for the whole duration (e.g., "We analyzed all patients receiving at least

one dose of the drug"). If dropout rates are small, this may have only minimal effects on estimates of adverse events. However, when dropout rates are higher, this could lead to underestimates of risk of harms, though such concerns may be alleviated if analyses are performed to assess potential effects of dropouts.

Similarly, the timeframe and relationship to drug exposure may be critical for evaluating harms data. In a recent FDA report on suicide risk associated with SSRIs, adverse events were counted only if they occurred while the patient was on active medication (FDA 2006). If a patient discontinued treatment for any reason and an event occurred more than 1 day later, it was not counted. Given the high discontinuation rates with antidepressants, this analytic approach could lead to serious underreporting of suicidal events, particularly if significant carryover effects from the drugs are present, or if suicidal thoughts are associated with discontinuation of therapy. On the other hand, patients are usually unblinded following discontinuation of treatment, making interpretation of subsequent adverse events challenging.

When evaluating the quality of harms assessment, EPCs should also consider whether adverse events are prespecified and defined. Accurate and complete assessment of common or expected adverse events is more likely when the outcomes are clearly defined *a priori*. Although this criterion will not be met for unanticipated adverse events, studies reporting such outcomes can be very valuable for identifying previously unrecognized harms. Data on specific defined adverse events are also likely to be more accurate and informative than are generic statements, such as "no adverse events were noted" or "the interventions were well tolerated." If a specific adverse event is not reported, it is generally safer for systematic reviewers to assume that they were not ascertained or not recorded than to assume that the prevalence or incidence was zero (Loke, Price et al. 2007). Trials should also predefine the qualifiers "serious" and "severe" to describe adverse events. Otherwise, it is impossible for readers to determine whether these labels were applied consistently within and across trials.

Standardized criteria for grading severity of adverse events are available for certain conditions (NCI 1999; NIAID 2004; NCI 2006). CERs should note when grading severity or seriousness of adverse events is based on nonstandardized or poorly defined criteria, as such classifications may not be comparable across studies and may be less reproducible than classifications based on standardized definitions.

"Withdrawals due to adverse events" is commonly reported in trials and often used in systematic reviews as a marker for intolerable or severe adverse events. However, the Cochrane Handbook (version 4.2.5) suggests caution in interpreting withdrawals attributed to adverse events in this manner, for the following reasons (Loke, Price et al. 2007).

1. Attribution of reasons for discontinuation is likely to be imprecise and to vary across trials. Reasons for discontinuation include mild but irritating side effects, lack of efficacy, nonmedical reasons, severe or serious adverse events, or any combination of these factors.
2. Pressure to keep dropouts low in trials may result in rates that do not reflect real-world practice.

3. Unblinding often takes place before the decision to withdraw, which can lead to distortion of estimates of an intervention's effect on withdrawal (e.g., symptoms are less likely to lead to withdrawal if the patient is found to be on placebo).

Nonetheless, withdrawals due to adverse events are often reported even when serious or severe adverse events are not reported or are poorly defined, and they may provide some useful information.

Reviewers should also consider what methods were used to ascertain adverse events. Several studies have shown that active methods for identifying adverse events (such as querying patients using a checklist or standardized laboratory tests) are more likely to identify adverse events than passive methods such as relying on patient self-report (Olsen, Klemetsrud et al. 1999; Bent, Padula et al. 2006). In drug trials, use of an independent external endpoint committee may provide less biased estimates of harms than are outcomes assessment performed by investigators connected to the study (Juni, Nartey et al. 2004; Sydes, Spiegelhalter et al. 2004).

Observational Studies

Because observational studies lack randomization, they should adhere to higher methodological standards to be considered valid (Egger, Schneider et al. 1998; Lawlor, Davey Smith et al. 2004). Randomized controlled trials are expected to have outcomes recorded by blinded personnel and to include all participants who were randomized in the analysis of results. Use of blinded outcome assessors and an inception cohort (e.g., "new users") is at least as important when assessing observational studies.

Instruments for assessing risk of bias in observational studies vary greatly in scope, in the number and types of items used, and in developmental rigor (Deeks, Dinnes et al. 2003). Further study is needed to determine which methodological shortcomings are consistently associated with bias in assessment and reporting of harms. In addition, none was specifically designed to assess quality of harms assessment and reporting.

Some consensus exists on the major domains that should be considered when evaluating the overall validity of an observational study. For cohort studies, for example, important factors include assembly of an inception cohort, complete follow-up, appropriate assessment of potential confounders, accurate determination of exposures and outcomes, and blinded assessment of outcomes (West, King et al. 2002; Deeks, Dinnes et al. 2003).

Several systematic reviews have empirically evaluated effects of specific methodological characteristics on estimates of harms from observational studies. They found that prospective or retrospective design (Rothwell, Slattery et al. 1996; Dalziel, Round et al. 2005), case-control compared with cohort studies (Ofman, MacLean et al. 2002; Juni, Nartey et al. 2004), and smaller compared with larger case series (Dalziel, Round et al. 2005) had no clear effects on estimates of harms. Two studies found that industry-funded studies tended to report more favorable outcomes than did nonindustry-funded studies (Juni, Nartey et al. 2004; Shah, Albert et al. 2005). Because all of these studies generally evaluated fairly limited samples of studies, wider applicability of their findings to other datasets and interventions is uncertain.

Observational studies based on evaluations of large administrative databases should follow the same general principles to reduce bias as observational studies that directly collect data from patients. In these cases, reviewers should pay particular attention to the methods used for ascertaining exposures and outcomes and for measuring and analyzing potential confounders, as these issues are more likely to be problematic in studies relying on administrative claims (although not unique to them) (Schneeweiss and Avorn 2005).

For all observational studies, estimates of harms are less likely to be confounded when evaluating unintended or unknown adverse events than when evaluating known or intended effects. NSAIDS are a case in point. Before publication of VIGOR, prescribing of cyclo-oxygenase-2-selective versus nonselective NSAIDs was unlikely to be influenced by considerations about patients' risk for myocardial infarction. Clinicians were more likely to prescribe selective NSAIDs in patients at higher risk for gastrointestinal bleeding, as this was a well-known risk of nonselective NSAIDs. Because such patients are at higher risk of developing gastrointestinal bleeding independent of drug use, this led to the appearance of an apparent association between selective NSAID use and bleeding in epidemiologic studies (Laporte, Ibanez et al. 2004). In some cases, such spurious associations may remain despite adjustment for known confounders ("residual confounding").

Uncontrolled Studies

Studies of surgery, devices, and noninvasive interventions are often uncontrolled series of patients who received the therapy and then were followed prospectively over a period of time. Such studies can provide information about rates of adverse events in clinical practice, and they may be most informative when the background rate of such events in untreated patients is low. Unfortunately, such studies frequently do not meet standards for accurate and comprehensive reporting of complications (Martin, Brennan et al. 2002). In addition, as in other types of clinical research, authors are more likely to submit for publication studies showing the best outcomes.

For some interventions, reviewers must consider including uncontrolled studies for assessment of adverse events, as little or no other evidence may be available. Adapting risk of bias criteria for other types of observational studies from West et al. (West, King et al. 2002), Carey and Boden have proposed several criteria for evaluating risk of bias in case series (Carey and Boden 2003):

- clearly defined question,
- well-described study population,
- well-described intervention,
- use of validated outcome measures,
- appropriate statistical analyses,
- well-described results,
- discussion/conclusions supported by data, and
- funding source acknowledged.

In addition, the Cochrane Non-Randomised Studies Methods Group suggests that the initial, critical risk-of-bias criterion that reviewers can (or should) apply to any nonrandomized study is whether the study sample was systematically selected (i.e., enrollment or attempted enrollment of all patients meeting prespecified inclusion criteria) (Oleson 1999). Without unbiased

selection of subjects, determining how representative even well-described results may be is impossible.

Instruments for Assessing Risk of Bias (Quality) in Studies on Harms

Development of instruments for assessing risk of bias specifically in studies of harms is still in an early stage of development. Two points remain unclear at present: whether to use a specific rating instrument to evaluate harms assessment and reporting, or whether using instruments for rating the overall risk of bias of a study is sufficient (as long as particular attention is paid to how well adverse events are defined, ascertained, and reported).

Chou et al. empirically developed and tested a quality-rating instrument for assessing quality of harms assessment and reporting in randomized trials and observational studies (cohort studies and uncontrolled surgical series) of carotid endarterectomy for symptomatic carotid artery stenosis (Chou, Fu et al. 2006). This approach involved four criteria: nonbiased selection, low loss to follow-up, adverse events prespecified and defined, and adequate duration of follow-up. Studies meeting at least three of the four criteria reported a rate of postsurgical complications of 5.7 percent (95% CI, 4.8 percent to 6.6 percent) compared with 3.7 percent (95% CI, 3.1 percent to 4.3 percent) for studies meeting fewer than three such criteria: The generalizability of this instrument to other interventions is unclear, as it did not predict differences in estimates of risk of myocardial infarction associated with rofecoxib (Chou, Fu et al. 2006).

Santaguida et al. have also developed a quality-rating instrument (McHarm) for evaluating studies reporting harms (Santaguida and Raina May 2005). The tool was developed from items generated by a review of the literature on harms and previous quality assessment instruments. A formal Delphi consensus exercise was used to reduce the number of items. The subsequent list of quality criteria specific to harms was tested for reliability and face, construct, and criterion validity. This quality-assessment tool is intended for use in conjunction with another standardized quality-assessment tool that captures design-specific internal validity issues. However, the association between quality scores on the McHarm and differences in summary estimates from meta-analyses has not yet been evaluated.

Case reports may provide valuable information about the possibility of rare or previously unrecognized adverse events. Of 47 case reports published in 1963 in four major general medical journals, one study 25 years ago judged that 35 of them were subsequently proved to be "clearly" correct (Venning 1982). However, the methods used to determine reliability of case reports in this study were subjective, and results have not been replicated. A recent study, in fact, found that only 18 percent of case reports of suspected adverse drug reactions have been subjected to rigorous evaluation in subsequent studies (Loke, Price et al. 2006). A statistical modeling study suggested that the likelihood of more than one to three spontaneously reported cases is very unlikely to be coincidental when the adverse event is rare or uncommon (Begaud, Moride et al. 1994).

Several disease-specific (Danan and Benichou 1993; Maria and Victorino 1997) and nondisease-specific (Naranjo, Busto et al. 1981; Michel and Knodel 1986) methods for assessing the probability of causality from case reports of adverse events have been proposed. These methods

represent expert opinion and have not been validated empirically. Factors believed to increase the likelihood of causality include:

- temporal relationship (exposure preceding adverse event and adverse event appearing at an appropriate time interval after exposure);
- lack of alternative causes;
- drug levels in body fluids or tissues;
- resolution or improvement after discontinuation;
- dose-response relationship;
- recurrence following rechallenge (that is, restarting the drug to see whether the adverse reaction recurs) (Benichou, Danan et al. 1993); and
- confirmation of adverse event by objective information.

Guidelines for improving the reporting of suspected adverse drug events in case reports (similar to CONSORT guidelines for reporting harms in randomized trials) have recently been proposed (Aronson 2003). In 35 reports of 48 patients published in *BMJ,* the median number of recommended items that were reported was nine (range 5-12) of 19, although effects of missing information on the validity of case reports have not been studied.

Synthesizing Evidence on Harms

The following issues (also discussed in Chapter 9) are especially relevant for analysis of adverse events:

Meta-analysis for uncommon or rare adverse events

A common problem in randomized trials and systematic reviews is interpreting a nonsignificant probability value, for testing a "superiority" hypothesis, as indicating no difference in risk for rare adverse events, particularly when the confidence intervals are wide and encompass the possibility of clinically important risks (Goodman and Berlin 1994; Jonville-Bera, Giraudeau et al. 2006). For example, one trial concluded that, in patients with meningitis, "treatment with dexamethasone did not result in an increased risk of adverse events" compared with placebo for treatment of hyperglycemia, herpes zoster, or fungal infection because *P* values for all three outcomes were > 0.20 (de Gans and van de Beek 2002). However, the 95% confidence intervals for estimates of relative risks for these three conditions encompassed clinically significant increases in risk (-13.5 percent to 77.6 percent, -60.4 percent to 377.7 percent, and -43.6 percent to 496.2 percent, respectively). A more meaningful analysis would perhaps acknowledge the lack of statistical power to assess risk adequately and include an interpretation of confidence intervals, including the possibility or probability of excess harm.

Pooling data on harms from different populations or interventions

In the case of drugs, class effects (similar benefits or harms across different drugs or interventions) are often assumed because of similar chemical structures or mechanisms of action, but such pathophysiologic rationales can be misleading. (McAlister, Laupacis et al. 1999) Clinical trials directly evaluating comparative risks are the most useful source of evidence for determining whether a class effect is present and whether data from interventions can be appropriately pooled. If systematic reviewers choose to pool results for two or more

interventions, they should clearly state their basis for assuming a class effect. Statistical tests for heterogeneity can be helpful for evaluating situations when assumptions about a class effect should be rejected, though lack of statistical heterogeneity does not necessarily mean that class effects are present.

In some cases, EPCs may consider including evidence on harms from populations other than those evaluated for benefits (Loke, Price et al. 2007). An advantage of including studies of other populations for harms is that this approach increases power to evaluate uncommon adverse events. A disadvantage is that it can complicate assessments of the balance of the risk and benefit because risk may vary across different populations.

The decision to include studies of harms from other populations should depend on whether they are known, or are likely, to differ systematically from the population included for assessment of benefits in risk for adverse events. Another factor to consider is whether relative estimates of the harms being evaluated are likely to differ depending on the indication for initiating the therapy of interest. In such cases, data on harms from different populations should not be combined. Reviewers should clearly indicate when data are from mixed or disparate populations and, at a minimum, include an analysis for heterogeneity to test assumptions about similarity of risk for harms across the populations.

Equivalence and noninferiority

Systematic reviewers should draw conclusions about "equivalence" or "noninferiority" of competing interventions with regard to harms only when appropriate data justify such statements (Piaggio, Elbourne et al. 2006). In fact, few reviews will have the statistical power to assess adequately the noninferiority of competing interventions when the risk of an adverse event is on the order of 1 percent or lower. For example, Ware and Antman showed that about 100,000 patients would have been needed in the COBALT or GUSTOIII trials to rule out an excess relative death rate of five percent with 80 percent power (Ware and Antman 1997). Smaller event rates would require even greater sample sizes.

Combining data from different types of studies

Most CERs will include data on harms from different types of studies. Although methods for combining data from different types of studies are being developed (Wald and Morris 2003), statistical combination of data from observational studies is often inappropriate and should be avoided unless there is clear rationale to do so, which should be reported if such analyses are undertaken (Egger, Schneider et al. 1998).

Discrepancies between randomized trials and observational studies

A separate challenging situation is when results from randomized trials and observational studies are discordant. Potential reasons for discrepancies between randomized trials and observational studies include the following:

- differences in study risk of bias;
- differences in applicability (study populations, interventions, or settings);
- differences in methods used to define or measure outcomes;
- differential effects of publication or selective outcomes reporting bias; and

- differential effects related to funding source (observational studies less likely to be funded by industry) (Papanikolaou, Christidi et al. 2006).

A reasoned analysis of potential sources of discrepancy is more helpful than simply presenting the different results.

Reporting Evidence on Harms

As when reporting evidence on benefits, CERs should emphasize the most reliable information for the most important adverse events. Summary tables should generally present data for the most important harms first, with more reliable evidence preceding less reliable evidence. Evidence on harms from each type of study should be clearly summarized in summary tables, narrative format, or both (GRADE Working Group 2004).

Elements to focus on include the following:

- descriptions of important factors related to risk of bias (quality) assessment (study design, number of studies, study quality, consistency of evidence, directness of evidence, and other modifying factors);
- issues related to applicability (population characteristics, interventions, comparisons, and outcomes);
- results (number of patients and absolute and relative estimates of risks);
- assessments regarding the likelihood of publication bias or incomplete outcomes data (e.g., when an adverse event is only assessed in a small subgroup of studies); because many observational studies evaluate patients from the same database, possible effects of "double counting" should also be assessed (McGettigan and Henry 2006); and
- additional analyses, if performed (e.g., sensitivity analyses, subgroup analysis, meta-regression).

CERs should emphasize the most reliable information for the most important adverse events. Another critical role of CERs is to report clearly on the limitations of the evidence on harms and to analyze thoughtfully how these limitations may affect estimates of the balance of benefit and harm.

9. QUANTITATIVE SYNTHESIS

Meta-analysis is the most commonly used statistical method for quantitative synthesis in comparative effectiveness reviews (CERs). It combines the results from two or more studies and if used appropriately, is a powerful tool to summarize results from multiple studies, provides insights into heterogeneous studies, and assists in deriving meaningful conclusions. The purposes of a meta-analysis include:

- improving the power to detect a small difference if the individual studies are small,
- improving the precision of the effect measure,
- comparing the efficacy of multiple drugs within a drug class or evaluating the consistency and differences in effect measures across study characteristics,
- helping to settle controversy arising from conflicting studies or generating new hypotheses to explain these conflicts, and
- gaining insights into statistical heterogeneity in effect sizes.

Some potential disadvantages of meta-analysis arise from combining dissimilar studies or unrepresentative studies. Other pitfalls arise from an incorrect choice of a statistical model and from biases that can be introduced by statistical procedures for pooling, particularly when there are few studies or few events. In this chapter, we describe approaches to avoid or mitigate potential problems. We address meta-analyses conducted by using aggregated or summary data at the study level. Patient-level meta-analyses and meta-analyses for diagnostic tests will not be considered here.

A plan for meta-analysis should start when key questions are formulated. There should be a compelling reason to pool studies. In the methods section of a CER, the authors should state explicitly the purpose of the meta-analysis, demonstrate that they considered the potential disadvantages, and outline the plan for exploring heterogeneity.

This next section of this chapter discusses when to combine studies, focusing on common situations in which the decision can be difficult. Whenever investigators decide to combine studies, they must choose an effect measure and a statistical model, and choose a method to explore heterogeneity; the second, third, and fourth sections, respectively, discuss these decisions. These steps are summarized in Appendix 9-1. The remaining sections address combining studies of mixed designs, sensitivity analysis, and interpreting and presenting a meta-analysis.

When to Combine Individual Studies

Box 9-1. Key Points (When to Combine Studies)

Variability should be categorized into three types: clinical diversity, methodological diversity and statistical heterogeneity.

Clinical and methodological diversity should not be ignored. Conclusions should not be drawn based on inconsistent results across studies.

When statistical tests indicate there is statistically significant heterogeneity, studies can still be combined unless there are systematic differences among studies or the studies are too heterogeneous to produce a meaningful combined estimate.

A common criticism of meta-analysis is that studies that are too heterogeneous are combined. Even if a group of studies meets the criteria for a carefully formulated research question, substantial variability among studies is often observed. In deciding whether to combine studies, the most important considerations are whether the studies asked similar questions and whether the study populations are similar enough to yield a meaningful result when they are combined.

Unfortunately, no commonly accepted standard exists for "similar enough." Judgment of the similarity among studies depends on the scope of the research question. A more general question may allow more variation among studies than a more focused question. For example, it sometimes makes sense to combine studies from a class of drugs instead of a particular drug -- if the drug class in general is of interest, where the included studies were conducted in similar populations in a similar manner, and the drugs in the class affect the outcome in question through similar mechanisms.

Decisions to combine studies should also be based on thorough investigation of variability among studies, which can be categorized into three types (Higgins and Thompson 2002):

1. Variability in study population characteristics, interventions and outcomes is considered *clinical diversity*.
2. Variability in study design and quality, such as blinding and concealment of allocation, is considered *methodological diversity*.
3. Variability in the observed treatment effects being evaluated in different trials is considered *statistical heterogeneity.*

Clinical and methodological diversity across studies is common. A wide variety of factors, such as evolving disease, evolving diagnostic criteria, change in standard care, time-dependent care, difference in baseline risk, and dose-dependent effects may cause seemingly similar studies to be different. Diversity in clinical characteristics will cause statistical heterogeneity if the true treatment effect varies depending on those characteristics. Methodological diversity can also cause statistical heterogeneity in the observed treatment effects. In this case, statistical heterogeneity suggests that the studies are not all estimating the same effect, but it does not necessarily mean that the true treatment effect varies. In particular, heterogeneity associated solely with methodological diversity would indicate that the studies suffer from different degrees of bias.

Statistical tests of heterogeneity help analysts to identify variation among effect estimates (see *Exploring Heterogeneity* below). However, basing the decision to combine or not on statistical tests alone is ill-advised. Cochran's Q is the standard test for statistical heterogeneity among studies. This test has low power to detect heterogeneity when the number of studies is relatively low or when individual studies are small; it is sensitive for detecting unimportant heterogeneity when the number of studies is high (Hardy and Thompson 1998). Because of its low power, a P-

value of 0.10 instead of 0.05 is routinely used to determine statistical significance (Higgins and Thompson 2002).

