Methods Guide for Effectiveness and Comparative Effectiveness Reviews

September 2013

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Methods Guide for Effectiveness and Comparative Effectiveness Reviews

September 2013

This document was written with support from the Effective Health Care Program at the Agency for Healthcare Research and Quality (AHRQ). None of the authors has a financial interest in any of the products discussed in this document. The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ, the Veterans Health Administration, or the Health Services Research and Development Service. Therefore, no statement in this report should be construed as an official position of these entities, the U.S. Department of Health and Human Services, or the U.S. Department of Veterans Affairs.

Methods Guide for Effectiveness and Comparative Effectiveness Reviews

AHRQ Publication No. 10(13)-EHC063-EF
September 2013

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted for which further reproduction is prohibited without the specific permission of copyright holders.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(13)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2013. Chapters available at: www.effectivehealthcare.ahrq.gov.

Preface

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 (EHC) 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 Effectiveness and 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. This Guide presents issues key to the development of Effectiveness and Comparative Effectiveness Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The Methods Guide for Effectiveness and Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves. Comments and suggestions on the Methods Guide for Effectiveness and Comparative Effectiveness Reviews and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

Richard Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, Evidence-based Practice Center Program
Agency for Healthcare Research and Quality

Contents

Foreword. Comparing Medical Interventions: AHRQ and the Effective Health Care Program 1
Chapter 1. Principles in Developing and Applying Guidance for Comparing Medical Interventions 5
Chapter 2. Identifying, Selecting, and Refining Topics.....15
Chapter 3. Finding Evidence for Comparing Medical Interventions.....32
Chapter 4. Selecting Observational Studies for Comparing Medical Interventions.....56
Chapter 5. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions.....69
Chapter 6. Assessing the Applicability of Studies When Comparing Medical Interventions.....98
Chapter 7. Assessing Harms When Comparing Medical Interventions 112
Chapter 8. Expanded Guidance on Selected Quantitative Synthesis Topics.....145
Chapter 9. Conducting Quantitative Synthesis When Comparing Medical Interventions 160
Chapter 10. Grading the Strength of a Body of Evidence When Comparing Medical Interventions 178
Chapter 11. Using Existing Systematic Reviews To Replace De Novo Processes in Conducting Comparative Effectiveness Reviews.....196
Chapter 12. Updating Comparative Effectiveness Reviews: Current Efforts in AHRQ's Effective Health Care Program.....214
Chapter 13. Avoiding Bias in Selecting Studies.....228
Chapter 14. Handling Continuous Outcomes in Quantitative Synthesis 249

Foreword. Comparing Medical Interventions: AHRQ and the Effective Health Care Program

Jean Slutsky, David Atkins, Stephanie Chang, Beth A. Collins Sharp

Health care expenditures are growing faster than incomes for most developed countries, jeopardizing the stability of health care systems globally.[1] This trend has led to interest in knowledge about the most effective use of health care worldwide. To increase the value of health care services, many countries have established programs or independent agencies that inform health care decisionmaking through systematic reviews of technologies, pharmaceuticals, and other health care interventions. A few examples include the National Institute for Health and Clinical Excellence (NICE) in the United Kingdom, the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany, the Haute Autorité de Santé (HAS) in France, and the Canadian Agency for Drugs and Technologies in Health (CADTH). Some international consortiums and collaborations are also committed to increasing the use of evidence in health care decisionmaking. The Cochrane Collaboration has received international recognition for its sustained efforts at developing and disseminating systematic reviews. Additionally, Health Technology Assessment International (HTAi) is an organization with global membership that promotes evidence-based technology assessments.

By any measure, health care expenditures in the United States are increasing much faster than the health of the population and at a faster rate than in any other industrialized nation. Driven by the same goals as other countries and organizations—improving the quality, effectiveness, and efficiency of health care delivery—the U.S. Agency for Healthcare Research and Quality (AHRQ) created the Effective Health Care (EHC) Program in 2005.

A series of articles to be presented here in upcoming months give guidance on the methods to be used in conducting systematic reviews of technologies and interventions under the EHC Program, and together they form the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. While the various international programs and agencies mentioned here are united in their goal of providing objective assessments of effective health care interventions through systematic reviews, the varied health care system environments necessitate differences among the programs. For example, with the presence of a universal health system, NICE conducts cost-effectiveness studies, which are more difficult in a decentralized health care system. It is important to understand the context, principles, and philosophies of each program or agency, since they carry implications for the various approaches, methods, and end products of systematic reviews from the various groups.