In addition to the test for heterogeneity, measures have been developed to quantify the magnitude of heterogeneity. The most easily interpretable one is I^2, which expresses the percentage of between-study variability attributable to heterogeneity rather than sampling error (chance). I^2 ranges between 0 and 100 percent; a value greater than 50 percent may be considered substantial heterogeneity (Higgins and Thompson 2002; Higgins, Thompson et al. 2003).

Even when these statistical measures do not suggest significant heterogeneity, combining studies with very diverse outcomes produces results that are difficult to interpret. For example, the outcome of mood symptoms may include depression, negative mood, anxiety, feelings of panic, tearfulness, or irritability; moreover, mood symptoms can be measured using various instruments. Because these outcomes do not measure the same aspects of mood, combining the studies wouldn't make sense.

Because tests for heterogeneity are insufficient to determine similarity among studies, ultimately the decision to combine study results or not may be subjective and qualitative. Investigators can minimize the potential for bias in these decisions by recognizing certain situations that arise frequently in conducting CERs and making consistent decisions in those situations. The remainder of this section describes approaches to some common situations.

Combining a small number of studies

No general rule exists to decide the minimum number of studies for a meta-analysis. The main issue lies in the interpretation of the results. Few will argue with the reliability of a meta-analysis of two mega-trials (10,000 patients or more). However, a meta-analysis of two RCTs with a total of 37 patients will not produce a reliable estimate. Even if the combined estimate is statistically significant, the confidence interval is likely to be very wide and could change dramatically with the addition of more studies. Thus, the results of meta-analyses of a small number of studies should be interpreted cautiously or meta-analysis should be deferred until more studies are available.

Combining studies to examine an idiosyncratic adverse event

If an adverse event is idiosyncratic and unrelated to patient characteristics, combining studies across different diseases or dissimilar populations may make sense. For example, if the outcome of interest is allergic skin rash secondary to penicillin use, trials may be combined irrespective of the type or severity of disease in patients receiving penicillin. In this example, the assumption is that there is no relationship between allergic rash and disease. This would not be true, of course, for all skin reactions and all diseases (e.g., a Jarisch-Herxheimer reaction is specific to syphilis).

This approach is most credible when there is prior reason to believe that the adverse effect is not related to underlying patient characteristics and where pooling clinically diverse studies would provide power to detect and quantify a relatively uncommon effect than cannot be detected in individual studies. In another example, Bohlius and colleagues conducted a meta-analysis to compare the effect of erythropoietin and darbepoetin on hematologic responses, red blood cell transfusions, thromboembolic events, and overall survival by combining patients with various

types of cancer (Bohlius, Langensiepen et al. 2005). Combining studies enrolling different cancer patients for hematologic outcomes and red blood cell transfusions is reasonable as long as the hematologic response is not related to cancer type—this may apply to different solid tumors but not to combining solid tumors and hematologic malignancies. Combining various cancer patients for thromboembolic event and overall survival isn't appropriate, however, because of the association between both mortality and thrombotic risk vary substantially with different cancer types.

Combining studies when selective reporting is suspected

Sometimes, the outcome of interest is reported in only a minority of otherwise eligible studies and is incompletely reported, or not reported at all, in the remaining studies. Providing a summary of a highly biased sample of studies may be misleading

In particular, adverse events are usually not as systematically reported as efficacy and effectiveness measures. (Gotzsche 1989; Hayashi and Walker 1996; Ioannidis and Lau 2001; Ioannidis and Lau 2002). Empirical evidence suggests that such nonreporting (and incomplete reporting) may be selective and guided by the nature of the findings for the pertinent outcomes (Chan, Hrobjartsson et al. 2004; Chan and Altman 2005; Chan, Krleza-Jerić et al. 2005). Secondary outcomes are more likely to be inconsistently or incompletely reported. Although primary outcomes are almost always reported, some studies may report only favorable outcomes at a selected time point.

Reviewers have two options in dealing with this problem. They can refrain from presenting a summary estimate that is likely to be misleading and qualitatively summarize the available evidence. Even without a pooled estimate, however, users may try to get an overall estimate using a crude calculation, which may be as misleading as a meta-analysis summary estimate. Alternatively, reviewers may perform a meta-analysis of the available data and provide a summary estimate, along with appropriate cautions that this result may have limited applicability. If analysts suspect selective nonreporting, they may use a sensitivity analysis to assess how biased the summary estimate may be. Such calculations may be done by modeling the process of selective nonreporting (Williamson and Gamble 2007) and are similar to sensitivity analyses for the impact of publication bias and other selection biases.(Copas and Jackson 2004) However, Williamson's method has not been validated with real (nonsimulated) data.

As noted above, the Food and Drug Administration Amendments Act of 2007 mandates inclusion of results of trials of approved drugs and devices in the U.S. government's clinical trial registry, http://www.clinicaltrials.gov. The legislation gives the National Institutes of Health until September, 2008, to implement this requirement. When it is implemented, routine searching of Clinicaltrials.gov will improve our ability to compare published results with the protocol of the study as originally planned. This will provide a more objective way to determine if trials selectively reported only some of their intended outcomes.

Combining studies that use composite outcomes

Composite outcomes need to be viewed carefully in meta-analysis. Composite outcomes bring together two or more events to be considered as a single outcome. The events could be from the same domain—for instance, cardiovascular events such as cardiovascular mortality, non-fatal

myocardial infarction, and revascularization. They also can be from different domains with a common cause—for example, a composite endpoint of adverse drug events may include gastrointestinal effects and headache. Finally, they may reflect a common endpoint caused by competing factors—for example, all-cause mortality following coronary artery bypass includes perioperative deaths as well late cardiac deaths (which may be reduced by surgery).

Two situations need to be distinguished here: (1) meta-analysis of composite outcomes reported by the primary studies and (2) meta-analysts creating composite outcomes out of individual outcomes reported by primary studies. In a meta-analysis, one should consider only composite outcomes that are generally agreed upon and in wide usage by the primary studies. Here, creating *de novo* composite outcomes should be avoided.

A composite outcome has the advantage of better statistical power, but it has to make clinical sense. Analysts evaluating the appropriateness of using a composite outcome must take the research question into consideration. A composite outcome with events from the same domain may be justifiable in certain cases, as when included studies reported rare but related adverse events. By contrast, a composite outcome with events from different domains is generally avoided. A statement that an intervention reduces a composite outcome of cardiovascular mortality, myocardial infarction, and revascularization is appropriate if the intervention has similar effect on each of these events. Conversely, it is misleading if revascularization procedures were more common outcomes than were death or infarction, or if the intervention had a large apparent treatment effect on revascularization but not on death or infarction (Freemantle and Calvert 2007).

Combining studies with different comparators

Even when populations and the intervention under study are homogeneous, comparators may not be. For example, trials might use a "usual care" comparator, a specified comparison therapy, or a placebo group comparator. Not only are these distinct kinds of control groups, but "usual care" may differ across settings and countries or over time. The assumption that all comparators are similar enough to combine needs to be carefully considered.

In another situation, a co-intervention is added in all comparison arms of some studies but no co-interventions (or different co-interventions) are added in other studies. For example, anticoagulation might be added to both arms in a trial that compares drug-eluting stents with bare metal stents. More generally, the group of trials comparing A vs. B could be described as one group of studies of A + X vs. B + X; others with A+Y vs. B+Y, and so on. Summarizing studies with different comparisons makes the implicit assumption that no interactions occur between the common added components X or Y and any of the interventions of interest. This assumption needs to be evaluated before quantitative synthesis. This type of interaction applies to evaluation of harms as well.

A special, common case is when the comparators are different drugs in the same class. This situation is discussed briefly in the context of pooling studies of adverse effects (**Chapter 8**). Analysts need to consider both the similarity of the comparator drugs and their dosing before deciding to pool different trials.

Choice of Effect Measures

Box 9.2 Key Points (Choice of Effect Measures)

For dichotomous outcomes, relative effects metrics (i.e., relative risk, odds ratio) are generally more appropriate for meta-analyses than the risk difference. The risk difference may be considered when the control rates are reasonably similar.

When using a relative measure, the risk or rate differences should be calculated using the combined relative effect measure and control (comparison) event rate.

Use hazard ratios for meta-analysis of time to event data.

Effect measures quantify differences in outcomes between treatments in trials or exposure groups in observational studies. Effect measures can be broadly classified in two ways: (1) *absolute*, e.g., risk difference, rate difference, or mean difference; or (2) *relative*, e.g., odds ratio (OR), relative risk, or relative hazard ratio. The number needed to treat (NNT) and number needed to harm (NNH), which are the inverse of the risk difference, may also be considered an effect measure. Estimating and interpreting the NNT or NNH in a meta-analysis may not be straightforward, however (Altman and Deeks 2002; Cates 2002).

The choice of effect measure in meta-analysis is often prescribed by two factors:

1. the type of outcome data used, e.g., continuous, dichotomous, ordinal, interval, counts, or time to event
2. the corresponding measure reported, e.g., mean difference or standardized mean difference, relative risk, rate ratio, odds ratio, risk difference, or hazard ratio. The NNT (or NNH) can be derived from the risk difference from each study but is usually not combined in meta-analysis because its standard error is rarely calculated or reported.

Dichotomous Outcomes

A dichotomous outcome (e.g., death, stroke, or loss of vision) is the most common type of outcome reported. It offers straightforward interpretation—an event has occurred or not. The corresponding proportions of study participants are reported, and relative and absolute differences in proportions or event rates have well-defined meanings.

Both absolute and relative effect measures convey important aspects of evidence. Commonly used relative effect measures for dichotomous outcomes include the relative risk, incidence rate ratio, and odds ratio. The distinction between relative risk and incidence rate ratio is subtle and for practical purposes the two are similar. Odds ratios are the only effect measure directly calculable from a case-control study. The relative risk has a clear clinical interpretation as the ratio of two probabilities, but the odds ratio has more tractable statistical properties. Odds ratios overestimate relative risks as events in the referent or control group become common (e.g., more than 10 percent) (Zhang and Yu 1998).

Absolute measures for dichotomous outcomes, such as the risk difference, are easier for patients and clinicians to interpret than are relative measures (Covey 2007). The risk difference, or absolute difference in risk or event rates between groups, is the basis for calculating NNT or NNH.

Because they are easy to interpret, absolute measures such as the risk differences would be preferred, but some empirical evidence suggests that an absolute effect measure is usually less consistent than a relative measure across studies (Deeks 2002). Relative measures (relative risk, odds ratio) are more likely to be homogeneous across studies, particularly when variation among control group rates is large. When the control event rates are similar among trials, combining risk differences may be used. When rates among control groups vary widely among studies, the investigators can use the relative measure, or calculate pooled estimates of both measures to see whether pooling the risk differences introduces bias. When meta-analysis of the risk difference is not appropriate, investigators can use the summary relative effect size and an applicable control group rate to estimate the predicted absolute difference for a specific population.

Continuous outcomes

For studies reporting outcomes on a continuous scale (e.g., blood pressure, quality of life measurements), measurements are typically available at baseline and one or more follow-up times. The mean difference between groups at follow-up or the difference in change from baseline can be combined—standardized or not. The distinction between a mean difference at follow-up and the difference in change from baseline is important, but it may be overlooked because in trials with similar baseline values their magnitudes are generally similar.

The choice of effect measure to combine for continuous outcomes is determined primarily by the form of data available. If multiple trials report results using the same or similar scale, mean differences or differences in change between groups can be combined. Standardized effect sizes expressed as differences, or differences in change over time, divided by standard deviations are sometimes combined. This method is typically used when outcome measures are reported in different scales. Although this measure can incorporate multiple scales, the results can be difficult to interpret.

For some continuous outcomes, a small change can be judged clinically meaningful on an individual level; someone achieving that minimal change can be considered a responder (Tubach, Dougados et al. 2006). Under these circumstances, a fundamental limitation of continuous effect measures is that they fail to identify the proportion of patients experiencing a meaningful clinical response (Senn 1997). Analysts should avoid inferring individual responses from a summary weighted mean difference for a continuous outcome measure. However, when a meaningful clinically important improvement has been defined, it is reasonable to estimate the proportion of patients responding with a meaningful difference in a continuous outcome measure.

In the QOL literature, simulation studies suggest that a relationship can be found between the standardized effect size, minimally important difference, and NNT (Norman, Sridhar et al. 2001; Schunemann, Jaeschke et al. 2006). The relationship is based on normality assumptions and does not consider whether response varies with different baseline values. Furukawa developed a table converting standardized effect size to NNT when a response rate is known for one arm of a trial

(Furukawa 1999). The conversion assumes normally distributed data with equal variances in each arm. Furukawa and colleagues also suggest an imputation method (Furukawa, Cipriani et al. 2005). Still, our understanding of the relationship between continuous effect measures and clinical response is not complete and requires analysts to make certain assumptions. That is, inferring clinical response from differences in combined continuous outcomes may be explored, but the methods have not been well scrutinized and should therefore be applied carefully.

Table 9.1 illustrates considerations involved in choosing an effect measure. The investigators included eight trials examining the efficacy of an intervention for treating knee osteoarthritis. Pain was measured on a 100mm visual analogue scale (VAS) (0 for no pain and 100 for the worst imaginable pain) at baseline and following treatment. Although the trials reported only mean changes, using simulated data they were able to explore relative and absolute effects with response defined by a 20mm or greater decrease in VAS pain—a magnitude representing a minimum clinically important improvement (Tubach, Ravaud et al. 2005). All combined results were estimated from random effects models.

Table 9.1 Different effect measures and magnitudes combined from simulated data based on eight trials assessing treatment efficacy for knee osteoarthritis on a 100 mm visual analog scale; response for relative and absolute effects defined by ≥ 20 mm improvement.

Relative and Absolute Effects	Effect Magnitude	95% CI	P-value
Weighted Mean Differences (mm)			
Post-test difference	-11.8	-22.2 to -1.3	0.03
Difference in change	-11.6	-21.5 to -1.7	0.02
Standardized Mean Differences			
Post-treatment difference	-0.57	-1.21 to 0.08	0.09
Difference in change	-0.57	-1.16 to 0.02	0.06
Relative Effect Measures			
Odds Ratio	1.78	0.73 to 4.31	0.20
Relative Risk	1.27	0.90 to 1.78	0.17
Absolute Effect			
Risk Difference*	0.13	-0.07 to 0.32	0.20

*For illustration only. Generally the risk difference should not be combined.

The various effect measures differ in magnitude and levels of statistical significance, and they may convey different meanings. For example, the combined mean difference in change of 11.6mm is statistically significant but accompanied by a wide confidence interval. The corresponding standardized effect size (-0.57) is modest in magnitude and did not reach statistical significance at the 0.05 level. Neither relative risk nor risk difference was statistically significant and the odds ratio overestimated the relative risk substantially. Conclusions and clinical interpretations could vary based on the choice of effect measure.

Time to events

The effect measure for analyzing time-to-event or survival data is the hazard ratio (HR). The hazard ratio, an estimate of the relative risk, is also referred to as the relative hazard (RH).

Probably the most common survival analysis yielding a HR is the Cox proportional hazards model. The model assumes that the HR is constant over time, yielding a single value for a given study. Studies reporting a HR from the model should state explicitly whether or not the proportional hazard assumption was satisfied. However, a HR may not always be available or explicitly reported. Commonly, event rates are reported at various times during follow-up. Under such circumstances, the HR and its variance can be calculated if observed and expected events can be extracted (Parmar, Torri et al. 1998).

Choice of Model for Combining Studies

Box 9-3 Key points (Choice of Model for Combining Studies)

Broadly speaking, two types of model are available to combine studies: fixed effects model and random effects model. Both types of models can be used to combine effect measures for dichotomous data, continuous data or time -to -event data.

A fixed effects model assumes a single treatment effect across studies, and provides the best estimate of the treatment effect, if there were a single common treatment effect.

A random effects model assumes that the treatment effects across studies follows some distribution and the combined estimate is the center and most likely estimate of the distribution of treatment effects.

Choice of a model should not be solely based on tests of heterogeneity.

A random effects model is generally preferred since clinical and methodological diversity are inevitable among included studies.

For rare dichotomous outcomes, a fixed effects model, such as Peto odds ratio, or the Mantel-Haenszel method, is preferable to inverse variance syntheses and to some random effects models.

For rare or zero events, use relative measures and include studies with zero events in one arm in the meta-analysis. The relative measure can be estimated with the addition of 0.5 or alternative values as a correction factor.

Studies with 0 events in both arms should be excluded from the main analyses but should be summarized quantitatively.

Sensitivity analyses using both fixed effects and random effects model may be conducted to examine how model choice affects the combined estimates and conclusions.

Meta-analysis can be performed using either a fixed or a random effects model. A fixed effects model assumes a single treatment effect across studies, whereas the random effects model assumes that the treatment effects across studies are not identical but rather follow some

distribution. A common assumption is that the distribution is normal. The combined effect estimate from a fixed effects model is usually interpreted as being the best estimate of the treatment effect, if in fact a single common treatment effect exists. From random effects analysis, the combined estimate represents the center of the distribution of treatment effects, the most likely estimate from within the distribution of treatment effects.

Generally, a fixed effects model should not be used in the presence of significant heterogeneity. Moreover, some argue that clinical and methodological diversity is always present across studies and that variation among studies is inevitable whether or not the test of heterogeneity detects it. Therefore, random effects models are often suggested as preferable to fixed effects models. When heterogeneity is present, a random effects model gives more weight to smaller studies, and it incorporates the unexplained heterogeneity across studies in estimating the confidence interval of the combined estimate to produce a wider confidence interval than a fixed effects model will produce. When no statistical heterogeneity is present among studies, the random and fixed effects models yield identical or near-identical results.

A common criticism of the random effects model is that it is difficult to validate the assumption that treatment effects are normally distributed, especially when the number of studies is small (although there is no commonly accepted rule as what number is too small). When the results of small studies are systematically different from the results of the large ones, which can happen because of publication bias or differences in study quality between small and large studies, a random effects model will accentuate bias (Poole and Greenland 1999; Kjaergard, Villumsen et al. 2001). In this case, where the normality assumption is not justified, a fixed effects model would provide a less biased effect estimate, but it is not entirely appropriate either. In this situation, as suggested by the Cochrane handbook (Higgins and Green 2005), it may be wise not to present any summary estimate; alternatively, one can perform a sensitivity analysis excluding small studies or studies with poor quality. For dichotomous outcomes, when the events are rare, a random effects model would provide a biased estimate of between-study variance. The Mantel-Haenszel method will provide a more robust estimate of combined effect, at the cost of disregarding the observed heterogeneity.

Both fixed and random effects models can be used to combine dichotomous measures. A marginal analysis to estimate a summary effect size—that is, summing all events in each intervention across all studies and treating all studies as a single mega-study, or not including the study strata in a logistic regression—is generally not correct. The marginal approach neglects the differences among studies and the possible confounding effect by the study strata.

Fixed effects model

Several fixed effects methods are commonly used to combine effect measures for dichotomous outcomes: the Mantel-Haenszel method, the inverse variance method, the Peto method, and logistic regression. The latter two methods apply only to combining odds ratios.

Both the Mantel-Haenszel and the inverse variance methods can be used to combine odds ratios, relative risks, and risk differences. When studies report only few events, the Mantel-Haenszel method works better than inverse variance and has better statistical properties. In other situations the two methods give similar estimates. Logistic regression also works similarly to the Mantel-Haenszel method. The Peto method can be used only to combine ORs. It works well when

treatment effects are small (i.e., odds ratios are close to 1), events are not particularly common, and the trials have similar numbers in experimental and control groups.

In short, the Mantel-Haenszel method is a better choice than the other approaches in most cases. EPCs should keep in mind that for the inverse variance method, odds ratios and relative risks need to be log-transformed before they are combined, and combined odds ratios and relative risks are obtained by transforming the combined estimate back to its original scale.

Random effects models

Random effects models incorporate variation among studies into the estimate of the combined effect measure, and the combined estimate has a wider confidence interval. The most commonly used method was proposed by DerSimonian and Laird (DerSimonian and Laird 1986). The DerSimonian and Laird approach is a variation of the inverse variance method; it adjusts the weight to incorporate heterogeneity across studies.

Alternative estimates are derived by using simple or profile likelihood methods. The DerSimonian and Laird method does not adequately reflect the error associated with parameter estimation, especially when the number of studies is small. The profile likelihood method usually provides an estimate with better coverage probability and should be used when possible (Brockwell and Gordon 2001).

For odds ratios, a logistic random effects model could be used to combine results, although it may underestimate the uncertainty (Smith, Spiegelhalter et al. 1995).

Combining continuous outcomes

For a fixed effects approach to combine continuous outcomes, analysts should generally use the inverse variance method. For random effects models, a DerSimonian and Laird approach or likelihood approaches can be used.

Combining counts, rates, or time to event outcomes

These outcome measures are often expressed as risk ratio (count data), rate ratio (rate data) or hazard ratio (time to event). These measures are comparable in most cases. Both fixed effects (inverse variance method) and random effects models (DerSimonian and Laird, likelihood methods) could be used.

Bayesian models

Both fixed and random effects models have been developed within a Bayesian framework for dichotomous and continuous outcomes. The Bayesian fixed effects model provides good estimates when events are rare for dichotomous data (Sweeting, Sutton et al. 2004). A full Bayesian random effects model takes account of uncertainty in all parameters. For complex meta-analyses, the Bayesian methods can provide a flexible modeling framework.

The main criticism of Bayesian meta-analysis is that the specification of prior distributions is subjective. When the prior distributions are noninformative, Bayesian estimates are usually comparable to estimates using conventional methods.

Choice of model for sparse data (rare dichotomous outcomes).
When the outcome of interest is relatively rare, few or zero events may occur in one or both arms in several studies. Examples include an important but uncommon adverse event or mortality in populations with low baseline risk. In these cases, the normal approximation to the binomial distribution does not hold, and commonly used meta-analysis methods may not yield correct confidence intervals (Sweeting, Sutton et al. 2004; Bradburn, Deeks et al. 2007).

This situation occurs frequently when trials designed to test efficacy are pooled to estimate the rate of rare adverse events. Although such trials usually have smaller sample sizes, systematic reviewers should recognize that the presence of a large number of such trials could result in a distorted estimate of harms because a substantial number of patients experiencing no events have been excluded from the analyses (Nissen and Wolski 2007). Unless individual patient data are available for analysis, such trials are often excluded from pooled estimates of harms because a relative risk or odds ratio cannot be calculated (Table 9.2. Meta-analysis methods that utilize the effect sizes of each study—such as the fixed effects inverse variance method or the DerSimonian and Laird random effects method—necessitate the use of correction factors. The Mantel-Haentszel method, the Peto method, and Bayesian approaches do not explicitly need the addition of a correction factor in studies with zero events.

Table 9.2 Effect sizes that become inestimable if there are zero events and no correction factors are used.

Situation	Effect size			Standard error		
	logOR	LogRR	RD	logOR	LogRR	RD
Zero events in one arm	Not estimable	May be not estimable*	Estimable	Not estimable	Not estimable	Estimable**
Zero events in 2 arms	Not estimable	Not estimable	Estimable (0)	Not estimable	Not estimable	(0)**

RD=risk difference, OR, odds ratio, RR, relative risk
* depending on whether the 0 gets in the denominator of the RR
** Normal approximation does not apply in this situation. Therefore, one cannot use this standard error estimate to calculate 95% CI based on a normal distribution.

Rare events. For nonzero rare events (event rate < 1 percent), the Peto odds ratio method provides combined estimates that are least biased and have best confidence interval coverage. This is true provided that no substantial imbalance exists between treatment and control group sizes within trials and that treatment effects are not exceptionally large. The bias in the Peto method is evident in extreme imbalances (e.g., 8:1) and for large effects (e.g., OR <=0.2 or OR>=5) (Greenland and Salvan 1990). These circumstances are not likely to be observed in most medical meta-analyses of medical interventions.

For more frequent events (rates ~5 percent), the Mantel-Haenszel OR method and logistic regression perform similarly, and they are less biased than the Peto method. For the risk difference, both Mantel-Haenszel method and the inverse variance method provide biased combined estimates, with conservative confidence interval coverage and low statistical power; for that reason, these methods are unsuitable for meta-analysis of rare events (Bradburn, Deeks et

al. 2007). This advice also applies to studies with zero events in either one arm or both arms. Other measures are better choices for rare events, based on current evidence.

Zero events in both arms. When both arms have no (zero) events, the relative measures (OR and RR) cannot be defined. Some experts consider these studies to be noninformative and propose excluding them from the calculations (Sweeting, Sutton et al. 2004; Bradburn, Deeks et al. 2007). However, others consider including such "zero" studies in the analyses to be important (Sankey, LA et al. 1996; Friedrich, Adhikari et al. 2007) because excluding them biases the results in the direction of a higher event rate. This approach accords with the general preference that all available data be used.

If the sample sizes are small, including or excluding such studies does not change the summary effect size (OR or RR) appreciably because they receive very small weight in the synthesis. Nonetheless, inferential changes may be observed (Friedrich et al. BMC Med Res Methodology 2007). When EPCs use relative measures in a CER, they should exclude studies without any events from the main analyses. The excluded studies could be qualitatively summarized, as in the hypothetical example below (Table 9-3), to provide information on the confidence interval for the proportion of events in each arm. In addition, in sensitivity analysis reviewers can add these studies and look for any changes in the magnitude or the variance of the summary effect or in heterogeneity estimates and testing.

Table 9-3 Qualitative summary of studies with no events in both groups

	Intervention A		**Intervention B**	
Studies with zero events in both arms	Counts	95% exact confidence interval for the proportion of events	Counts	95% exact confidence interval for the proportion of events
Study 1	0/10	(0, 0.31)	0/20	(0, 0.168)
Study 2	0/100	(0, 0.036)	0/500	(0, 0.007)
Study 3	0/1000	(0, 0.004)	0/1000	(0, 0.004)

Explore Heterogeneity

Subgroup Analyses

Subgroup analyses are encouraged to explore causes of heterogeneity. Unfortunately, subgroup analyses are often misused leading to inferential leaps and overinterpretation that may result in erroneous conclusions (Yusuf, Wittes et al. 1991; Oxman and Guyatt 1992; Assmann, Pocock et al. 2000; Pocock, Assmann et al. 2002; Rothwell 2005; Hernandez, Boersma et al. 2006).

Box 9-4.1. Key Points (Heterogeneity and Subgroup Analysis)

Explore heterogeneity using one or more of the following methods: subgroup analysis and meta-regression with sensitivity analyses.

Consider excluding studies. This is a useful approach when heterogeneity is caused by one or two "outlier" studies. A clear and defensible rationale should be provided for identifying "outlier" studies.

Distinguish between pre-specified and post hoc subgroup analyses. When possible, use pre-specified subgroup analyses based on the key questions to explore heterogeneity.

Be as rigorous in the subgroup analyses as in the primary analyses. One should become familiar with the data, assess them for systematic differences in the effects at different control rate and assess the impact of additional factors on outcomes.

Be conservative in the interpretation of subgroup differences. Between-subgroup differences may be due to chance (data-dredging) and biases (e.g., ecological fallacy, outcome reporting bias).

If the results of subgroup analysis are to be considered valid, they should be clinically plausible and supported by other external or indirect evidence.

The use of random effects meta-regression is preferred to check for subgroup differences. Alternatively, a z-score can be used to compare between-subgroup effects.

The qualitative difference between subgroups defined *a priori* or *a posteriori* in a clinical trial or an observational study is substantial. To a certain extent this distinction pertains to other research designs as well. For CERs, *a priori (ad hoc)* specified subgroup analyses are those that are decided on during the planning of the systematic review and before data analysis has occurred. Factors that are expected to account for clinical or methodological heterogeneity are typically included in such analyses (e.g., differences in populations, differences in the interventions or their comparators, or variability in the study design). Good knowledge of the clinical and biological background of the topic and key questions is important in delineating a succinct set of useful and informative subgroup analyses. Ideally, the most important subgroup analyses are laid out in the key questions of the systematic review or CER after careful consideration of the topic and taking into account information from outside experts and previous reviews. Common examples of factors that are considered in subgroup analyses are age categories, sex, and other topic-specific factors such as device type, site of lesion, or disease severity.

Analyses of subgroups defined *a posteriori (post hoc)* are those that are guided by the data. Typically, they are those suggested only after preliminary analysis. Such analyses are potentially

a form of data dredging (Yusuf, Wittes et al. 1991; Oxman and Guyatt 1992), and they may result in uncontrolled type I error, i.e., false-positive associations (Yusuf, Wittes et al. 1991; Oxman and Guyatt 1992; Rothwell 2005).

However, meta-analysis is usually a retrospective design. In contrast to a randomized study, the actual data are often known to a greater or lesser extent when the analyses are being planned. Indeed, for some topics, the reviewers may be more or less familiar with the individual studies. Therefore, the distinction between pre- and post-specified subgroup analyses may not be that clear. For example, someone who is very familiar with a set of studies may be in a position to recognize a strong pattern between studies with different characteristics before performing the actual meta-analysis. Two teams that perform essentially the same effectiveness review might, therefore, consider the same subgroup analysis differently: it may be specified *a priori* by one team and *post hoc* by the other team. Similarly, a *post hoc* subgroup analysis in an early meta-analysis may be specified as an *a priori* analysis in its update; unless the update is extensive and substantial, the qualitative distinction between the two kinds of subgroup analysis is less clear.

Therefore, these distinctions may not be that clear for CERs. As a general rule, differences among subgroups (specified either *a priori* or *a posteriori*) should be clinically plausible and supported by other external or indirect evidence, if they are to be convincing.

Subgroup analyses may be performed with various strategies. First, reviewers may derive the summary effect measures in each subgroup and then compare the subgroups pair-wise using z-scores, calculated as the difference in the effect size divided by its standard error. For more than two subgroups, an ANOVA (analysis of variance) could be used.

Subgroup analyses should follow a process that is as rigorous as that in the primary analysis. The analyst should begin with graphical assessments of the data and evaluate systematic differences in the effect size across different control rates. Alternatively, one may perform the subgroup analyses in the context of meta-regression where the subgroup variable is coded as an indicator variable. In meta-regression (discussed below), more than two subgroups can be assessed in the same analysis.

Meta-regression

As noted in the section on heterogeneity, meta-regression is a valuable tool to investigate the contribution of specific factors to between-study heterogeneity. In a meta-regression, the effect size, for example the (log) odds ratio, is regressed against study characteristics such as dosages, durations of treatment, proper blinding, and patients' characteristics such as demographic factors or severity of illness. Without the benefit of individual patient data, these meta-regression models must rely on the summary results of published studies. These summary results describe only *between-study*, not

Box 9.4-2. Key Points (Meta-regression)

Use pre-specified factors (see Box 9.5-1).

Meta-regressions on summarized patient-level covariates such as mean age, proportion of males, may provide useful insight. Interpretation of the findings should be conservative, because of the potential ecological fallacy.

between-patient, variation in the risk factors. For that reason, they are most useful for characteristics *that differ across studies* and that are shared by all participants in the same study.

Study-level versus patient-level predictors in meta-regressions. Meta-regression models describe associations between the summary effects of a study and study-level data. There are two distinct types of study-level data: (1) study-level factors that apply equally to all patients in a study, and (2) study-level summary statistics representing the aggregate of individual patient-level data (Lau, Ioannidis et al. 1998; Schmid, Lau et al. 1998; Schmid, Cappelleri et al. 2004). Examples of the former are study design, timing of measurement of variables (e.g., time of measurement of serum creatinine), and definition of outcomes. Examples of the latter are mean age, the mean baseline serum creatinine value for all patients, and the percentage of diabetic patients.

Meta-regression on a study-level factor that has the same value for all patients in a study is mathematically equivalent to one based on individual patient data. On the contrary, this is not true for a meta-regression on summaries of patient-level factors (Schmid, Cappelleri et al. 2004).

Ecological fallacy. A meta-regression on summarized patient-level factors may be subject to *ecological fallacy*, a phenomenon in which associations present at the patient level are not necessarily true at the study level. This problem arises because the summary statistic does not necessarily adequately describe the individual values of each patient (Lau, Ioannidis et al. 1998). First, group averages among studies may vary only by a little even though the range within each study is wide. For example, in many studies the average age of patients is very similar even though the age range in some studies is wide. When a study-level variable has a small range, a regression analysis has trouble picking up an association. Second, the group average may not account for within-study variation. Two studies may have the same average but very different distributions of values with different implications for the outcome.

In summary, meta-regressions on study-level covariates may provide useful insight and help formulate hypotheses. They should be interpreted with caution, however, taking into account the aforementioned caveats.

A note on selecting factors in meta-regression analyses. Ideally, factors included in meta-regressions should be prespecified. Prespecifying characteristics reduces the likelihood of spurious findings by limiting the number of factors to analyze and preventing knowledge of the trials' results from influencing the choice of factors analyzed. True prespecification ideally occurs when the key questions are formulated. However, in doing the qualitative synthesis of studies, reviewers may identify factors that they did not think of prior to starting the study. Such factors can be included in a meta-regression, but investigators should state clearly that they were not prespecified.

How many studies are needed for a meta-regression? There is no single correct answer to this question. Meta-regressions are weighted linear models, not ordinary least squares regressions. Empirical research studies have employed a minimum of six studies for a meta-regression (Schmid, Lau et al. 1998), although there was no definitive methodological reason behind that choice (i.e., the authors opted for at least 4 degrees of freedom in their meta-regressions).

Caveats on the use of meta-regression for subgroup analyses. Many characteristics that might have important effects on the outcome may not be able to investigated with meta-regression due to inadequate data. Certain important risk factors may be reported in only a subset of studies. Analysis of this subset alone reduces the ability to find associations because of the loss of statistical power, but it can lead to bias if studies were more likely to report risk factors when an association with an outcome was found.

Meta-regression also cannot handle factors that vary by patient within studies. These patient-level factors, such as age or diabetes status, can be included in meta-regression only using study-level summaries, such as mean age or percentage of subjects who are diabetic. Associations present at the patient level will not necessarily be seen with study-level data.

Most random effects meta-regression models assume that between-study heterogeneity (τ^2) is common across all "subgroups" defined by the explanatory variables. If this is not true, then the power to detect significant associations may diminish; this situation is analogous to the t-test with equal vs. unequal variances. In such cases, results obtained with the simple z-score method and results obtained from a meta-regression may differ.

Appropriate interpretation of subgroup analyses and meta-regressions requires caution. Subgroup analyses and meta-regressions are entirely observational in nature. These analyses investigate differences among trials, and although individuals are randomized to one group or other within a trial, they are not randomized to go in one trial or another. Hence, subgroup analyses suffer the limitations of any observational investigation, including possible bias through confounding by other trial-level characteristics. Further, even a genuine difference among subgroups is not necessarily a result of the classification of the subgroups.

Control Rate Meta-regressions

Patients at different baseline risks may experience different benefits and harms (Glasziou and Irwig 1995). For studies with dichotomous outcomes, the event rate in the control group ("control rate") is affected by disease severity, concomitant treatments, follow-up duration, and other factors that may differ across studies (Lau, Ioannidis et al. 1998; Schmid, Lau et al. 1998). In an empirical evaluation, control rate effects were seen in 14 percent, 13 percent, and 31 percent of 115 meta-analyses of dichotomous outcomes when the measure of choice was the odds ratio, the risk ratio, or the risk difference, respectively (Schmid, Lau et al. 1998). The differences in the proportions between the relative measures (odds ratio, risk ratio) and the absolute measures (risk difference) are not surprising: a risk ratio of 1.5 corresponds to very different risk differences at various levels of baseline risk (0.5 percent at a 1 percent control rate, and 5 percent at a 10 percent control rate).

Box 9-4.3. Key concepts (Control Rate)

- Always examine the relation of effect size to control rate differences, which may reflect the difference in patient characteristics.
- Results are measure-specific, so assess control rate effects with the effect size measure you plan to use.

- Use graphical methods to assess how control rates influence the treatment effects. forest plots ordered by increasing control rate, L'Abbe plots, or scatter plots of the point estimates of the effect size versus the control rate (a smoothed line may offer additional insights)
    - In forest plots and scatter plots, search for a systematic change in the effect size at different control rates.
    - In L'Abbe plots, note whether the line that connects the points of the estimates does not pass through (or near) the origin.
- For formal inference, use a simple weighted regression of the effect size on the control rate.
    - If the slope is not significantly different than 0, it is most likely that the formal methods would agree.
    - If the slope is significantly different than 0, advanced methods must be used to obtain the correct level of statistical significance (see text).

Technical issues arise in the assessment of control rate effects because of the regression-to-the-mean phenomenon (Schmid, Lau et al. 1998; Sharp and Thompson 2000; Thompson and Higgins 2002). The control rate is correlated with the effect size, because effect size is calculated using information on control rate (McIntosh 1996; Thompson, Smith et al. 1997). For this reason, simple weighted regressions tend to identify significant control rate effects twice as often as more suitable approaches (Schmid, Lau et al. 1998). Formal approaches to this problem include hierarchical meta-regression models (Schmid, Lau et al. 1998) and Bayesian meta-regressions (Thompson, Smith et al. 1997).

Graphs are valuable to assess the presence of control rate effects. L'Abbe plots, cumulative meta-analyses that order studies by their control rate, or even scatter plots with smoothed interpolation lines may be used. A quick way to assess the presence of control rate effects uses simple regressions of the effect size on the control rate. A negative finding with a simple regression would be most likely replicated by the more complicated methods; a positive finding would need to be verified by a more comprehensive method. As always, EPCs should not rely automatically on the formal significance levels; rather, they should critically evaluate all findings.

Indirect Comparisons

Placebo-controlled trials can be helpful for evaluating absolute rates of benefits and harms associated with an intervention. However, evidence from head-to-head comparisons is always preferable to adjusted indirect analyses from placebo or active-controlled trials for evaluating comparative efficacy and harms.

For indirect analyses to be reliable, studies should be similar in terms of quality, factors related to applicability (population, interventions, settings), measurement of outcomes, and incidence of adverse events (Bucher, Guyatt et al. 1997; Song, Altman et al. 2003; Glenny, Altman et al. 2005; Chou, Fu et al. 2006).

In addition, formal methods for adjusted indirect comparisons incorporate additional variance to account for increased uncertainty when combining different sets of data, which frequently render apparent differences in estimates nonsignificant (Bucher, Guyatt et al. 1997; Glenny, Altman et al. 2005). A more informative approach would be to explore reasons for the discrepancies in rates of arrhythmias in the control arms and how they may have affected results.

Box 9-5. Key points (Indirect Comparisons)

- In the absence of sufficient direct head-to-head evidence, indirect comparisons can be considered as an additional analytic tool. How reliable results are, however, is uncertain and interpretations of findings must be made carefully with the limitation of indirect comparison in mind.
- Indirect comparisons have limitations under real world conditions. The validity of adjusted indirect comparison method depends on the consistency of treatment effects across a set of studies.
- Simple methods (such as Bucher) for making adjusted indirect comparisons have been validated with extensive simulations and can be easily implemented in commonly used statistical packages such as R, S-plus, Stata or SAS. These are the methods of first choice.
- Do not use unadjusted indirect comparisons of outcomes, which are subject to confounding effects from factors that differ between compared groups (e.g. severity of disease, control event rates).
- Avoid overstating findings based on qualitative indirect comparisons. A qualitative indirect comparison may be useful when there is a large degree of overlap in confidence intervals.
- Decisions regarding the number of studies needed should be made on a case-by-case basis, depending on sample sizes and event rates.
- If results from direct and indirect evidence conflict, evaluate the studies' similarities and dissimilarities.

Qualitative and informal indirect comparisons

Informal indirect comparisons (e.g., concluding that intervention A is safer than intervention B because confidence intervals relative to placebo do not overlap) should be avoided, as they assume that conditions for reliable indirect comparisons are met. A naïve comparison—studying A vs. B by obtaining the summary event rates in A from one set of RCTs and comparing them with the summary event rates in B from another set of RCTs—is generally wrong. This naïve method ignores the randomized nature of the data, and it is subject to a variety of confounding factors. The confounders will bias the naïve estimate for the indirect comparison in an unpredictable direction and with uncertain magnitude (Song, Altman et al. 2003). For example, a meta-analysis of COX-2 selective NSAIDs found rofecoxib associated with an increased risk of arrhythmia compared with control treatments; celecoxib was not (Zhang, Ding et al. 2006). However, the rate of arrhythmia in the control arms was 10-fold higher (0.27 percent or 18 of 6,568 subjects) with celecoxib than with rofecoxib (0.02 percent or 2 of 10,174 subjects). An implicit or naive indirect comparison about relative safety of celecoxib compared with rofecoxib is likely, therefore, to be problematic.

Investigators often make judgments on the indirect comparison between A and B by observing the effects of A vs. C and B vs. C, respectively. They may then use the degree of overlap in the confidence intervals of A vs. C and B vs. C to claim "similar" effects for A and B against a

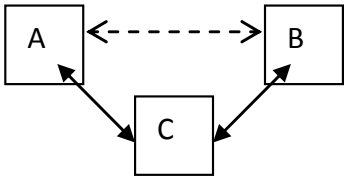
"common" comparator and therefore "equivalence" of A and B. This approach utilizes the summary effect sizes of A vs. C and B vs. C comparisons, but in a nonquantitative way. The extent of overlap of the confidence intervals is not, however, a very reliable substitute for formal testing. If the overlap in confidence intervals is large, formal testing is likely to provide results consistent with qualitative indirect comparison. However, when confidence intervals have a small degree of overlap, the formal testing may show significant differences. The reverse can also be true.

Quantitative methods for indirect comparison

Researchers who conduct CERs often have to decide whether to employ statistical methods to compare competing interventions indirectly when head-to-head RCT data are sparse or unavailable. (Ioannidis 2006; Lumley 2002).