The United States spent an estimated $1.8 trillion in 2005 on health care, including $342 billion under its Medicare program, with an annual estimated cost growth of 2.4 percent above the Gross Domestic Product.[2] Potential solutions for long-term solvency of the Medicare program for seniors and the disabled have been the cause of much political debate. This debate led to a series of Medicare reforms passed by Congress in 2003.[3] These reforms included a new drug benefit for seniors as well as new funding of $15 million annually for AHRQ (subsequently doubled to $30 million) to conduct and support research with a focus on the outcomes, comparative clinical effectiveness, and appropriateness of pharmaceuticals, devices, and health care services. Underlying this effort is a realization that improving value and controlling Medicare costs can be achieved only by understanding the relative effectiveness of the different

health care interventions at our disposal—both old and new. The EHC Program is guided by 14 priority conditions that are important to beneficiaries of the Medicare, Medicaid, and State Children's Health Insurance Program but would resonate with health care programs throughout the world.

The EHC Program involves the collaborative efforts of three major activities: systematic review, new research, and translation of findings for different audiences. Like the majority of the programs throughout the world, the EHC Program relies on systematic review methods to provide guidance on the effectiveness of therapeutics. The EHC program commissions 14 Evidence-based Practice Centers to perform the systematic reviews that provide an essential foundation from which to understand what we know from existing research and what critical research gaps remain. The Evidence-based Practice Centers undertake a broad variety of reviews that assess the effectiveness, comparative effectiveness, and comparative harms of different health care interventions. Some of these reviews are especially challenging in breadth and depth because the questions of most interest to decisionmakers often require complex comparisons. The EHC Program is supported by a Scientific Resource Center, which provides scientific and technical support to maintain consistency in the methods used across the different centers.

The EHC Program reflects in many ways the decentralized nature of the U.S. health care system. The audience includes not only policymakers in government and private health plans but also clinicians, patients, and members of industry, all of whom play a major role in health care decisionmaking. All of these stakeholders provide input and guidance to the program, all may contribute suggestions of new topics for assessment, and all have provided comments on drafts of the guidance given in this series. The EHC Program is meant to provide understandable and actionable information for patients, clinicians, and policymakers.

In order to provide useful information on effective health care interventions, the EHC Program follows three key principles that guide the EHC Program and, thus, the conduct of systematic reviews by the Evidence-based Practice Centers. First, reviews must be *relevant and timely* in order to meet the needs of decisionmakers. The questions being addressed in reviews must answer emerging and complex health care questions at the time when decisionmakers need the information. This means identifying the most important issues under the priority conditions and the optimal time to initiate a review. It also requires a conscientious effort to complete the review as quickly as possible without sacrificing the quality of the product.

Second, reviews must be *objective and scientifically rigorous*. To maintain the objectivity of a review, lead authors on the reports are barred from having any significant competing interests. In addition, although Evidence-based Practice Center staff, consultants, subcontractors, and other technical experts may not be disqualified from providing comments, they must disclose any financial, business, and professional interests that are related to the subject matter of a review or other product or that could be affected by the findings of the review. With respect to the types of financial interests to be disclosed, AHRQ is guided by the U.S. Department of Health and Human Services Regulations 45 CFR Part 94. Directors of the Evidence-based Practice Centers are responsible for the scientific integrity of all members of the review team by ensuring that they comply with AHRQ policy and by providing opportunities for training in rigorous scientific methods. There are a variety of sources for training in systematic review scientific methods in the United States and elsewhere. In addition to having the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* as a resource, AHRQ and the Scientific Resource Center have regularly scheduled conference calls with Evidence-based Practice Centers and face-to-face meetings biannually to discuss scientific methods and other aspects of

producing scientifically sound and credible systematic reviews. The Evidence-based Practice Centers participate in many scientific forums, and the work they do in methods informs the process and helps in collaborating with the work of similar groups in other countries.

Finally, *public participation and transparency* increase public confidence in the scientific integrity and credibility of reviews and provide further accountability to the Evidence-based Practice Centers. Reviews commissioned under the EHC Program are posted publicly at different stages of the review process, including the stage of proposed Key Questions and the draft report stage. Public posting of the processes and methodological approaches used in developing systematic reviews ensures that the reports are accessible, clear, and credible. The publication of this series of methods articles in the *Journal of Clinical Epidemiology* and the posting of the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* on the EHC Web site (www.effectivehealthcare.ahrq.gov) are fundamental ways of clearly laying out the EHC approach to conducting systematic reviews of comparative effectiveness.

The Evidence-based Practice Centers' work on Comparative Effectiveness Reviews builds on nearly 10 years of experience doing systematic reviews of diverse topics, including drugs and devices, diagnostic tests, and health care system interventions.[4] Unlike many other programs or agencies producing systematic reviews, which focus on evaluating individual interventions, the AHRQ EHC Program focuses on health care questions that require comparisons of alternative interventions for a given clinical condition.