To illustrate the situation, we consider the simple case of three interventions A, B, and C, in which RCTs compare A vs. C and B vs. C, but not A vs. B. Figure 9-1 depicts this situation.

Figure 9-1. A simple network of three interventions.



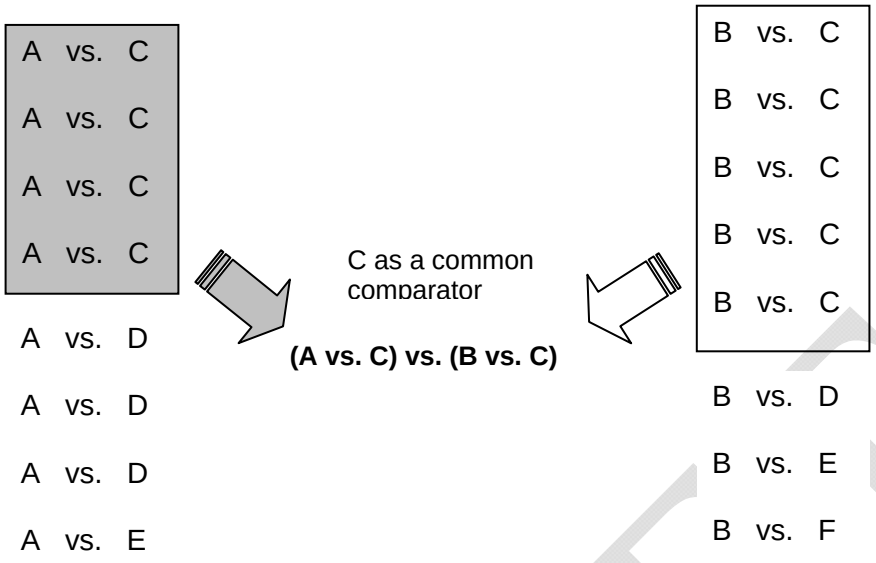
Solid arrows: direct (head-to-head) comparisons; dashed arrow: implied indirect comparison

An acceptable way to get estimates for the indirect comparison of A vs. B is through an adjusted indirect comparison. The rationale behind the adjusted indirect comparisons as described by Bucher (Bucher, Guyatt et al. 1997) is simple and can be approached by the following thought experiment:

- Assume that the effects of A vs. C (e.g., the RR_{AC} of A vs. C) are invariable across all possible event rates in C.
- Similarly, assume that the effects of B vs. C (RR_{BC}) are invariable across all possible event rates in C.
- The RR_{AB} is RR_{AC} / RR_{BC}.
- The variance for the indirect effect is the sum of the variances of the log direct effects because RR_{AC} and RR_{BC} are independent:
  $var(\log(RR_{AB}) = var(\log(RR_{AC}) + var(\log(RR_{BC})$

This approach is a special case of multi-treatment meta-analysis (MTM) models (Gleser and Olkin 1994; Berkey, Anderson et al. 1996; Lumley 2002; Lu and Ades 2004). The adjusted indirect comparison method relies on the invariance of the treatment effects across study populations and retains the benefits of randomization in the original RCTs (Bucher, Guyatt et al. 1997; Song, Glenny et al. 2000). Figure 9-2 outlines the principle of an adjusted indirect comparison.

Figure 9.2. Adjusted indirect comparisons



In one special case, the point estimate from the naïve method and the point estimate from the indirect method will be very similar. This would be expected when:

the effects for all A vs. C studies are consistent and so are the studies of B vs. C, *and* all studies in both comparisons have the same (or very similar) proportion of events in the C arm.

Even in this case, however, the confidence interval of the "A vs. B" indirect comparison with the naïve method will be too narrow. In the event EPCs ever use such an approach, this problem with interpretation must be addressed.

Validity of indirect comparison

Studies over the past decade have evaluated the validity and reliability of various statistical methods to conduct indirect comparisons (Bucher, Guyatt et al. 1997; Baker and Kramer 2002; Lumley 2002; Song, Altman et al. 2003; Caldwell, Ades et al. 2005; Glenny, Altman et al. 2005). A Health Technology Assessment, conducted by Glenny and colleagues (Glenny, Altman et al. 2005) for the National Health System Review and Dissemination Health Technology Program is the largest, and most thorough, empirical evaluation of a simple "networks" method (Glenny, Altman et al. 2005).

The validity of the result from an indirect comparison depends greatly on the consistency of the treatment effect across the two different sets of trials. In practice, trials can vary in numerous ways including population characteristics, interventions and cointerventions, length of follow-up, loss to follow-up, and study quality. Although some of these factors can easily be assessed and controlled for, others require assumptions that may not be verifiable.

Because indirect comparisons essentially constitute an observational study, residual confounding can always be present. Differences in factors between the two sets of trials that could influence the prognosis will bias indirect comparison results. In addition, all caveats that have to be

considered for meta-analyses, such as heterogeneity, publication bias, or differing control event rates, also apply for indirect comparisons.

In general, indirect comparisons have low power and often lead to indeterminate results. Four times as many equally sized studies are necessary for an adjusted indirect approach to have the same power as a direct comparison (Bucher, Guyatt et al. 1997; Glenny, Altman et al. 2005). Reviewers that consider indirect comparisons for CERs need to keep the limitations of indirect comparisons under "real world" conditions in mind. First, given the limited information in many publications, the validity of indirect comparisons often has to be based on an unverifiable assumption of similarity. Second, indirect comparisons are underpowered, frequently leading to indeterminate results with wide confidence intervals. Inferences based on such findings may be limited. Third, no consensus exists on how to interpret results that differ substantially from direct evidence or on how to weigh findings of indirect comparisons against those results from nonrandomized direct evidence.

Incorporation of mixed study designs in a meta-analysis

The most common design seen in randomized trials is the parallel trial. Other designs—such as crossover, factorial, or cluster-randomized design—are also common choices. In principle, trials from different designs may be combined in a single meta-analysis. Whether to combine them depends on the features of the studies in a CER.

Box 9-6. Key Points (Combining Studies of Mixed Design)

- If crossover trials are appropriate for the intervention and medical condition in question, and there are no systematic differences between the two types of design, crossover designs can be combined with parallel trials.
- For meta-analysis of crossover trials, use estimates from within-individual comparisons when available. When estimates from within-individual comparisons are not reported, or when there is a carryover effect, use between-group estimates or data from the first treatment period.
- Cluster randomization trials can be combined with individual randomized trials.
- When available, effect measures from an analysis that appropriately accounts for the cluster design should be used for meta-analysis.
- When the best estimate is not available, analyses using effective sample size or corrected standard error are also an option.
- Investigators should explicitly state what kind of data are available in the paper, how they have dealt with data from crossover trials or cluster randomization trials, and how the decision of combining or not combining crossover trials has been made.
- In general, randomized trials and non-randomized studies should not be combined.

Generally, producing a combined estimate for crossover trials and parallel trials separately is advisable, when such estimates are appropriate, whether or not crossover trials and parallel trials are combined.

Incorporation of crossover trials in a meta-analysis

In a crossover trial, each participant receives two or more interventions in random order and serves as his or her own control. The design is most appropriate to compare interventions with a

reversible, temporary effect on symptoms in patients with a stable, chronic disease, such as multiple sclerosis or rheumatoid arthritis. It is not appropriate for a disease with rapid progression or for interventions with long-lasting effects such that, upon entry to subsequent phases, patients systematically differ from their initial state owing to the treatment effects of interventions. For example, if the primary outcome is irreversible such as death, or pregnancy in a fertility study, a crossover design is generally inappropriate.

In addition, some secondary outcomes may not be properly evaluated from crossover trials. One such example is the withdrawal rate in the assessment of adverse events. If one patient drops out in the first period, this patient cannot be evaluated for withdrawal in the later periods.

The strength of the crossover design is that it allows comparison of interventions at the individual rather than group level. The major concern with the crossover design is the risk of a carryover effect, when the effect of an intervention in the first period persists and influences the patient's response in the subsequent intervention period. For this reason, a "washout" period between treatment periods is often included to reduce the risk of a carryover effect.

In considering whether to include crossover trials in a meta-analysis, investigators should first evaluate whether a crossover trial is appropriate for the intervention and medical condition in question and whether it may provide useful information to answer the research question. The risk of carryover and the adequacy of the washout period should be fully evaluated. Combining crossover and parallel trials is reasonable if these trials are estimating the same intervention effect and if the choice of trial design has not been dictated by any differences in therapeutic indication or clinical conditions that could potentially influence the observed treatment effect (Elbourne, Altman et al. 2002).

Approaches to include crossover trials into meta-analyses. The most frequently used crossover design has two interventions with two periods. The ideal estimates for meta-analyses are those from within-individual comparisons for which the standard errors are estimated appropriately. EPCs may sometimes be able to calculate these estimates if the article does not report them (Elbourne, Altman et al. 2002). Estimates from within-individual comparisons from a crossover trial could be combined with results from parallel trials.

Methods of combining have been developed for continuous data (Curtin, Altman et al. 2002; Curtin, Elbourne et al. 2002), dichotomous data (Curtin, Elbourne et al. 2002), and even when carryover occurs, although carryover effects may exist but not be detected (Curtin, Elbourne et al. 2002). Elbourne and colleagues provided examples on how to apply some of these methods (Elbourne, Altman et al. 2002). Unfortunately, the reporting of estimates from crossover is often very variable and incomplete, and the ideal estimate for meta-analysis is not reported. Frequently, extracting suitable data is difficult or impossible, even if the appropriate analysis was performed.

When results are reported only for each intervention, a simple approach is to ignore the crossover design and use reported estimates as if they came from a parallel trial. This approach, which ignores the within-patient correlation, is a conservative approach that is likely to produce a confidence interval wider than it should be. It also reduces the weight given each crossover trial,

with the possible consequence of disguising clinically important heterogeneity. If this approach is used, sensitivity analysis may be done assuming a range of within-patient correlations to check the robustness of the results and whether the results may approximate those from a paired analysis.

Another approach is to include only the data from the first period, on the grounds that the first period of a randomized crossover trial is, in effect, a parallel group trial. This method is often recommended when either a carryover effect is a problem or the crossover design is considered inappropriate for the condition or outcome being investigated. However, available data from the first period is a biased subset of all first-period data as first-period data are often reported only when there is evidence of carryover effect. In addition, excluding later periods loses some of the information collected.

In all cases, investigators should explicitly state what kind of data are available, how they have dealt with data from crossover trials, and how the decision whether to combine crossover trials has been made. Use sensitivity analysis to investigate the robustness of conclusions.

Incorporation of cluster randomized trials in a meta-analysis

In cluster randomized trials (also known as group-randomized trials), a group of individuals in intact social units (rather than a single individual) is randomized to different interventions. The unit of allocation is the cluster; the clusters may be, for example, schools, communities, families, or practice settings. Cluster randomization trials involve several additional potential sources of heterogeneity such as choice of cluster randomization schemes (e.g., matched pair vs. complete randomization), the nature of the randomization unit (e.g., households vs. worksites), and the sizes of clusters. Analysis of cluster randomized trials must take the correlation of results within groups; if it does not, false significance may be claimed and the study may receive more weight than is appropriate in a meta-analysis.

A CER may include cluster randomized trials and individual randomized trials that address the same scientific question. To decide whether combining the results from both cluster randomized trials and individual randomized trials is appropriate, EPCs should assess whether the type of randomization unit affects the intervention and outcome. The presence of such interaction is less likely when the intervention is a pharmacological agent with biological effects than when the intervention is intended to shape attitudes or behaviors (Donner and Klar 2002). One approach is to begin by performing separate meta-analyses. If the results of separate meta-analysis agree, the investigators should have more confidence that the findings are robust. If the results do not agree, investigators should note the lack of consistent findings.

Important differences and characteristics between different types of trials should be fully considered.

The best estimate from a cluster randomized trial appropriately accounts for the cluster design. Such estimates could be obtained from a mixed effects model, or a model using generalized estimating equations (GEEs), among other techniques.

Unfortunately, many cluster randomized trials have not performed or reported appropriate analyses. For example, the analysis may have been done as if the randomization were on the

individual-patient level. In this case, approximately correct analyses may be performed if data on cluster size, summarized outcome results (ignoring cluster design), and an estimate of the intra-cluster correlation coefficient (ICC) are all available (Donner and Klar 2002). The idea is to calculate an "effective sample size." A corrected standard error based on estimates of ICC and cluster size can be used in the meta-analysis. However, ICC is seldom available in published reports. A common approach then is to use external estimates obtained from similar studies (Ukoumunne, Gulliford et al. 1999), supplemented by sensitivity analysis assuming plausible values of ICC.

Donner and colleagues provide a detailed discussion of incorporating cluster-randomized trials in a meta-analysis (Donner and Klar 2002) and a more technical treatment of the problem (Donner, Piaggio et al. 2001). White and Thomas examine special considerations for analysis of standardized mean differences from cluster-randomized trials (White and Thomas 2005).

Synthesis of information from randomized and nonrandomized studies

Observational studies and randomized trials are complementary sources of evidence. Randomized and nonrandomized evidence often agree in their results (Concato, Shah et al. 2000; Benson and Hartz 2000; Ioannidis, Haidich et al. 2001). However, discrepancies are not infrequent (Ioannidis 2005). Examples where findings from observational studies were not replicated by large clinical trials are numerous (Ioannidis 2005). Conversely, knowledge based on a few small randomized trials may be refuted by subsequent large, well-designed and well-conducted observational trials. For example, a meta-analysis of small RCTs of aprotinin in cardiac surgery failed to find an increase risk for renal failure (Carless, Moxey et al. 2005), whereas such a risk was discovered in a large observational study (Mangano, Tudor et al. 2006).

Currently, we recommend against combining randomized and nonrandomized studies including observational data. Statistical methods concerning how to incorporate observational data have not been well developed. Further research is needed to investigate whether combining nonrandomized and randomized trials is appropriate and, if so, under what conditions. Other issues still needing examination include how to assess systematically the consistency between nonrandomized and randomized trials. Finally, further development of statistical methods to combine observational and trial data is needed (see Chapter 8).

Sensitivity Analyses

Completing a CER is a structured process. Decisions and assumptions are made in the process of conducting the review and meta-analysis; these decision and assumptions may affect the main findings.

Sensitivity analysis is an approach to investigate how the decisions and

Box 9-7. Key points

- Sensitivity analysis is an approach to examine the robustness of the combined estimates to decisions and assumptions made in the process of review.
- A CER with a meta-analysis should include a sensitivity analyses.
- Planning of sensitivity analysis should start at the early stage of a CER.
- Investigators should describe key decisions and assumptions for sensitivity analysis.

assumptions influence the main findings and how robust the results are to these decisions and assumptions. Various experts have argued that sensitivity analysis should be performed to reflect the decision made at all stages of a meta-analysis (Olkin 1994) and always be performed to assess the robustness of combined estimates (Egger and Smith 2001).

Sensitivity analysis is a necessary step for a meta-analysis in CERs. Planning of sensitivity analysis should start at the early stage of a CER and should include tracking decisions made along the way. Investigators should identify key decisions and assumptions for sensitivity analysis, such as:

- the effect of including certain types of excluded trials;
- whether the combined estimates are consistent across the subgroups of study population;
- whether the combined estimates are consistent across intervention types and settings; studies that have been stopped early and studies that have run their planned course; and studies with different lengths of follow-up time;
- how inclusion or exclusion of studies rated as "poor quality" (i.e., having a high risk of bias) affects the combined estimates, and how the combined estimates are affected by individual factors contributing to risk of bias, such as use of blinding, concealment of allocation, objective ascertainment of outcomes;
- whether results derived from different effect measures (e.g., relative risk and risk difference) agree;
- whether results from a fixed effects model vs. a random effects model, or different formulations of a fixed or random effects model, agree;
- how different approaches for handling missing data, zero cells, and incomplete data reporting affect the results. [One example is the calculation of standard deviation (SD) for pre-post mean difference for a group based on reported means and SDs at baseline (pre) and endpoints (post). Estimating the correlation coefficient between baseline and endpoints is required for calculating SD but often is not reported. Then a reasonable range of values for the correlation coefficient could be assumed with influence on the results assessed.]; and
- whether the combined estimates are consistent with the sample size of the study.

Robust estimates provide more confidence in the findings in the review. In a meta-analysis on the effect of beta-blockers on mortality after myocardial infarction, Egger and Smith presented a good example of how sensitivity analysis can increase confidence in the results (Egger and Smith 2001). The sensitivity analysis examined the robustness of the combined estimate to choice of statistical model (fixed effects vs. random effects), concealment of allocation, double-blinded vs. other blinding method, trial size, length of follow-up, and exclusion of trials stopped early. These factors had little influence on the combined estimate.

When the results are not robust, it indicates the need to interpret results cautiously or employ alternative approaches for presenting a grand combined estimate. For example, if the results are sensitive to length of follow-up, then combined estimates based on different length of follow-up should be reported.

Statistical methods for sensitivity analysis are readily available. A new analysis could be done by including or excluding certain studies. EPCs can study the influence of each study by excluding one study at a time, but investigating the influence of key decisions and assumptions is more important. Subgroup analysis is often used for sensitivity analysis, and the cautions for subgroup analysis should also be applied here (see above).

Interpretation and Translation of Results of Meta-analyses

CERs should present summary effects in a way that makes it easy for readers to interpret and apply these findings appropriately. This section discusses different ways of presenting and interpreting various effect measures.

The most commonly used effect measures for dichotomous outcomes in meta-analyses are relative risks (RR) and odds ratios (OR). They provide the most stable estimates over various populations and differing study durations. The interpretation of RR is fairly straightforward as the ratio of probabilities of an event (risk or benefit) between two intervention groups. An RR of 2 can be interpreted as a twofold risk of an event in patients on a treatment compared with those not receiving the treatment. For example, in a study examining the adherence to prescribed inhalers for patients with chronic obstructive pulmonary disease, authors stated that patients on tiotropium were twice as compliant as patients using ipratropium (RR: 2.0; 95% CI,1.8-2.3) (Breekveldt-Postma, Koerselman et al. 2007). Likewise, authors of a meta-analysis reported that subjects exposed to crystalline silica had a twofold incidence of lung cancer compared with those not exposed to crystalline sylika (RR: 2.0, 95% CI, 1.8-2.3) (Smith, Lopipero et al. 1995).

Alternatively, presentating results as a relative risk reduction or relative risk increase may be more intuitive for readers, especially when the RR is below 2. For example, a CER on second-generation antidepressants compared nausea and vomiting between venlafaxine and the class of selective serotonin reuptake inhibitors (SSRIs) (Gartlehner, Hansen et al. 2007); the pooled RR was 1.50 (95% CI, 1.21-1.84). The authors expressed this finding as a relative risk increase and stated that venlafaxine had a 50 percent higher risk of causing nausea or vomiting than SSRIs as a class. Such statements, however, should be accompanied by a measure of absolute risk, such as the risk difference or a number needed to treat (NNT) or harm (NNH), to provide enough information for readers to assess the clinical relevance of such a finding. The CER on second-generation antidepressants reported a corresponding NNH of 9 (95% CI, 6-23).

The term “relative risk” can be confusing, however, if it refers to a beneficial outcome. Substituting “relative risk” with “relative benefit” may help readers avoid confusion with contradicting terminology. For example, the term "relative benefit" was used in a systematic review on the efficacy and safety of second-generation antidepressant to describe the beneficial response to treatment (Hansen, Gartlehner et al. 2005).

Although ORs have mathematical advantages over RRs, they are more difficult to interpret because they describe the ratio of the odds of an event among those exposed to an intervention to the odds among those not exposed. Frequently, odds and odds ratios are interpreted inappropriately as risks and risk ratios. Because odds and risks differ substantially when event rates are high (> 30 percent), incorrect interpretations can lead to an overstatement of the actual effect size. For example, a study examining physician diagnostic practices for patients with chest

pain noted a statistically higher rate of cardiac catheterizations for men than for women (OR 1.7, 95%CI, 1.1-2.5) (Schulman, Berlin et al. 1999), causing concerns in the media about gender disparities. Schwartz et al. reanalyzed the same data and found a substantially smaller effect size when using RRs (RR 1.07 95%CI 1.01-1.16) (Schwartz, Woloshin et al. 1999). Thus, converting ORs to RRs may be advisable to allow easier interpretation.

The clinical relevance of risk differences (RD) depends on the underlying event rates. A RD of 2 percent could be clinically significant if the change is from 3 percent to 1 percent of an event, and less significant if the intervention reduces the rate of events from 78 percent to 76 percent. Therefore, when reporting a risk difference, the underlying event rate should be reported as well. For example, in a placebo-controlled RCT, the authors presented event rates and relative and absolute risk reduction to summarize differences in the risk of ventricular fibrillation and arrhythmic death (Table 9-4) (Cairns, Connolly et al. 1997).

Table 9.4. Absolute and relative risk reduction of resuscitated ventricular fibrillation and arrhythmic death for amiodarone vs. placebo in all patients and subgroups with different baseline rates of events.

Patient Characteristics	Amiodarone		Placebo		Relative-risk Reduction (%)	Absolute risk reduction
	Events	Rate per year (%)	Events	Rate per year (%)		
All patients	25	2·29	39	3·71	38·2	1.42
Concomitant use of beta -blockers	2	0.38	16	2.93	87·1	2.55
Previous myocardial infarction	6	2.33	20	6.89	66.2	4·56
Previous congestive heart failure	8	5.13	15	7.8	34	2.67

NNTs and NNHs are frequently used because they portray the absolute effect of an intervention in an intuitive way. NNTs and NNHs themselves do not reflect variations attributable to underlying event rates; and they do not have a standardized unit of time. These drawbacks should be considered when NNTs or NNHs are presented. EPCs should report these measures with an appropriate time frame and confidence intervals and make clear that they are based on an average estimate, for example, "On average, 10 patients would have to be treated for 3 years with treatment A to observe one fewer event after 3 years" (Hutton 2000). If substantial variations in NNTs (NNHs) exist based on different event rates, dosages, or subgroups, then EPCs should report them separately for each group.

Smeeth and colleagues calculated NNT with statins to prevent one cardiovascular event (Smeeth, Haines et al. 1999). Although the authors pooled studies to achieve a summary NNT, they also presented NNTs for individual studies with varying baseline risks (**Table 9.5**.). The pooled NNT to prevent one death was 113 over 5 years. NNTs of individual studies, however, ranged from 41 to 167 corresponding to baseline risks of cardiac death from 1.4/100 to 0.1/100 person-years.

Continuous Outcomes

The weighted mean difference (WMD) and the standardized effect size can be used for meta-analyses of continuous data. WMD can be used when outcome measurements in all trials are assessed on the same scale. The summary effect has the same unit as the scale employed in the included studies. For example, the CER on second-generation antidepressants conducted a meta-analysis of differences in points on the Montgomery-Asberg Depression Scale (MADRS) between escitalopram and citalopram (WMD of 1.13 [95% CI, 0.18 – 2.09]) (Gartlehner, Hansen et al. 2007). Because the unit of the pooled summary effect is the same as that of the original scale, findings can be interpreted as escitalopram having an additional treatment effect of 1.13 points on the MADRS. Although this finding was statistically significant, the clinical significance of a difference of 1.13 points must be determined independently.

Standardized effect size meta-analyses can be used if the same outcome was assessed on different measurement scales. Results however, are expressed in units of standard deviations, rather than in units of any measurement scales and can be difficult to interpret. For example, Hansen et al. pooled functional outcomes in placebo-controlled studies of Alzheimer's drugs (Hansen, Gartlehner et al. 2007) and reported the pooled findings in units of standard deviations. Although they had interpreted results based on a classification of Cohen's *d*, the clinical significance of the additional treatment effect of Alzheimer's drugs compared with placebo is difficult to determine. An approximation of the size of the effect on the included measurement scales can be achieved by multiplying the standardized effect sizes by the pooled standard deviation for each included scale.

Table 9.5. Number needed to treat with statins to prevent one cardiovascular event in 5 years.

Trials	No of subjects	Baseline risk of CHD mortality per 100 person-years	Rate ratios			Number needed to treat (5 years)		
			Total mortality	CHD mortality	All CV events	Total mortality	CHD mortality	All CV events
Primary prevention								
AFCAPS/TexCAPS7	6,605	0.1	1.04	1.36	0.69	167	1000	28
WOSCOPS8	6,595	0.4	0.78	0.67	0.7	118	182	28
Secondary prevention								
Scandinavian simvastatin survival study trial	4,444	1.6	0.71	0.59	0.64	33	31	8
CARE	4,159	1.2	0.92	0.81	0.75	133	95	11
Long-term intervention with pravastatin in ischaemic disease	9,014	1.4	0.78	0.77	0.8	41	64	17
Pooled effects (95% CI)			0.80 (0.74 to 0.87)	0.73 (0.66 to 0.81)	0.74 (0.71 to 0.77)	113 (77 to 285)	500 (222 to -)	20 (17 to 25)

CHD, coronary heart disease; CV, cardiovascular.

Reporting the Quantitative Synthesis of Studies

The following summary of headings (Tables 9.6 and 9.7) for reporting quantitative syntheses of studies may ensure some degree of uniformity in how EPCs can present CER methods results. The summary is not entirely prescriptive because CERs only need to include headings relevant to analyses included in the review. If a review touches upon an area encompassed by a heading or subheading, then the heading or subheading should be included in the review. Reporting of elements pertaining to the heading or subheading should be done in accordance with the explanations provided in the "required reporting" column of the table below. For additional information, we identify the section of the guide that discusses the pertinent issues. The exact titles of headings and subheadings are left to the discretion of authors.

For example, if the authors decide to conduct a meta-analysis, then they will have to include a heading in the methods section of their report that pertains to "method of combining studies." Under this heading, they will have to describe and justify the statistical procedure used to combine effect measures from individual studies. In the results, a graphical summary of individual and combined study effect estimates will have to be provided in accordance with the recommendations enumerated below.

Table 9.6. Summary of Headings for Reporting the Quantitative Synthesis of Studies: Methods Section

Headings	Subheadings	Required Reporting
Rationale to combine (see 9.1 Decisions to combine or not combine individual studies)	Clinical heterogeneity	Specify important clinical characteristics which may differ among studies (e.g. intervention, dosage, baseline disease severity, length of follow-up) and how they will affect the decision to combine. Define the threshold for acceptable differences in clinical characteristics which could be combined in a meta-analysis based on the scope of the research question.
	Methodological heterogeneity	Specify important methodological characteristics which may differ among studies (e.g. mechanism of randomization, extent and handling of withdrawals and losses to follow up) and how they will affect the decision to combine. Define the threshold acceptable differences in methodological characteristics which could be combined in a meta-analysis based on the scope of the research question.
Criteria for selecting outcomes for combining (see 9.1 Decisions to combine or not combine individual studies)	Outcome definitions	Specify whether outcome definitions or the way outcomes were measured differed among studies. Specify whether surrogate outcomes or combined endpoints were used. If observational studies are included, specify the definition and measurement of confounding factors/effect modifiers.
	Primary vs. secondary outcomes	Specify whether outcomes were primary or secondary outcomes in the original studies. Specify benefit and harm outcomes and their combinations.
	Outcome assessment in RCTs	Specify whether ITT, per protocol, last observation carried forward, etc. was used to handle outcomes in each study. If estimates from different outcome definitions were combined, then subgroup and/or sensitivity analyses should also be undertaken
Types of studies included (see 9.1 Decisions to combine or not combine individual studies)	Study design	Specify what type of study designs are being combined (e.g. RCT (crossover, cluster randomized, factorial), observational (cohort, case-control, cross-sectional))

Headings	Subheadings	Required Reporting
	Rationale for inclusion of observational studies	If observational studies are included, then provide a rationale (e.g. to broaden generalizability, to examine longer follow up periods, etc.)
Type of comparisons (see chapters 9.A1, 9.6, 9.7)	Direct comparisons (9.A1 An approach to the meta-analysis of aggregate data using direct comparisons	Specify what types of comparisons are being made. Specify what methods are used to combine study data if indirect or mixed comparisons are being made (e.g. logistic regression, meta-regression, or adjusted indirect comparisons)
	Indirect comparisons (9.6 Indirect comparison)	If indirect comparisons are being made then clearly state the rationale. Unadjusted indirect comparisons can lead to bias and should not be conducted
	Mixed designs (9.7 Combining studies of mixed designs)	Specify what types of RCT study designs are being combined (e.g. parallel group, crossover, cluster randomized). If observational studies are included, compare stratified results from RCT and observational studies either qualitatively or quantitatively. There are currently no standard methods to combine RCT and observational studies.
Explanation of choice of effect measure (see 9.2 Choice of effect measures)		Specify what type of outcome data is being combined (e.g. dichotomous, continuous, ordinal, counts, time to event) and the measure(s) of effect chosen (e.g. RR, OR, RD, HR, mean difference, standardized mean). This should be done for each outcome considered. If the study design allows a choice of effect measure then choose the one that best answers the research question and provide a rationale for that choice.
Methods for combining study estimates (see 9.3 Choice of model for combining studies)	Statistical procedure and justification of model chosen	Describe and justify the statistical model used to combine effect measures (e.g. random effects model, fixed effects model, Bayesian model)
	Special considerations (see 9.3.1 Special considerations)	Describe and justify the statistical methods used when there are sparse data or selective reporting of outcomes
Statistical heterogeneity (see 9.4 Heterogeneity, 9.5 Exploring heterogeneity)	Statistical tests	Specify how statistical heterogeneity is assessed and the threshold used to identify “important” heterogeneity
	Quantifying heterogeneity	Specify methods used to quantify statistical heterogeneity (e. g., I^2, H, tau^2)

Headings	Subheadings	Required Reporting
	Exploring heterogeneity	Specify the methods used to explore important clinical, methodological, or statistical heterogeneity (e.g., meta-regression, control rate meta-regression, subgroup analysis). Distinguish between pre-specified and post hoc analysis. The total number of subgroups examined should be reported.
Additional analyses (see 9.8 and chapter 10)	Assessment of selection bias including publication bias and selective reporting within studies (see chapter 10 Explore potential bias) Sensitivity analysis (see 9.8 Sensitivity analyses)	Specify how selection bias including publication bias and selective reporting within studies is assessed (e.g. funnel plots, L'abbe plots, Egger test); and how this information is to be used in any data syntheses. Specify what sensitivity analyses are being done and how they relate to key decisions and assumptions made in the systematic review. Sensitivity analyses should be specified *a priori*.

Table 9.7 Summary of Headings for Reporting the Quantitative Synthesis of Studies: Results Section

Headings	Subheadings	Recommendations
Descriptive study information (see 9.9 Interpretation and translation of results of meta-analyses)		Include information for each study describing the sample size, intervention, outcome, study design, target population, study population, baseline risk and other important PICOS study characteristics that are related to clinical, methodological or statistical heterogeneity. Sponsorship of the studies and reported conflict of interest should be reported.
Level of Evidence and Quality of the Studies		Specify the level of evidence given feasibility of different designs to investigate the research question. Specify the scale to estimate the quality of the study and how internal and external validity of the studies are assessed.
Graphical summary of individual and overall study estimates		For each outcome present tables or a graphical representation of the data (forest plot) including: the comparison type, sample size for each study, weight given to each study, measure of effect and confidence interval for each study, and a summary measure of effect and confidence interval for all studies combined. A p-value for a test and quantification of statistical heterogeneity should be included in the figure or in the figure legend. If study results are not quantitatively combined, a forest plot without a summary estimate can still be provided.

Headings	Subheadings	Recommendations
Reporting of additional analyses	Assessment of selection bias including publication bias and selective reporting within studies	Report a graphical assessment and/or the results of a statistical test. If a graph is reported, then include the results of the statistical test in the legend in the whole group and in subgroups.
	Other additional analyses (exploring heterogeneity, sensitivity analyses, inclusion of observational studies)	Report the results of all additional analyses undertaken.

Appendix 9-1. An Approach to the Meta-analysis of Aggregate Data using Direct Comparisons

If a selection of two or more studies has reasonably similar patient populations, design, comparisons, and outcome definitions to warrant a meta-analysis, the following steps are suggested to approach the analysis.

Step I. Examine the data and evaluate whether intervention effects vary with different rates for control groups (control rates)

1. Use graphical assessments if possible including.
    a. L'Abbe plots, or
    b. Forest plots ordered by increasing control rate,
    c. Scatter plots of the point estimates of the effect size versus the control rate (a smoothed line may offer additional insights).
2. Decide on which measure to use. Usually, relative (multiplicative) measures (OR, RR) rather than absolute (additive) measures (RD) are preferred for dichotomous data.
    a. General guidelines for dichotomous data:
        2.a.1 Use RR for RCTs or prospective comparative studies
        2.a.2 Use OR for case-control studies and related designs
        2.a.3 Use RD for RCTs or prospective comparative studies when the control rates are reasonably similar.
    b. General guidelines for continuous data:
        2.b.1 Use mean differences if outcome measures are reported on the same scale.
        2.b.2 Use standardized mean differences if outcome measures are reported on different scales. Standardized effect sizes are seldom used in meta-analyses in medical literature. However, whenever the use of a standardized effect size is warranted, Hedges's *g* is preferred over other metrics of standardized effect size (Hedges 1981) e.g., Cohen's *d*, Glass's δ.
3. If control rates vary, use simple (Ordinary Least Squares, weighted) regression of effect size on control rate to assess their correlation.
4. If the regression slope is statistically nonsignificant, chances are that more formal methods will yield the same inference. Skip to **Step II, below**.
5. If the regression slope is statistically significant, perform formal control rate meta-regression using a hierarchical model or Bayesian analyses. Go to **Step IIIA, below**.

Step II. Provide a grand mean

1. Test for between-study heterogeneity (Q) and assess its extent (I^2).
    a. If heterogeneity is extensive or substantial and there are clinical, methodological, and epidemiological reasons to expect systematic differences in the effects between the different studies, go to **Step IIIA, below**.
    b. If heterogeneity is not substantial or if heterogeneity is statistically significant or extensive but there are no strong clinical, methodological, and epidemiological reasons for systematic differences in the effects between the different studies, estimate a grand mean.
2. Because clinical diversity can always be anticipated, estimate a grand mean using a random effects model.

a. If between-study variability (τ^2) is 0, use the inverse variance fixed effects model, which yields the same results as the DerSimonian and Laird random effects model.

Step IIIA. Perform a meta-regression on control rates

1. Perform a (random effects) control rate meta-regression and describe the results along a range of control rates.
    a. Provide the predicted summary effect size for typical control rates in different settings.
    b. Corroborate with individual studies that have similar control rates.
2. If many studies exist and there is a strong reason to consider additional study-level covariates, include them in the model of the control rate meta-regression analyses.
3. If many studies exist and there is a strong reason to consider additional study-level covariates, include them in the model of the control rate meta-regression analyses.

Step IIIB. Perform a meta-regression on patient and study-level covariates

3. Identify patient and study-level covariates. Although aggregated patient-level factors (such as mean age or mean blood pressure) do not necessarily describe all patients accurately, they should be routinely explored.
4. Remember:
    a. Determine whether there is enough variation in the study-level covariates to run a meaningful meta-regression.
    b. Separate pre-specified analyses from post-hoc analyses.
    c. Because of the anticipated small number of studies, consider examining one covariate at a time.
5. If there are statistically significant associations
    a. Check their clinical plausibility
    b. Check for external evidence consistent with the identified association
    c. Make conservative statements.
    d. Provide adjusted estimates at different levels of the covariate, which may provide better insight than a grand summary estimate.
6. If no significant associations are present, provide the grand summary estimate.

Step IV. Conduct subgroup analyses

1. Conduct a subgroup analysis if not already performed to assess control rate differences, as described in **Step I above.**
2. Use meta-regression in place of the simple comparison of the summary estimates for subgroup analyses. A random effects meta-regression is preferred; meta-regression and *z*-score comparisons of subgroup summary estimates should give roughly similar results in most cases unless substantial heterogeneity within subgroups is present).

Step V. Conduct sensitivity analyses

1. Use sensitivity analysis to evaluate the robustness of the quantitative answers to key decisions and assumptions that were made in the process of the review. Depending on the exact topic, different sensitivity analyses may be preferred at different levels of the review process. Briefly (the list is not exhaustive), sensitivity analyses can be done for
    a. contextual review-specific decisions such as

i. Selection of included studies, e.g.:
    1. Subtleties in the definition of population, intervention, comparators and outcomes
    2. Language of publication, country of origin
    3. Studies that received different quality rating

b. technical decisions:
    i. Consider different effect measures
    ii. Consider exploring the effects of dropouts on individual studies, e.g. by using a best case/worst case analysis for trials that have dropouts (assuming that none or all of the dropouts experienced the event).
    iii. Consider different methods for assessing the influence of publication bias or other selection biases.

10. AVOIDING POTENTIAL BIASES IN CONDUCTING SYSTEMATIC REVIEWS AND META-ANALYSES

This chapter is deliberately omitted.

11. GRADING THE STRENGTH OF A BODY OF EVIDENCE

Comparative effectiveness reviews (CERs) are essential tools for summarizing information to help make well-informed decisions about health care options (Helfand 2005). Synthesizing data to provide robust conclusions, and doing so consistently, is a critical part of the health technology assessment process. Reviews should provide clear judgments about the strength of the evidence that underlies CER conclusions to enable decision makers to use CERs effectively. This chapter explores the rationale for grading strength of evidence, defines the domains of concern for evidence strength, and describes the grading system for CERs.

Box 11-1 summarizes the principal recommendations from this chapter. For ease of reference, the approach that EPCs should take for CERs is labeled an "EPC GRADE" approach, because several domains and elements of the overall grading approach are based on the basic GRADE approach (Atkins, Briss et al. 2004).

As noted in Box 11-1, we are recommending that EPCs grade strength of evidence only for the findings on major outcomes for the main comparisons of interest. These will be determined in part by the topic nomination process (Chapter 2) and further refinement of key questions early in the review and in part by actual findings once analyses are completed. As discussed in Chapter 2, we recommend use of analytic frameworks (causal pathways or logic models) for CERs. We anticipate that, ideally, the principal outcomes of interest will be those in the box of any analytic framework that specifies the ultimate or most important health outcomes (e.g., death, clinical morbidity, quality of life).

Box 11-1. Key Points

1. Grade strength of evidence for all comparisons of interest for the most important outcomes (benefits and harms).
2. Grade first for each major health outcome, then (if desired) each surrogate or intermediate outcome.
3. Grade for each major harm separately.
4. Apply required domains in all cases.
5. Apply additional domains when appropriate (largely with respect to observational studies)
6. Use two or more independent reviewers to assign grades. Resolve discrepancies or disagreements by consensus (first) or by use of a third, independent rater. Then report a single grade for the outcome.
7. In a table, record domain-specific and overall strength of evidence grades for each reviewer and retain this information at the EPC.
8. Describe how overall grade is determined from individual domains (e.g., qualitatively or with a numerical weighting system)
9. Record overall grade in tabular form (e.g., discussion chapter; executive summary), along with narrative text that explains the rationale for the grade.
10. Ensure that all major findings featured in the CER are graded and that findings for which a grade is given is highlighted in text.

Strength of Evidence: Domains

In drawing conclusions about strength of evidence, a growing number of organizations have adopted a systematic approach to making judgments about the strength of evidence. A wide variety of grading systems is available (West, King et al. 2002), and different organizations may weigh features, or domains, of a body of evidence differently. Consequently, discrepant,

contradictory, or variable ratings may arise, and results may not be helpful to some organizations.

A major challenge for AHRQ is to ensure some consistency in how reports from different EPCs grade the strength of evidence. Attaining this goal rests in part on consistency and predictability in the domains that EPCs use in this effort. Although no one system for reporting results and grading the related strength of evidence is likely to suit all users, documentation and consistent reporting of the most important summary information about a body of literature will make CERs more useful to a broader range of potential audiences.

The EPC approach to grading evidence begins with a set of agreed-upon domains pertaining to entire bodies of evidence about key outcomes (benefits and harms) and comparisons. In selecting these domains, we relied on work by the US Preventive Services Task Force (Harris, Helfand et al. 2001), the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (GRADE Working Group 2004) and other work in this area by EPCs (West, King et al. 2002; Treadwell, Tregear et al. 2006). Judgments about those domains are then aggregated into an overall evidence grade (explained below) for each key outcome. **Tables 11-1 and 11-2** present two sets of domains: required and optional (respectively).

Required Domains

The first set, "required domains," comprises four major constructs that EPCs should use for all main outcomes and comparison(s) of interest in a CER: risk of bias, consistency, directness, and precision. Table 11-1 defines these and indicates how to assess or apply them. These four domains are discussed in more detail below.

Before assessing the required domains, an EPC must first identify the studies that address the outcomes and comparisons of interest. When no studies are available on an outcome or comparison of interest, the evidence should be graded simply as ***insufficient***.

For the remaining major outcomes and comparisons of interest, the grade of evidence will depend on the required domains and not on the number of studies; the EPCs have decided that focusing on consistency, directness, and precision is more informative than emphasizing just the number of studies. Nevertheless, EPCs should note in their CERs the number of studies overall and for potential (or at least observed) comparisons, and they should indicate the number of studies that form the basis of given findings or conclusions. In this way, readers can better understand the available evidence.