In addition to the familiar issues raised in a systematic review or meta-analysis of a single intervention, there are specific challenges encountered in conducting Comparative Effectiveness Reviews. The methods papers in this series were written in response to these specific challenges.

The aim of a Comparative Effectiveness Review is to depict how the relative benefits and harms of a range of options compare, rather than to answer a narrow question of whether a single therapy is safe and effective. This requires a clear understanding of the clinical context to ensure that the review focuses on the appropriate population and interventions among which clinicians are currently choosing. As an example, our review of coronary artery bypass surgery vs. percutaneous coronary intervention for stable coronary disease focused on patients who have stable angina and two-vessel disease and on other subgroups for which clinicians might currently consider either option. It did not address patients at either clinical extreme, for whom the benefits of one option might be clear cut.

There is rarely a sufficient body of head-to-head trials to support easy conclusions about comparative benefits and harms. Providing useful information requires examining a broader array of literature, including placebo-controlled trials and observational studies; the latter are especially useful for looking more completely at harms, adherence, and persistence. In addition, reviews may examine whether, in the absence of head-to-head trials, indirect comparisons may be useful (e.g., comparing results of placebo-controlled trials of A and placebo-controlled trials of B).

Carefully examining the applicability of evidence is especially important. A useful review compares the tradeoffs of multiple alternatives, each of which may vary with the underlying population and setting. For example, the results of trials comparing the abilities of different oral diabetes drugs to control blood glucose may depend in important ways on the populations being studied. Evidence on harms is often hard to determine from tightly controlled randomized trials. Observational studies provide another check on whether results observed in trials appear to hold up under more representative settings and populations.

Finally, the interpretation of the evidence and the limits of interpretation are important. Equivalence of different treatments for a group of patients on average does not necessarily imply they are equivalent for all individuals. Attempts to explore subgroups for which benefits or harms of specific interventions vary may be needed. Often, however, there is limited evidence to support strong conclusions about the specific benefits of a particular intervention for subgroups.

The articles in this series reflect the final individual chapters of the EHC *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Written by AHRQ Evidence-based Practice Center investigators with the intention of improving both consistency and transparency in the EHC program, they were initially posted as one draft document for public comment on the EHC Web site in late 2007 and have been revised in response to public comment. Where there is an inadequate empiric evidence base, the articles review the existing guidance produced by different organizations and collaborations and build on these activities, focusing on issues specific to conducting Comparative Effectiveness Reviews. As the research methodologies develop, the EHC Program will continue to assess the need to update the current *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*.

Building a stronger empiric base for methods will increase transparency and consistency within and among the various groups that produce reviews of comparative effectiveness. In areas where empiric research is lacking, collaboration is paramount to determine best practices and to set a methods research agenda. Uniform guidance based on validated methods is essential to providing quality and consistent evidence for patients, clinicians, and policymakers, no matter where they live.

Author Affiliations

Agency for Healthcare Research and Quality, Rockville, MD, (JS, SC, BACS). Veterans Health Administration, Health Services Research and Development Service, Washington, DC, (DA).

This paper has also been published in edited form: Slutsky J, Atkins D, Chang S, et al. AHRQ Series Paper 1: Comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2010;63:481–483.

References

1. Kaiser Family Foundation. Healthcare Spending in the United States and OECD Countries. Available at: www.kff.org/insurance/snapshot/chcm010307oth.cfm. Accessed January 2007.
2. Congress of the United States, Congressional Budget Office. The Long-Term Outlook for Health Care Spending. Available at: www.cbo.gov/ftpdocs/87xx/doc8758/11-13-LT-Health.pdf. Accessed November 2007.
3. Medicare Prescription Drug, Improvement, and Modernization Act of 2008. Sec. 1013. Research on Outcomes of Health Care and Services. Public Law 108–173. 108th Congress.
4. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center Program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005;142:1035–41.

Chapter 1. Principles in Developing and Applying Guidance for Comparing Medical Interventions

Mark Helfand, Howard Balshem

Key Points

To be useful, Comparative Effectiveness Reviews must:

- Approach the evidence from a clinical, patient-centered perspective.
- Fully explore the clinical logic underlying the rationale for a service.
- Cast a broad net with respect to types of evidence, placing high-quality, highly applicable evidence about *effectiveness* at the top of the hierarchy.
- Present benefits and harms for different treatments and tests in a consistent way so that decisionmakers can fairly assess the important tradeoffs involved for different treatment or diagnostic strategies.
- CERs are empirically based whenever possible. When empirical evidence is not available or is inadequate, best practices should be defined to reduce variation among reviewers.