Table 11-1. Required Domains and their Definitions

Domain	Definition and Elements	Rating /Score or Application
Risk of Bias	Risk of bias is the degree to which the included studies for any given outcome or comparison have a high likelihood of adequate protection against bias (i.e., good internal validity), assessed through two main elements: • Study design (e.g., RCTs or observational	Use one of three levels of aggregate risk of bias: • Low Risk of Bias • Medium Risk of Bias • High Risk of Bias

Domain	Definition and Elements	Rating /Score or Application
	studies) • Aggregate quality of the studies under consideration. Information for this determination comes from the grading of quality (good/fair/poor) done for individual studies	See Chapter 6
Consistency	The principal definition of consistency is the degree to which reported effect sizes from included studies appear to go in the same direction. This can be assessed through two main elements: • Effect sizes have the same sign (that is, are on the same side of "no effect") • The range of effect sizes is narrow.	Use one of 3 levels of consistency: • No inconsistency • Inconsistency present • Unknown or not applicable (eg, single study) As noted in the text, single-study evidence bases (even mega-trials) cannot be judged with respect to consistency.
Directness	The rating of directness relates to whether the evidence links the compared interventions directly to health outcomes. For a comparison of two treatments, directness implies that there are head-to-head trials that measure the most important (health or ultimate) outcomes. Specifically, evidence is indirect if • It uses intermediate or surrogate outcomes instead of health outcomes; in this case, one body of evidence links the intervention to intermediate outcomes, and another body of evidence links intermediate to most important (health or ultimate) outcomes. • It uses two or more bodies of evidence to compare interventions A and B; for example, studies of A vs. placebo and B vs. placebo. Indirectness always implies that more than one body of evidence is required to link interventions to the most important health outcomes. Directness may be contingent on the outcomes of interest; EPC authors are expected to make clear the level of outcomes involved.	Score dichotomously as two levels of directness • Direct • Indirect If indirect, specify which of the two types of indirectness account for the rating (or both, if that is the case), and comment on the potential weaknesses caused by, or inherent in, the indirect analysis. The EPC should note if both direct and indirect evidence was available, particularly when . indirect evidence supports a small body of direct evidence.
Precision	Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (i.e., for each outcome separately) If a meta-analysis was performed, this will be the confidence interval around the summary effect size.	Score dichotomously as two levels of precision: • Precise • Imprecise A precise estimate is an estimate that would allow a clinically useful conclusion. An imprecise estimate is one for which the confidence interval is wide enough to include clinically distinct conclusions (for example, both clinically important superiority and inferiority (i.e., the direction of effect is unknown), a circumstance that will preclude a conclusion.

Risk of bias

As noted in Table 11-1, the risk of bias for an evidence base will be based on the assessment of the risk of bias in individual studies (see **Chapter 6 Assessing the Quality...**). If studies differ substantially in the risk of bias, greater weight could be given to the studies with a lower risk of bias.

Consistency

Main considerations. Consistency refers to the degree of similarity in the effect sizes of different studies within an evidence base. If effect sizes indicate the same direction of effect and the range of effect sizes is narrow, an evidence base can be judged to be "consistent." If meta-analysis is appropriate, EPCs can evaluate consistency using statistical tests and measures of heterogeneity (such as chi-square tests or I^2 statistics) as described in Chapter 9**.** Some evidence bases may show statistical heterogeneity in effect sizes but consistency in the direction of effect; if the heterogeneity cannot be explained, the evidence base can be judged to be consistent. With substantial unexplained heterogeneity, one cannot (or at least should not) determine a precise estimate of treatment effect, but one may still be confident in the direction of effect.

EPCs should designate an evidence base to be "inconsistent" when different studies show statistically significant effect sizes in opposite directions, In the absence of statistical testing or measurement of heterogeneity, judgment of consistency becomes more subjective.

Evaluation of a single-study evidence base. Evaluation of consistency ideally requires an evidence base with independent replication of findings; therefore, EPCs cannot properly evaluate consistency in an evidence base with a single study. Even if the study is a large multicenter trial (i.e., a mega-trial), findings from different centers within such a study are rarely reported separately. If the results are reported separately for each center, EPCs may be able to evaluate consistency within the overall trial, but this is not truly independent replication. Any flaw (reported or not reported) in the trial design or conduct will likely be replicated at every center. Even pairs of mega-trials addressing the same clinical question (i.e., the same patient-intervention-outcome combinations) may report discrepant results (Furukawa, Streiner et al. 2000), and the methodology of mega-trials has been further questioned (Charlton 2001).

Thus, EPCs can never be certain that a single trial, no matter how large or well-designed, presents the definitive picture of any particular clinical benefit or harm for a given treatment. Accordingly, we recommend that single-study evidence bases should not be judged with respect to consistency, because the consistency of findings cannot be adequately evaluated. The recommended judgment in this case is "consistency unknown (single study)."

Directness

As described in the section on analytic frameworks in Chapter 2**,** if direct evidence linking an intervention to the most important outcomes is lacking, then several bodies of evidence are needed to link the intervention to health outcomes. When several bodies of evidence are involved, the ultimate decision about using a service may depend on the strength of evidence for every link in the causal chain.

Some links in the causal chain will be more important than others, and thus the final assessment of directness will require consideration of the strength of evidence for each link as well as the importance of each link in the chain. For example, in the enteral feeding example in Figure 2-1, a large body of well-conducted randomized trials might demonstrate that enteral supplementation improved nutritional status and delivery of nutrients to the area of the wound. If, however, evidence of an association between a richer nutritional milieu and complete healing is weak, and experts agree that this is one of the more important links in the causal chain, then the decision might be to grade that overall body of evidence as indirect and the strength of evidence low.

The point about directness is that having a single body of evidence, as in this example that links enteral supplementation specifically with wound healing, is preferred to needing to use two linked bodies of evidence, particularly if the strength of evidence for those two bodies of evidence differ in material ways. Assessing directness clarifies the degree to which evidence between the intervention and the ultimate or health outcome does or does not meet the "ideal" set of studies addressing the overarching question.

Precision

Precision of an effect estimate is related to the boundaries of its confidence interval in relation to a threshold that would allow a judgment about the treatments being compared. Such thresholds include the boundary of statistical significance, or boundaries related to a conclusion about whether one treatment is clinically noninferior, equivalent, or superior to another (Sackett 2004; Sackett 2005). These boundaries may depend on the importance of the outcome being measured. Substantial heterogeneity does not necessarily render an estimate imprecise. A truly imprecise estimate is one with a confidence interval that does not rule out the superiority or inferiority of either treatment being compared. In this case, no conclusion can be reached.

Optional Domains

The second set of domains ("optional domains") consists of secondary constructs that EPCs should use and report if they are relevant to a particular CER. These domains also derive from the sources noted earlier. Table 11-2 provides their definitions and ways to score or apply them.

Generally, we expect these domains to be applied more often to evidence from observational studies (of all types) than from RCTs. However, these domains also will be relevant to assessing the strength of evidence from RCTs that have important limitations.

Table 11-2 Optional Domains and their Definitions

Domain	Definition and Elements	Scoring or Application
Coherence	Coherence is the degree of plausibility of results in relation to epidemiology or, in some cases, biology and pathophysiology.	This additional domain does not need to be described or noted unless something "implausible" has emerged, in which case EPC authors should comment on it.
Dose-response association	This association, either across or within studies, refers to a pattern of a larger effect with greater exposure (dose, duration, adherence)	Score as three levels: • Present: Dose-response pattern observed • Not present: No dose-response pattern observed (dose-response relationship *not* present) • NA (not applicable or not tested)
Residual confounding	Occasionally, in an observational study, residual confounders would work in the direction *opposite* that of the observed effect. A case in point is when a study is biased *against* finding an effect and yet it finds an effect. Thus, had these confounders not been present, the observed effect would have been even larger than the one observed.	Score as three levels: • Confounding unlikely to explain observed effect: Plausible residual confounders are more likely to have decreased the observed effect than to have increased the observed effect • Confounding may explain observed effect: Plausible residual confounders are unlikely to have decreased the observed effect and could be responsible for observed effect • Cannot assess
Strength of association (magnitude of effect)	Strength of association refers to the likelihood that the observed effect is large enough that it cannot have occurred solely as a result of bias from potential confounding factors.	Score as two levels: • Strong: large effect size that is unlikely to have occurred in the absence of a true effect of the intervention (e.g., relative risk > 5). • Weak: small enough effect size that it could have occurred solely as a result of bias from confounding factors (e.g., relative risk < 5).

Other Pertinent Issues

Publication Bias

Publication bias indicates that studies may have been published selectively with the result that the estimated effect of an intervention based on published studies does not reflect the true effect. Publication is regarded as separate from but related to strength of evidence. That is, the strength of a set of RCTs with consistent results depends on the assumption that similar (or better) RCTs with discrepant results were not systematically missed. The finding that only a small proportion of relevant trials has been published or reported in a results database may indicate a higher risk of publication bias, which in turn may undermine the overall robustness of a body of evidence. Although this is not a separate domain, publication bias can influence ratings of consistency and precision (and, to a lesser degree, risk of bias and directness.) For example, if the investigators identify unpublished trials, and if their results differ from those of published studies, investigators can take these factors into account in their rating for consistency and in calculating a summary confidence interval for an effect.

Applicability

Like publication bias, applicability is regarded as separate from but related to strength of evidence. Low applicability can reduce the strength of evidence for a particular decision maker. CERs should summarize characteristics that decision makers may need or want to consider in assessing the generalizability (sometimes denoted external validity) of the evidence. Decision makers may take into account how well the evidence maps to the patient populations, settings, diseases or conditions, interventions, comparators, and outcomes which are most relevant to their decisions. Because these factors differ with the perspective of decision makers, we have chosen to consider applicability as separate from but related to strength of evidence.

CERs should record information about applicability for the outcomes and comparisons for which they specify an overall strength of evidence rating. As described in Chapter 6**,** EPCs should summarize this information in a separate table that decision makers can review along with the strength of evidence table.

Rating Domains

EPCs should have two or more reviewers with the appropriate clinical and methodological expertise score each domain for each key outcome (benefit and harm). Differences should be resolved by consensus or mediation by an additional expert reviewer. Although the consensus judgments will appear in tables in the CERs, EPCs should record and save each reviewer's individual judgments about domains as background documentation for the CER.

Overall Strength of Evidence Grade

Four Strength of Evidence Levels

The overall grade for strength of evidence reflects a global assessment that takes the required domains above directly into account and, as needed, incorporates judgments about the optional domains as well. As noted, EPCs should rate strength of evidence for each major benefit (e.g., impact on health outcomes such as physical function or quality of life, or effects on laboratory measures or other surrogate variables) and each major harm (ranging from rare, serious, or life-threatening adverse events to common but bothersome effects) for each comparison of interest.

CERs can be broad in scope, encompassing multiple patient populations, interventions, and outcomes. EPCs are *not* expected to grade every possible comparison for every outcome. Rather, EPCs should set clear priorities, assigning grades to those combinations (patients-interventions-outcomes) that are likely to be of greatest interest to readers.

Table 11-3 summarizes the four levels of grades that EPCs should use. They have two components: (1) the principal definition concerns the level of confidence the authors place in the estimate of effect (benefit or harm)—i.e., that the evidence reflects the true effect; and (2) the subsidiary definition involves a subjective assessment of the likelihood that future research might affect the level of confidence in the estimate or actually change that estimate.

Grades are denoted high, moderate, low, and insufficient. They are not designated by Roman numerals or other terms.

Table 11-3. Strength of Evidence Grades and Definitions

Grade	Definition
High	**High confidence that the evidence reflects the true effect.** Further research is very unlikely to change our confidence in the estimate of effect.
Moderate	**Moderate confidence that the evidence reflects the true effect.** Further research may change our confidence in the estimate of effect and may change the estimate.
Low	**Low confidence that the evidence reflects the true effect .** Further research is likely to change the confidence in the estimate of effect and is likely to change the estimate.
Insufficient	Evidence either is unavailable or does not permit estimation of an effect.

High, Moderate, or Low Strength of Evidence

Assigning a grade of high, moderate, or low implies that evidence is available to estimate an effect size in the first place. EPCs should understand that, even when evidence is low, consumers, clinicians, and policymakers may find themselves in the position of having to make choices and decisions. The designations of high, moderate, and low should convey how secure reviewers feel about decisions based on evidence of differing grades.

Insufficient. In some cases, high, moderate, or low ratings will be impossible or imprudent to make; the reason is that the EPC cannot draw any conclusion for a particular outcome, specific comparison, or other question of interest. In these situations, EPCs should apply a grade of insufficient. Specifically, evidence for an outcome receives a grade of insufficient in one of two cases: (1) when no evidence is available or (2) when evidence on the outcome is too weak, sparse, or inconsistent to permit any conclusion to be drawn.

The former case is clear (no evidence at all is available from the included studies). The latter case is more complicated. It can arise for several reasons, such as unacceptably high risk of bias or a major inconsistency that cannot be explained (e.g., two studies with the same risk of bias that found opposite results, with no clear explanation for the discrepancy). Imprecise data may also lead to a grade of insufficient, specifically when the confidence interval is so wide that it includes two incompatible conclusions: that one treatment is clinically significantly better than the other, and that the difference is in the opposite direction. Indirect data based on only one study or comparison could also receive a grade of insufficient. If a single quantitative estimate is desired, the strength of evidence may be insufficient if an effect size cannot be calculated from reported information. This same evidence base, however, may still be sufficient to permit a conclusion about the general direction of the effect, but EPCs need to take care not to conflate "low" strength of evidence with "insufficient."

Incorporating Multiple Domains into an Overall Grade

To assign an overall grade to the strength of a body of evidence, EPCs must decide how to incorporate multiple domains into the overall grade. In some systems, such as that of the GRADE working group (Atkins, Briss et al. 2004), the overall evidence grade is calculated directly from the ratings for each domain using a transparent point system. Some groups, such as the American College of Physicians, have adopted many elements of GRADE but not an explicit numerical way to incorporate multiple domains into an overall grade.

A point system has the advantage of transparency, because it clearly delineates a direct path from the evidence to its grade. However, no empirical evidence supports the superiority (or inferiority) of such a numerical system compared with a more qualitative approach. Research is needed to compare the performance of point systems with other grading systems before we can recommend that EPCs use any specific system.

Although EPCs may use different approaches to incorporate multiple domains into an overall strength of evidence grade, several general principles are important. We recommend that EPCs first evaluate the risk of bias based on the study designs of the available evidence. For many types of outcomes, evidence that is based on randomized trials will have less risk of bias than does evidence based on observational studies. For these outcomes, if randomized trial data are available, the EPC may choose to start with a high grade for the strength of evidence and then downgrade the evidence based on other domains. If only observational data are available, the EPC may choose to start with a low grade for the strength of evidence, and then upgrade the data based on other domains. This overall approach is similar to, but more flexible than, the methods used in the GRADE system. For some outcomes, that is, the EPC may believe that observational studies have less risk of bias than do trials or that the available randomized trials have a substantial risk of bias. In such instances, the EPC may move up the initial rating of strength of evidence based on observational studies to moderate or move down the initial rating based on randomized trials to moderate.

The EPCs should use other domains to modify the overall grade for the strength of evidence. Inconsistency, indirectness, and imprecision should generally weaken the strength of evidence. EPCs should also consider the optional domains as appropriate. The strength of the evidence would be weakened by lack of coherence or evidence of residual confounding. In contrast, several factors may increase strength of evidence and are especially relevant for observational studies where one may begin with a lower grade based on the risk of bias. Presence of a clear dose-response association or a very strong association would justify increasing strength of evidence, as would a judgment that plausible biases (for example, residual confounding) would lead to underestimating the effect. The degree to which the overall strength of evidence is downgraded or upgraded based on these domains is a judgment call that the EPC should explain in the report.

EPCs should also take specific steps to ensure reliability and transparency within their own work (both in individual CERs and across them) when incorporating domains into an overall grade. The first step is to be explicit about whether the evidence grade will be determined by a numerical system for combining ratings of the domains or by a qualitative consideration of the domains

The second step is to identify the domains that are most important for the targeted body of evidence and to decide whether to use quantitatively or qualitatively different weights for the domains when assigning the evidence grade. For the sake of consistency across CERs, the domains should be defined using the terminology presented earlier in this chapter. In the absence of any evidence to support quantitative weighting of the domains, a qualitative approach generally will be reasonable; however, that does not mean that all domains should have the same weight. In general, the first or highest priority should be given to the domain for risk of bias, as it is well established that evidence is strongest when the study design has the lowest risk of bias.

The third step is to develop an explicit procedure for ensuring a high degree of inter-rater reliability for rating individual domains. As mentioned earlier, this assumes that at least two reviewers will rate each domain; ideally, they will have appropriate clinical and methodological expertise. In addition, EPCs should assess the resulting inter-rater reliability for each domain. Although EPCs generally will not include the details of the reliability assessment in the CER, the EPCs should keep records of this information for documentation. By recording this information, the EPCs will be able to increase knowledge about the reliability of the grading system.

The fourth step is to use the ratings of the domains to assign an overall strength of evidence grade according to the decisions made in steps 1-3. If this step involves a qualitative approach with subjective weighting of the domains, the EPCs should consider using at least two reviewers and should assess the inter-rater reliability of this step in the process. That will not be necessary if the approach involves a formulaic calculation or algorithm based on the ratings of the domains. However, the scoring system or algorithm should be specified in sufficient detail to permit replication by a reader.

The fifth step is to prepare a narrative explanation of the reasoning used to arrive at the overall grade for each body of evidence. This should include an explanation of what domains played important roles in the ultimate grades.

Reporting Strength of Evidence

As noted above, CERs should present information about all comparisons of interest for the outcomes that are most important to patients and other decision makers. Thus, strength of evidence should relate to those important outcomes in the comparative context.

Complete and perfect information is rarely available. For some treatments, data may be lacking about one or more of the outcomes. In other cases, the available evidence comes from studies that have important flaws, is imprecise, or is not applicable to some populations of interest. For this reason, CERs should also present information that would help decision makers judge the risk of bias in the estimates of effect, assess the applicability of the evidence to populations of interest, and take imprecision and other factors into account.

Table 11-4 illustrates one approach to providing actionable information to decision makers that does reflect strength of evidence. Specifically, it presents information pertinent to assessing evidence strength from different types of studies—specifically on the four required domains—and it displays estimates of the magnitude of effect (right column). For the outcome as a whole

(e.g., mortality or quality of life), the table also gives the overall rating. It shows, for instance, that one fair-quality RCT reported mortality, which was lower by one patient per 100 treated (i.e., 1 percent), a difference that was not statistically significant (95% confidence interval (CI), -4 percent to +3 percent). For the same comparison, 14 retrospective cohort studies had a wide range of effect sizes (range -7 percent to +5 percent). Had these estimates been precise (e.g., narrower CI for the RCT, consistent cohort studies to allow a summary effect size), one might have been able to reach a conclusion. Because these estimates are imprecise, however, the evidence is insufficient to allow a conclusion for mortality.

Table 11-4. Treatment 1 vs. Treatment 2: Numbers of studies and subjects, strength of evidence domains, magnitude of effect, and strength of evidence for key outcomes

Number of Studies; Subjects	Domains Pertaining to Strength of Evidence				Magnitude of Effect and Strength of Evidence
	Risk of Bias: Design/ Quality	Consistency	Directness	Precision	Absolute risk difference per 100 patients
Mortality					**Insufficient SOE**
1;80	RCT/Fair	Unknown	Direct	Imprecise	-1 (95% CI -4 to +3)
14;384	Retrospective cohort/ Fair	Inconsistent	Direct	Imprecise	-7 to +5 (range)
Myocardial infarction					**Low SOE**
7: 625	Retrospective cohort/ Low	Consistent	Direct	Precise	-3 (95% CI -5 to -1)
Severe diarrhea					**Moderate SOE**
4; 256	RCTs/ Fair	Consistent	Direct	Imprecise	-4 (95% CI -8 to +1)
14; 28,400	Cohort studies/ Fair	Consistent	Direct	Precise	-5 (95% CI -8 to -2)
Improved quality of life					**High SOE**
6; 265	RCTs/ Good	Consistent	Direct	Precise	-5 (95% CI -1 to -7)
Ulcer healing					**High SOE**
6; 265	RCTs/ Good	Consistent	Direct	Precise	+12 (95% CI +4 to +27)
5; 684	Retrospective cohort studies Good	Consistent	Direct	Precise	+17 (95% CI +12 to +22)

CI, confidence interval; RCT, randomized controlled trial; SOE, strength of evidence.

Although Table 11-4 illustrates how EPCs might organize information about the strength of evidence and magnitude of effect in ways useful to decision makers, it is incomplete. First, the table does not convey any information about the applicability of the evidence (Chapter 6). Second, a narrative summary of the results is also essential for interpreting the results of a literature synthesis.

REFERENCES

Altman, D. G. and J. J. Deeks (2002). "Meta-analysis, Simpson's paradox, and the number needed to treat." BMC Medical Research Methodology 2: 3.

Aronson, J. K. (2003). "Anecdotes as evidence." BMJ 326: 1346.

Aronson, J. K., S. Derry, et al. (2002). "Adverse drug reactions: keeping up to date." Fundamental and Clinical Pharmacology 16: 49-56.

Assmann, S. F., S. J. Pocock, et al. (2000). "Subgroup analysis and other (mis)uses of baseline data in clinical trials." Lancet 355(9209): 1064-9.

Atkins, D. (2007). "Creating and Synthesizing Evidence With Decision Makers in Mind. Integrating Evidence From Clinical Trials and Other Study Designs." Med Care 45: S16–S22.

Atkins, D., D. Best, et al. (2004). "Grading quality of evidence and strength of recommendations." BMJ 328(7454): 1490.

Atkins, D., P. A. Briss, et al. (2004). "Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group." BMC Health Services Research 4(38).

Atkins, D., K. Fink, et al. (2005). "Better information for better health care: The Evidence-based Practice Center Program and the Agency for Healthcare Research and Quality." Ann Intern Med 142(12_Part_2): 1035-1041.

Baker, S. G. and B. S. Kramer (2002). "The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C?" BMC Health Services Research 2: 13.

Begaud, B., Y. Moride, et al. (1994). "False-positives in spontaneous reporting: should we worry about them?" British Journal of Clinical Pharmacology 38(5): 401-404.

Benichou, C., G. Danan, et al. (1993). "Causality assessment of adverse reactions to drugs--II. An original model for validation of drug causality assessment methods: case reports with positive rechallenge." Journal of Clinical Epidemiology 46: 1331-1336.

Bennett, C. L., J. R. Nebeker, et al. (2005). "The Research on Adverse Drug Events and Reports (RADAR) project." JAMA 293 (17): 2131-40.

Bennett, D. A. and A. Jull (2003). "FDA: untapped source of unpublished trials." Lancet 361: 1402-3.

Benson, K. and A. J. Hartz (2000). "A comparison of observational studies and randomized, controlled trials." New England Journal of Medicine 342(25): 1878-86.

Bent, S., A. Padula, et al. (2006). "Brief communication: Better ways to question patients about adverse medical events: a randomized, controlled trial." Annals of Internal Medicine 144(4): 257-61.

Berkey, C. S., J. J. Anderson, et al. (1996). "Multiple-outcome meta-analysis of clinical trials." Stat Med 15(5): 537-57.

Bhandari, M. M., MSC, P. J. M. Devereaux, et al. (2002). "An Observational Study of Orthopaedic Abstracts and Subsequent Full-Text Publications
" Journal of Bone & Joint Surgery - American Volume 84-A(4): 615-621.

Bigby, M. (2001). "Challenges to the hierarchy of evidence: does the emperor have no clothes?" Archives of Dermatology 137(Mar): 345-346.

Bohlius, J., S. Langensiepen, et al. (2005). "Recombinant human erythropoietin and overall survival in cancer patients: results of a comprehensive meta-analysis." J Natl Cancer Inst 97(7): 489-98.
Bombardier, C., L. Laine, et al. (2000). "Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group.[see comment]." New England Journal of Medicine 343(21): 1520-8.
Bornhoft, G., S. Maxion-Bergemann, et al. (2006). "Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity." BMC Med Res Methodol 6: 56.
Bradburn, M. J., J. J. Deeks, et al. (2007). "Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events." Stat Med 26(1): 53-77.
Bravata, D. M., K. M. McDonald, et al. (2005). "Challenges in systematic reviews: synthesis of topics related to the delivery, organization, and financing of health care." Ann Intern Med 142(12_Part_2): 1056-1065.
Breekveldt-Postma, N. S., J. Koerselman, et al. (2007). "Enhanced persistence with tiotropium compared with other respiratory drugs in COPD." Respir Med 101(7): 1398-405.
Brockwell, S. E. and I. R. Gordon (2001). "A comparison of statistical methods for meta-analysis." Statistics in Medicine 20(6): 825-40.
Bucher, H. C., G. H. Guyatt, et al. (1997). "The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials." Journal of Clinical Epidemiology 50(6): 683-91.
Cairns, J. A., S. J. Connolly, et al. (1997). "Randomised trial of outcome after myocardial infarction in patients with frequent or repetitive ventricular premature depolarisations: CAMIAT. Canadian Amiodarone Myocardial Infarction Arrhythmia Trial Investigators. ." Lancet 349(9053): 675-682.
Caldwell, D. M., A. E. Ades, et al. (2005). "Simultaneous comparison of multiple treatments: combining direct and indirect evidence." BMJ 331(7521): 897-900.
Carey, T. S. and S. D. Boden (2003). "A critical guide to case series reports." Spine 28: 1631-1634.
Carless, P., A. Moxey, et al. (2005). "Are antifibrinolytic drugs equivalent in reducing blood loss and transfusion in cardiac surgery? A meta-analysis of randomized head-to-head trials." BMC Cardiovasc Disord 5: 19.
Cates, C. J. (2002). "Simpson's paradox and calculation of number needed to treat from meta-analysis." BMC Med Res Methodol 2: 1.
Chan, A. and D. Altman (2005). "Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors." BMJ 330(7494): 753.
Chan, A., A. Hrobjartsson, et al. (2004). "Empirical evidence for selective reporting of outcomes in randomized trials." JAMA 291(20): 2457-65.
Chan, A., K. Krleza-Jerić, et al. (2005). "Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research." CMAJ 171(7): 735-40.
Charlson, M. E. and R. I. Horwitz (1984). "Applying results of randomised trials to clinical practice: impact of losses before randomisation." British Medical Journal Clinical Research Ed 289(6454): 1281-4.
Charlton, B. G. (2001). "Fundamental deficiencies in the megatrial methodology." Current Controlled Trials in Cardiovascular Medicine 2: 2-7.

Chou, R., R. Fu, et al. (2006). "Methodological shortcomings predicted lower harm estimates in oneof two sets of studies of clinical interventions." J Clin Epidemiol 60(1): 18-28.
Chou, R., R. Fu, et al. (2006). " Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses." Lancet 368(9546): 1503-15.
Chou, R. and M. Helfand (2005). "Challenges in systematic reviews that assess treatment harms." Annals of Internal Medicine 142(12 Pt 2): 1090-9.
Cochrane Collaboration (2006). "The Cochrane Handbook for Systematic Reviews if Interventions 4.2.6."
Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, N.J., L. Erlbaum Associates.
Concato, J., N. Shah, et al. (2000). "Randomized, controlled trials, observational studies, and the hierarchy of research designs." New England Journal of Medicine 342(25): 1887-92.
Copas, J. and D. Jackson (2004). "A bound for publication bias based on the fraction of unpublished studies." Biometrics 60(1): 146-53.
Counsell, C. (1997). "Formulating questions and locating primary studies for inclusion in systematic reviews." Annals of Internal Medicine 127(5): 380-7.
Covey, J. (2007). "A Meta-analysis of the Effects of Presenting Treatment Benefits in Different Formats." Medical Decision Making 27: 638.
CRD, N. (1996). Undertaking systematic reviews of research on effectiveness (CRD Report 4). 1996. NHS Centre for Reviews and Dissemination. . York, The University of York.
CRD, N. (2001). Undertaking systematic reviews of research on effectiveness (CRD Report 4, 2nd ed). York, UK, University of York, Centre for Reviews and Dissemination (CRD).
CRD, N. (2007, 1-26-07). "Review methods and resources." from http://www.york.ac.uk/inst/crd/crdreview.htm.
Crumley, E. T. and N. Wiebe (2005). "Which resources should be used to identify RCT/CCTs for systematic reviews: a systematic review." BMC Medical Research Methodology 5(24).
Curtin, F., D. G. Altman, et al. (2002). "Meta-analysis combining parallel and cross-over clinical trials. I: Continuous outcomes." Stat Med 21(15): 2131-44.
Curtin, F., D. Elbourne, et al. (2002). "Meta-analysis combining parallel and cross-over clinical trials. II: Binary outcomes." Stat Med 21(15): 2145-59.
Dalziel, K., A. Round, et al. (2005). "Do the findings of case series studies vary significantly according to methodological characteristics?" Health Technol Assessment 9(2): 1-146.
Danan, G. and C. Benichou (1993). "Causality assessment of adverse reactions to drugs--I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries." J Clin Epidemiol 46(11): 1323-30.
Davis, C. E., W. B. Applegate, et al. (1995). "An empirical evaluation of the placebo run-in." Controlled Clinical Trials 16(1): 41-50.
de Gans, J. and D. van de Beek (2002). "Dexamethasone in adults with bacterial meningitis." N Engl J Med 347: 1549-1556.
Deeks, J. J. (2002). "Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes." Statistics in Medicine 21(11): 1575-600.
Deeks, J. J., J. Dinnes, et al. (2003). "Evaluating non-randomised intervention studies " Health Technology Assessment (Winchester, England) 7(27): iii-x, 1-173.

Derry, S., Y. Kong Loke, et al. (2001). "Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials." BMC Medical Research Methodology 1(1): 7.
DerSimonian, R. and N. Laird (1986). "Meta-analysis in clinical trials." Controlled Clinical Trials 7(3): 177-88.
Devereaux, P. J. and S. Yusuf (2003). "The evolution of the randomized controlled trial and its role in evidence-based decision making." Journal of General Internal Medicine 254(2): 105-13.
Dieppe, P., C. Bartlett, et al. (2004). "Balancing benefits and harms: the example of non-steroidal anti-inflammatory drugs." BMJ 329(7456): 31-34.
Donner, A. and N. Klar (2002). "Issues in the meta-analysis of cluster randomized trials." Stat Med 21(19): 2971-80.
Donner, A., G. Piaggio, et al. (2001). "Statistical methods for the meta-analysis of cluster randomization trials." Stat Methods Med Res 10(5): 325-38.
Easterbrook, P. J., J. A. Berlin, et al. (1991). "Publication bias in clinical research." Lancet 337(8746): 867-72.
Edwards, I. R. and J. K. Aronson (2000). "Adverse drug reactions: definitions, diagnosis, and management." Lancet 356(9237): 1255-1259.
Edwards, J., H. McQuay, et al. (1999). "Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain." J Pain Symptom Manage 18: 427-437.
Egger, M., P. Juni, et al. (2003). "How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study." Health Technology Assessment (Winchester, England) 7(1): 1-76.
Egger, M., M. Schneider, et al. (1998). "Spurious precision? Meta-analysis of observational studies." BMJ 316(7125): 140-4.
Egger, M. and G. D. Smith (2001). Principles of and procedures for systematic reviews [book chapter]. Systematic Review in Health Care: Meta-analysis in Context. M. Egger, G. D. Smith and D. G. Altman. London, England, BMJ Publishing Group: 23-42.
Elbourne, D. R., D. G. Altman, et al. (2002). "Meta-analyses involving cross-over trials: methodological issues." Int J Epidemiol 31(1): 140-9.
Ernst, E. and M. H. Pittler (2001). "Assessment of therapeutic safety in systematic reviews: literature review." BMJ 323(7312): 546-547.
Etminan, M. (2004). "Pharmacoepidemiology II: the nested case-control study--a novel approach in pharmacoepidemiologic research." Pharmacotherapy 24(9): 1105-1109.
Etminan, M. and A. Samii (2004). "Pharmacoepidemiology I: a review of pharmacoepidemiologic study designs." Pharmacotherapy 24(8): 964-969.
FDA (2006). Clinical review: Relationship between antidepressant drugs and suicidality in adults. , Department of Health and Human Services (DHHS), Public Health Service (PHS), Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER): 140.
Fergusson, D., S. Doucette, et al. (2005). "Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials." BMJ 330(7488): 396-.
Ferreira-Gonzalez, I., G. Permanyer-Miralda, et al. (2007). "Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials." BMJ 334(7597): 786.

Fraser, C., A. Murray, et al. (2006). "Identifying observational studies of surgical interventions in MEDLINE and EMBASE." BMC Medical Research Methodology 6: 41.
Freemantle, N. and M. Calvert (2007). "Composite and surrogate outcomes in randomised controlled trials [editorial]. ." BMJ 334(7597): 756-57.
Friedrich, J., N. Adhikari, et al. (2007). "Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data." BMC Med Res Methodol 7(5).
Furlan, A. D., E. Irvin, et al. (2006). "Limited search strategies were effective in finding relevant nonrandomized studies." J Clin Epidemio 59: 1303-11.
Fullerton, D. S. P. and D. S. Atherly (2004). "Formularies, Therapeutics, and Outcomes New Opportunities." Med Care 42: III-39 –III-44.
Furukawa, T. (1999). "From effect size into number needed to treat." Lancet 353(9165): 1680.
Furukawa, T., A. Cipriani, et al. (2005). "Imputing response rates from means and standard deviations in meta-analyses." Int Clin Psychopharmacol 20(1): 49-52.
Furukawa, T. A., D. L. Streiner, et al. (2000). "Discrepancies among megatrials." Journal of Clinical Epidemiology 53(12): 1193-9.
Furukawa, T. A., N. Watanabe, et al. (2007). "Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. [letter]." JAMA 297(5): 468-70.
Gartlehner, G., R. A. Hansen, et al. (2006). Criteria for distinguishing effectiveness from efficacy trials in systematic reviews. A. f. H. R. a. Q. (AHRQ). Rockville, MD, Evidence-based Practice Center: RTI-University of North Carolina
Gartlehner, G., R. A. Hansen, et al. (2006). "A simple and valid tool distinguished efficacy from effectiveness studies." Journal of Clinical Epidemiology 59(10): 1040-8.
Gartlehner, G., R. A. Hansen, et al. (2007). "Comparative Effectiveness of Second-generation Antidepressants in the Pharmacologic Treatment of Adult Depression. Comparative Effectiveness Review No. 7. (Prepared by RTI-UNC under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm."
Glanville, J. M., C. Lefebvre, et al. (2006). "How to identify randomized controlled trials in MEDLINE: ten years on." J Med Libr Assoc. 94(2): 130-6.
Glasgow, R. E., L. W. Green, et al. (2006). "External validity: we need to do more." Annals of Behavioral Medicine 31(2): 105-8.
Glasgow, R. E., D. J. Magid, et al. (2005). "Practical clinical trials for translating research to practice: design and measurement recommendations." Medical Care 43(6): 551-7.
Glasziou, P., I. Chalmers, et al. (2007). "When are randomised trials unnecessary? Picking signal from noise." BMJ 334: 349-51.
Glasziou, P., J. P. Vandenbroucke, et al. (2004). "Assessing the quality of research." BMJ 328(7430): 39-41.
Glasziou, P. P. and L. M. Irwig (1995). "An evidence based approach to individualising treatment." Bmj 311(7016): 1356-9.
Glenny, A. M., D. G. Altman, et al. (2005). "Indirect comparisons of competing interventions." Health Technology Assessment 9(26): 148.
Gleser, L. J. and I. Olkin (1994). Stochastically dependent effect sizes. . The Handbook of Research Synthesis. H. Cooper and L. V. Hedges. New York, Rullell Sage Foundation: 339-355.

Godwin, M., L. Ruhland, et al. (2003). "Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. ." BMC Med Res Methodol 3: 28.
Golder, S., H. M. McIntosh, et al. (2006). "Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE." Health Info Libr J 23(1): 3-12.
Goodman, S. N. and J. A. Berlin (1994). "The use of predicted confidence intervals when planning experiments andn the misuses of power when interpreting results." Ann Intern Med 121: 200-206.
Gotzsche, P. C. (1989). "Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis." Controlled Clinical Trials 10(1): 31-56.
GRADE Working Group (2004). "Grading quality of evidence and strength of recommendations." BMJ 328: 1490-.
Green, L. W. and R. E. Glasgow (2006). "Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology." Evaluation & the Health Professions 29(1): 126-53.
Greenland, S. and A. Salvan (1990). "Bias in the one-step method for pooling study results." Stat Med 9: 247-252.
Greenland, S., J. J. Schlesselman, et al. (1986). "The fallacy of employing standardized regression coefficients and correlations as measures of effect." Am J Epidemiol 123(2): 203-8.
Gunnell, D., J. Saperia, et al. (2005). "Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review." BMJ 330(488): 385.
Hansen, R. A., G. Gartlehner, et al. (2005). "Efficacy and safety of second-generation antidepressants in the treatment of Major Depressive Disorder." Ann Intern Med 143: 415-426.
Hansen, R. A., G. Gartlehner, et al. (2007). "Functional Outcomes of Drug Treatment in Alzheimer's Disease: A systematic review and meta-analysis." Drugs Aging 24(2): 155-167.
Hardy, R. J. and S. G. Thompson (1998). "Detecting and describing heterogeneity in meta-analysis." Stat Med 17(8): 841-56.
Harris, R. P., M. Helfand, et al. (2001). "Current methods of the US Preventive Services Task Force: a review of the process." American Journal of Preventive Medicine 20(3 Suppl): 21-35.
Hartling, L., F. A. McAlister, et al. (2005). "Challenges in systematic reviews of therapeutic devices and procedures." Annals of Internal Medicine 142(12 Pt 2): 1100-11.
Hayashi, K. and A. M. Walker (1996). "Japanese and American reports of randomized trials: differences in the reporting of adverse effects." Control Clin Trials 17(2): 99-110.
Haynes, B. (1999). "Can it work? Does it work? Is it worth it?" BMJ 319: 652-653.
Heart Protection Study Collaborative, G. (2002). "MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial." Lancet 360(9326): 7-22.
Hedges, L. V. (1981). "Distribution theory for Glass's estimator of effect size and related estimators." Journal of Educational Statistics 6: 107-128.
Helfand, M. (2005). "Using evidence reports: progress and challenges in evidence-based decision making " Health Affairs 24(1): 123-7.

Hernandez, A. V., E. Boersma, et al. (2006). "Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading?" American Heart Journal 151(2): 257-64.
Higgins, J. and S. Green. (2005). "The Cochrane Collaboration. The Cochrane handbook for systematic reviews of interventions." 2006, from http://www.cochrane.org/resources/handbook/handbook.pdf.
Higgins, J. G., S, Ed. (2006). Cochrane handbook for systematic reviews of interventions Chichester, UK, John Wiley & Sons, Ltd.
Higgins, J. P. and S. G. Thompson (2002). "Quantifying heterogeneity in a meta-analysis." Stat Med 21(11): 1539-58.
Higgins, J. P., S. G. Thompson, et al. (2003). "Measuring inconsistency in meta-analyses." BMJ 327(7414): 557-60.
Holloway, J. D. and D. D. Schocken (1988). "CASS in retrospect: lessons from the randomized cohort and registry. Coronary Artery Surgery Study." Am J Med Sci. 295(5): 424-32.
Hrachovec, J. B. and M. Mora (2001). "Reporting of 6-month vs 12-month data in a clinical trial of celecoxib." JAMA 286(19): 2398.
Hutton, J. L. (2000). "Number needed to treat: Properties and problems." Journal of the Royal Statistical Society. Series A: Statistics in Society 163(3): 403-419.
Ioannidis, J. P. (2006). "Indirect comparisons: the mesh and mess of clinical trials." Lancet 368(9546): 1470-2.
Ioannidis, J. P. A. (2005). "Contradicted and initially stronger effects in highly cited clinical research [see comments]." JAMA 294(2): 218-28.
Ioannidis, J. P. A., S. J. W. Evans, et al. (2004). "Better reporting of harms in randomized trials: an extension of the CONSORT statement." Annals of Internal Medicine 141(10): 781-788.
Ioannidis, J. P. A., A. B. Haidich, et al. (2001). "Comparison of evidence of treatment effects in randomized and nonrandomized studies." JAMA 286(7): 821-30.
Ioannidis, J. P. A. and J. Lau (2001). "Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas." JAMA 285(4): 437-43.
Ioannidis, J. P. A. and J. Lau (2002). "Improving safety reporting from randomised trials." Drug safety 25(2): 77-84.
Jollis, J. G., M. Ancukiewicz, et al. (1993). "Discordance of Databases Designed for Claims Payment versus Clinical Information Systems: Implications for Outcomes Research." Ann Intern Med 119(8): 844-850.
Jonville-Bera, A. P., B. Giraudeau, et al. (2006). "Reporting of drug tolerance in randomized clinical trials: when data conflict with authors' conclusions." Ann Intern Med 144: 306-307.
Juni, P., D. G. Altman, et al. (2001). "Systematic reviews in health care: Assessing the quality of controlled clinical trials." BMJ 323(7303): 42-6.
Juni, P., L. Nartey, et al. (2004). "Risk of cardiovascular events and rofecoxib: cumulative meta-analysis." Lancet 364(9450): 2021-9.
Kearney, P. M., C. Baigent, et al. (2006). "Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-infalmmatory drugs increase the risk of atherothrombosis? Meta-analysis of randomized trials. ." BMJ 332: 1302-1308.
Kjaergard, L. L., J. Villumsen, et al. (2001). "Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses." Ann Intern Med 135(11): 982-9.
Kleinbaum, D. G., L. L. Kupper, et al. (1982). Epidemiologic Research. Principles and Quantitative Methods. Belmont, CA, Wadsworth.

Kotaska, A. (2004). "Inappropriate use of randomised trials to evaluate complex phenomena: case study of vaginal breech delivery." BMJ 329(7473): 1039-1042.
Kuper, H., A. Nicholson, et al. (2006). "Searching for observational studies: what does citation tracking add to Pub Med? A case study in depression and coronary heart disease." BMC Medical Research Methodology 6(1): 4.
Laporte, J. R., L. Ibanez, et al. (2004). "Upper gastrointestinal bleeding associated with the use of NSAIDs: new versus older agents." Drug Safety 27(6): 411-420.
Lau, J., J. P. Ioannidis, et al. (1998). "Summing up evidence: one answer is not always enough." Lancet 351(9096): 123-7.
Lawlor, D. A., G. Davey Smith, et al. (2004). "Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence?" Lancet 363(9422): 1724-7.
Lee, S. (2002). "Statistical review. Center for Drug Evaluation and Research." Retrieved March 14, 2007, from www.fda.gov/cder/foi/nda/2002/21-042S007_Vioxx_statr.pdf.
Lemeshow, A. R., R. E. Blum, et al. (2005). "Searching one or two databases was insufficient for meta-analysis of observational studies." Journal of Clinical Epidemiology 58(9): 867-73.
Lewis, J. D., R. Schinnar, et al. (2006). "Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research.
." Pharmacoepidemiol Drug Saf.
Lindbloom, E., B. G. Ewigman, et al. (2004). "Practice-based research networks: The laboratories of primary care research." Medical Care 42(4 (suppl)): III45-49.
Lohr, K. (2004). "Rating the strength of scientific evidence: relevance for quality improvement programs." Int J Qual Health Care 16(1): 9-18.
Loke, Y. and S. Derry (2001). "Reporting of adverse drug reactions in randomised controled trials--a systematic survey." BMC Clin Pharmacol 1: 3.
Loke, Y. K., S. Derry, et al. (2004). "A comparison of three different sources of data in assessing the frequencies if adverse reactions to amiodarone." British Journal of Clinical Pharmacology 57(5): 616-21.
Loke, Y. K., D. Price, et al. (2006). "Case reports of suspected adverse drug reactions--systematic literature survey of follow-up." BMJ 332(7537): 335-9.
Loke, Y. K., D. Price, et al. (2007). Cochrane Handbook. Appendix 6b. Including Adverse Effects. Cochrane Colloboration, Cochrane Colloboration.
Lu, G. and A. E. Ades (2004). "Combination of direct and indirect evidence in mixed treatment comparisons." Statistics in Medicine 23(20): 3105-24.
Lumley, T. (2002). "Network meta-analysis for indirect treatment comparisons." Statistics in Medicine 21(16): 2313-24.
Mallen, C., G. Peat, et al. (2006). "Quality assessment of observational studies is not commonplace in systematic reviews." Journal of Clinical Epidemiology 59(8): 765-9.
Mangano, D. T., I. C. Tudor, et al. (2006). "The risk associated with aprotinin in cardiac surgery.[see comment]." New England Journal of Medicine 354(4): 353-65.
Maria, V. A. J. and R. M. M. Victorino (1997). "Development and validation of a clinical scale for the diagnosis of drug-induced hepatitis." Hepatology 26: 664-669.
Martin, R. C. G., M. F. Brennan, et al. (2002). "Quality of complication reporting in the surgical literature." Annals of Surgery 235: 803-813.
Martinez, C., S. Rietbrock, et al. (2005). "Antidepressant treatment and the risk of fatal and non-fatal self harm in first episode depression: nested case-control study." BMJ 330(7488): 389-.

Matchar, D. B., E. V. Westermann-Clark, et al. (2005). "Dissemination of Evidence-based Practice Center Reports." Ann Intern Med 142(12_Part_2): 1120-1125.
McAlister, F. A., A. Laupacis, et al. (1999). "Users' Guides to the Medical Literature: XIX. Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect." Jama 282(14): 1371-7.
McAlister, F. A., S. E. Straus, et al. (1999). "Why we need large, simple studies of the clinical examination: the problem and a proposed solution. CARE-COAD1 group. Clinical Assessment of the Reliability of the Examination-Chronic Obstructive Airways Disease Group." Lancet 354(9191): 1721-4.
McAuley, L., B. Pham, et al. (2000). "Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses?" Lancet 356(9237): 1228-31.
McDonagh, M., S. Carson, et al. (2006) "Drug class review on atypical antipsychotic drugs." Volume, DOI:
McDonagh, M., M. Helfand, et al. (2004). "Hyperbaric oxygen therapy for traumatic brain injury: a systematic review of the evidence." Arch Phys Med Rehabil 85(7): 1198-204.
McGettigan, P. and D. Henry (2006). "Cardiovascular risk and inhibition of cyclooxygenase: A systematic review of the observational studies of selective and nonselective inhibitors of cyclooxygenase 2 " JAMA 296(epub).
McIntosh, H. M., N. F. Woolacott, et al. (2004). "Assessing harmful effects in systematic reviews." BMC Medical Research Methodology 4: 19
McIntosh, M. W. (1996). "The population risk as an explanatory variable in research synthesis of clinical trials." Stat Med 15(16): 1713-28.
Medical Research Council (2000). A framework for development and evaluation of RCTs for complex interventions to improve health [report]. London, England, Medical Reseach Council: 1-19.
Michel, D. J. and L. C. Knodel (1986). "Comparison of three algorithms used to evaluate adverse drug reactions." American Journal of Hospital Pharmacy 43(7): 1709-1714.
Moja, L. P., E. Telaro, et al. (2005). "Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study." BMJ 330(7499): 1053-7.
Montori, V. M., N. L. Wilczynski, et al. (2005). "Optimal search strategies for retrieving systematic reviews from Medline: analytical survey." BMJ 330(7482): 68.
Mosteller, F. (1996). "The promise of risk-based allocation trials in assessing new treatments [editorial]." American Journal of Public Health 86(5): 622-3.
Mulrow, C., P. Langhorne, et al. (1997). "Integrating heterogeneous pieces of evidence in systematic reviews." Annals of Internal Medicine 127(11): 989-95.
Naranjo, C. A., U. Busto, et al. (1981). "A method for estimating the probability of adverse drug reactions." Clin Pharmacol Ther 30(2): 239-45.
NCI (1999). Common Toxicity Criteria Manual, version 2.0, National Cancer Institute Cancer Therapy Evaluation Program.
NCI (2006). Common Terminology Criteria for Adverse Events v3.0 (CTCAE) National Cancer Institute, Cancer Therapy Evaluation Program.
NIAID (2004). Division of AIDS table for grading the severity of adult and pediatric adverse events., National Institute of Allergy and Infectious Diseases.
Nissen, S. E. and K. Wolski (2007). "Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes." N Engl J Med 356(10.1056/NEJMoa072761).

Norman, G., F. Sridhar, et al. (2001). "Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life." Med Care 39(10): 1039-47.
Norris, S. L. and D. Atkins (2005). "Challenges in using nonrandomized studies in systematic reviews of treatment interventions." Annals of Internal Medicine 142(12 pt 2): 1112-9.
Ofman, J. J., C. H. MacLean, et al. (2002). "A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs.[see comment]." Journal of Rheumatology 29(4): 804-12.
Oleson, O. (1999). Types of study design. Draft chapters for the Guidelines on Non-randomised studies in Cochrane reviews, The Cochrane Non-Randomised Studies Methods Group (NRSMG): Chapter 2.
Olkin, I. (1994). "Re: A critical look at some popular meta-analytic methods [comment]." American Journal of Epidemiology 140(3): 297-9; discussion 300-1.
Olsen, H., T. Klemetsrud, et al. (1999). "Adverse drug reactions in current antihypertensive therapy: A general practice survey of 2586 patients in Norway." Blood Pressure 8: 94-101.
Oremus, M. M., M. Hanson, et al. (2006). The Uses of Heparin to Treat Burn Injury. Evidence Report Technology. M. U. E.-b. P. Center. Rockville, AHRQ: 1-95.
Oxman, A. D. and G. H. Guyatt (1992). "A consumer's guide to subgroup analyses." Annals of Internal Medicine 116(1): 78-84.
Papanikolaou, P., N, G. D. Christidi, et al. (2006). "Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies." CMAJ 174(5): 635-41
Papanikolaou, P. N., R. Churchill, et al. (2004). "Safety reporting in randomized trials of mental health interventions." Am J Psychiatry 161: 1692-1697.
Parmar, M. K., V. Torri, et al. (1998). "Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints." Stat Med 17(24): 2815-34.
Piaggio, G., D. R. Elbourne, et al. (2006). "Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement." JAMA 295(10): 1152-60.
Pignone, M., S. Saha, et al. (2005). "Challenges in Systematic Reviews of Economic Analyses." Ann Intern Med 142(12_Part_2): 1073-1079.
Pocock, S. J., S. E. Assmann, et al. (2002). "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems." Statistics in Medicine 21(19): 2917-30.
Poole, C. and S. Greenland (1999). "Random-effects meta-analyses are not always conservative." Am J Epidemiol 150(5): 469-75.
Psaty, B. M., C. D. Furberg, et al. (2004). "Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis." JAMA 292(21): 2622-31.
Psaty, B. M., T. Koepsell, et al. (1999). "Assessment and control for confounding by indication in observational studies." Journal of the American Geriatrics Society 47(6): 749-754.
Ray, W. A. (2003). "Evaluating medication effects outside of clinical trials: new-user designs." American Journal of Epidemiology 158(9): 915-20.
Ray, W. A. (2003). "Population-based studies of adverse drug effects." New England Journal of Medicine 349: 1592-1594.
Richardson, W. S., M. C. Wilson, et al. (1995). "The well-built clinical question: a key to evidence-based decisions." ACP J Club 123(3): A12-3.

Ridker, P. M. and J. Torres (2006). "Reported Outcomes in Major Cardiovascular Clinical Trials Funded by For-Profit and Not-for-Profit Organizations: 2000-2005." JAMA 295(19): 2270-2274.
Rosmarakis, E. S., E. S. Soteriades, et al. (2005). "From conference abstract to full paper: differences between data presented in conferences and journals." FASEB J. 19(7): 673-680.
Rothwell, P. M. (2005). "External validity of randomised controlled trials: "to whom do the results of this trial apply?"" Lancet 365(9453): 82-93.
Rothwell, P. M. (2005). "Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation." Lancet 365(9454): 176-86.
Rothwell, P. M., J. Slattery, et al. (1996). "A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis." Stroke 27(2): 260-265.
Royle, P. and R. Milne (2003). "Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches." International Journal of Technology Assessment in Health Care 19(4): 591-603.
Sackett, D. (2005). The principles behind the tactics of performing therapeutic trials. Clinical Epidemiology: How to Do Clinical Practice Research. R. B. S. Haynes, David L, Guyatt, Gordon H. and Tugwell, Peter. New York, Lippincott Williams & Wilkins.
Sackett, D. L. (2004). "Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis." ACP Journal Club 140(2): A11.
Sampson, M., L. Zhang, et al. (2006). "An alternative to the hand searching gold standard: validating methodological search filters using relative recall." BMC Medical Research Methodology 6: 33.
Sanderson, S., I. D. Tatt, et al. (2007). "Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography." Int J Epidemiol 36(3): 666-76.
Sankey, W. LA, et al. (1996). "An assessment of the use of the continuity correction for sparse data in meta-analysis." Communications in Statistics – Simulation and Computation 25: 1031-56.
Santaguida, P. L., M. Helfand, et al. (2005). "Challenges in systematic reviews that evaluate drug efficacy or effectiveness." Annals of Internal Medicine 142(12 Pt 2): 1066-72.
Santaguida, P. L. and P. Raina (May 2005). The Development of a Quality Assessment Scale Specific to Harms in Studies evaluating the efficacy of Health
Technologies. Report prepared for a funded study by the Canadian Centre for Health Technology Assessment (CCOHTA). Ottawa, Canada.
Savoie, I., D. Helmer, et al. (2003). "Beyond MEDLINE: reducing bias through extended systematic review search." International Journal of Technology Assessment in Health Care 19(1): 168-78.
Schmid, C. H., J. C. Cappelleri, et al. (2004). "Bayesian methods to improve sample size approximations." Methods in Enzymology 383: 406-27.
Schmid, C. H., J. Lau, et al. (1998). "An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials." Statistics in Medicine 17(17): 1923-42.
Schneeweiss, S. and J. Avorn (2005). "A review of uses of health care utilization databases for epidemiologic research on therapeutics." Journal of Clinical Epidemiology 58: 323-337.
Schulman, K. A., J. A. Berlin, et al. (1999). "The effect of race and sex on physicians' recommendations for cardiac catheterization." N Engl J Med 340(8): 618-626.

Schunemann, H. J., R. Jaeschke, et al. (2006). "An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations." Am J Respir Crit Care Med 174(5): 605-14.
Schwartz, L. M., S. Woloshin, et al. (1999). "Misunderstandings about the effects of race and sex on physician's referrals for cardiac catheterization." N Engl J Med 341(4): 279-83; discussion 286-7.
Senn, S. (1997). Statistical issues in drug development. Chichester ; New York, John Wiley.
Shadish, W. R., T. D. Cook, et al. (2002). Experimental and Quasi Experimental Designs for Generalized Causal Inference. Boston, Houghton-Mifflin.
Shah, R. V., T. J. Albert, et al. (2005). "Industry support and correlation to study outcome for papers published in Spine." Spine 30: 1099-1104.
Sharp, S. J. and S. G. Thompson (2000). "Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches." Stat Med 19(23): 3251-74.
Shekelle PG, Morton SC, et al. (2004). Pharmacological and Surgical Treatment of Obesity. Summary, Evidence Report/Technology Assessment: Number 103. AHRQ Publication Number 04-E028-1. Rockville, MD., Agency for Healthcare Research and Quality
Shekelle, P. G., S. C. Morton, et al. (2005). "Challenges in Systematic Reviews of Complementary and Alternative Medicine Topics." Ann Intern Med 142(12_Part_2): 1042-1047.
Shepherd, J., D. B. Hunninghake, et al. (2004). "Safety of rosuvastatin." The American Journal of Cardiology 94(7): 882-888.
Shojania, K. G. and L. A. Bero (2001). "Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy." Eff Clin Pract 4(4): 157-62.
Silverstein, F. E., G. Faich, et al. (2000). "Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: A randomized controlled trial. Celecoxib Long-term Arthritis Safety Study.[see comment]." JAMA 284(10): 1247-55.
Sjostrom, C. D., L. Lissner, et al. (1999). "Reduction in incidence of diabetes, hypertension and lipid disturbances after intentional weight loss induced by bariatric surgery: the SOS Intervention Study." Obes Res 7(5): 477-84.
Slavin, R. E. (1995). "Best evidence synthesis: an intelligent alternative to meta-analysis." Journal of Clinical Epidemiology 48(1): 9-18.
Smeeth, L., A. Haines, et al. (1999). "Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading." BMJ 318(7197): 1548-51.
Smith, A. H., P. A. Lopipero, et al. (1995). "Meta-analysis of studies of lung cancer among silicotics. ." Epidemiology 6(6): 617-624.
Smith, T. C., D. J. Spiegelhalter, et al. (1995). "Bayesian approaches to random-effects meta-analysis: a comparative study." Stat Med 14(24): 2685-99.
Song, F., D. G. Altman, et al. (2003). "Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses." BMJ 326(7387): 472.
Song, F., A. M. Glenny, et al. (2000). "Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery." Control Clin Trials 21(5): 488-97.

Song, F. J., A. Fry-Smith, et al. (2004). "Identification and assessment of ongoing trials in health technology assessment reviews." Health Technology Assessment 8(44): 1-87.
Steg, P. G., J. Lopez-Sendon, et al. (2007). "External validity of clinical trials in acute myocardial infarction." Arch Intern Med 167(1): 68-73.
Sterne, J. A., M. Egger, et al. (2001). "Investigating and dealing with publication and other biases in meta-analysis." BMJ 323(7304): 101-5.
Stricker, B. H. and B. M. Psaty (2004). "Detection, verification, and quantification of adverse drug reactions." BMJ 329(7456): 44-47.
Strom, B. (2005). What is pharmacoepidemiolgy? Pharmacoepidemiolgy. B. Strom. Chister, UK, John Wiley & Sons Ltd.: 3-16.
Strom, B. L. (2004). "Potential for Conflict of Interest in the Evaluation of Suspected Adverse Drug Reactions: A Counterpoint." JAMA 292(21): 2643-2646.
Sweeting, M. J., A. J. Sutton, et al. (2004). "What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data." Stat Med 23(9): 1351-75.
Sydes, M. R., D. J. Spiegelhalter, et al. (2004). "Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials." Clinical Trials 1: 60-79.
Tatsioni, A., D. A. Zarin, et al. (2005). "Challenges in systematic reviews of diagnostic technologies." Annals of Internal Medicine 142(12 pt 2): 1048-55.
Thompson, S. G. and J. P. Higgins (2002). "How should meta-regression analyses be undertaken and interpreted?" Stat Med 21(11): 1559-73.
Thompson, S. G., T. C. Smith, et al. (1997). "Investigating underlying risk as a source of heterogeneity in meta-analysis." Stat Med 16(23): 2741-58.
Toma, M., F. A. McAlister, et al. (2006). "Transition from meeting abstract to full-length journal article for randomized controlled trials." JAMA 295(11): 1281-7.
Treadwell, J. R., S. J. Tregear, et al. (2006). "A system for rating the stability and strength of medical evidence." BMC Medical Research Methodology 6(52).
Tubach, F., M. Dougados, et al. (2006). "Feeling good rather than feeling better matters more to patients." Arthritis Rheum 55(4): 526-30.
Tubach, F., P. Ravaud, et al. (2005). "Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement." Ann Rheum Dis 64(1): 29-33.
Tucker, J. A. and D. L. Roth (2006). "Extending the evidence hierarchy to enhance evidence-based practice for substance use disorders." Addiction 101(7): 918-32.
Tunis, S. R., D. B. Stryer, et al. (2003). "Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy." JAMA 290(12): 1624-32.
Ukoumunne, O. C., M. C. Gulliford, et al. (1999). "Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review." Health Technol Assess 3(5): iii-92.
Vandenbroucke, J. P. (2004). "Benefits and harms of drug treatments." BMJ 329(7456): 2-3.
Vandenbroucke, J. P. (2004). "When are observational studies as credible as randomised trials?" Lancet 363(9422): 1728-31.
Venning, G. R. (1982). "Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms." BMJ 284: 249-252.
Walach, H., T. Falkenberg, et al. (2006). "Circular instead of hierarchical: methodological principles for the evaluation of complex interventions." BMC Medical Research Methodology 6: 29.

Wald, N. J. and J. K. Morris (2003). "Teleoanalysis: combining data from different types of study." BMJ 327(7415): 616-8.
Ware, J. H. and E. M. Antman (1997). "Equivalence trials." N Engl J Med 337(16): 1159-1161.
wCurtin, F., D. Elbourne, et al. (2002). "Meta-analysis combining parallel and cross-over clinical trials. III: The issue of carry-over." Stat Med 21(15): 2161-73.
West, S., V. King, et al. (2002). Systems to rate the strength of scientific evidence. Rockville, MD, Agency for Healthcare Research & Quality.
White, I. R. and J. Thomas (2005). "Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis." Clin Trials 2(2): 141-51.
Whitlock, E. P., C. T. Orleans, et al. (2002). "Evaluating primary care behavioral counseling interventions: an evidence-based approach." American Journal of Preventive Medicine 22(4): 267-84.
Whittington, C. J., T. Kendall, et al. (2004). "Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data." Lancet 363(9418): 1341-5.
Wilczynski, N. L. and B. Haynes (2006). "EMBASE search strategies achieved high sensitivity and specificity for retrieving methodologically sound systematic reviews
" J Clin Epidemio.
Williamson, P. R. and C. Gamble (2007). "Application and investigation of a bound for outcome reporting bias." Trials 8: 9.
Witter, J. (2000). "Celebrex Capsules (Celecoxib) NDA 20-998/S-009 Medical Officer Review." Retrieved 21 Dec, 2005, from http://www.fda.gov/ohrms/dockets/ac/01/briefing/3677b1_03_med.pdf.
Wood, A. M., I. R. White, et al. (2004). "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals." Clinical Trials 1: 368-376.
Wood, L. and C. Martine (2004). "The General Practice Research Database: role in pharmacovigilance." Drug safety 27(12): 871-81.
Woolf, S. H. (1996). Manual for conducting systematic reviews. Agency for Health Care Policy and Research., AHRQ: 77p.
Woolf, S. H., C. G. DiGuiseppi, et al. (1996). "Developing evidence-based clinical practice guidelines: lessons learned by the US Preventive Services Task Force." Annual Review of Public Health 17: 511-38.
Yusuf, S., J. Wittes, et al. (1991). "Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials." Jama 266(1): 93-8.
Zarin, D. A., J. L. Young, et al. (2005). "Challenges to evidence-based medicine: a comparison of patients and treatments in randomized controlled trials with patients and treatments in a practice research network." Soc Psychiatry Psychiatr Epidemiol 40(1): 27-35.
Zhang, J., E. L. Ding, et al. (2006). "Adverse Effects of Cyclooxygenase 2 Inhibitors on Renal and Arrhythmia Events: Meta-analysis of Randomized Trials." JAMA 296: 1619-1632.
Zhang, J. and K. F. Yu (1998). "What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes " JAMA 280(19): 1690-1.
Zhang, L., I. Ajiferuke, et al. (2006). "Optimizing search strategies to identify randomized controlled trials in MEDLINE." BMC Medical Research Methodology 6: 23.