Introduction

Comparative Effectiveness Reviews (CERs) are summaries of available scientific evidence in which investigators collect, evaluate, and synthesize studies in accordance with an organized, structured, explicit, and transparent methodology. They seek to provide decisionmakers with accurate, independent, scientifically rigorous information for comparing the effectiveness and safety of alternative clinical options. CERs have become a foundation for decisionmaking in clinical practice and health policy. To play this important role in decisionmaking, CERs must address significant questions that are relevant to patients and clinicians, and they must use valid, objective, and scientifically rigorous methods to identify and synthesize evidence, applying these methods consistently and in an unbiased and transparent manner.

In this chapter, we describe the preliminary work and key principles that underlie the development of the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60). The chapters in this guide describe recommended approaches for addressing difficult, frequently encountered methodological issues. The science of systematic reviews is evolving and dynamic. However, excessive variation in methods among systematic reviews gives the appearance of arbitrariness and idiosyncrasy, which undercuts the goals of transparency and scientific impartiality.

Background and History

In 1997, the Agency for Healthcare Research and Quality (AHRQ) began its Evidence-based Practice Center (EPC) program. EPCs were established and staffed with personnel who had training and expertise in the conduct of systematic evidence reviews. From the inception of the program, the EPCs have been committed to developing methods for identifying and synthesizing evidence that minimize bias. EPCs adopted some precautions against bias in conducting evidence reviews that were extraordinary for their time. In 1996, for example, the

procedures used by EPCs, documented in AHRQ's *Manual for Conducting Systematic Reviews*,[1] included a requirement for the involvement of a technical expert panel to work with EPC scientists to develop the questions to be answered in the review as a way to protect against bias in framing or selecting questions. This approach helps ensure that a review will address important questions that decisionmakers need answered, and it also protects against bias in framing or selecting questions. Another protection against reviewer bias—using independent researchers, without conflicts of interest, to assess studies for eligibility—has also been used since the inception of the EPC program.

The *Methods Guide* is part of a broader system of safeguards to ensure that reviews produced by the EPCs are high quality, consistent, and fair.[2] Safeguards are needed because, as in any type of clinical research, the habits or views of investigators and funders can introduce bias, variation, or gaps in quality.[3-5] The framework for conducting systematic reviews includes strategies to reduce the possibility of bias at every step.[6,7]

The *Methods Guide* is a collaborative product of the 14 EPCs with oversight from the Scientific Resource Center (SRC). It serves as a resource for the Effective Health Care Program and scientists employed by AHRQ. To prioritize topics for the *Methods Guide*, we:

- Identified challenges in the production of AHRQ evidence reports and variation among EPCs.
- Examined public and peer-reviewed commentary on CERs.

In 2004 and 2005, each EPC analyzed published evidence reports and produced a series of articles identifying methodological challenges and areas of high practice variation among the EPCs. Topics included assessing beneficial[8] or harmful effects of interventions,[9] using observational studies,[10] assessing diagnostic tests[11] or therapeutic devices,[12] and others. When possible, the articles also suggested best practices.[13]

Through these approaches, we have identified concerns about inconsistent or poorly developed methods that are common across reports, such as:

- Inconsistency in approaches to quantitative synthesis, such as the choice of a fixed- or random-effects model.
- Inconsistency in the selection of data sources and evaluation of their quality for assessment of harms.
- A weakly developed approach to assessing the strength of evidence and a desire to begin to reconcile the EPC and GRADE (Grading of Recommendations Assessment, Development and Evaluation) approaches.
- A need to develop a consistent and structured approach to the assessment of applicability.

We used this preliminary work to select the key issues for the first version of the *Methods Guide*. To address these issues, AHRQ established five workgroups made up of EPC investigators, AHRQ staff, and SRC staff. The five workgroups developed guidance on observational studies, applicability, harms and adverse effects, quantitative synthesis, and methods for rating a body of evidence. The workgroups identified relevant methods papers and reviewed the published guidance from major bodies producing systematic reviews—most importantly, the Cochrane Collaboration Handbook[14] and the Centre for Reviews and Dissemination manual on conducting systematic reviews.[15,16]

Principles—Developing Guidance

The fundamental principle used in the development of the *Methods Guid*e and subsequent guidance has been that workgroups should use empirical, methodological research when available. However, when empirical evidence is not available or is inadequate, workgroups are asked to develop a structural, best-practice approach based on the principle that the approach will eliminate or reduce variation in practice and provide a transparent and consistent methodological approach.

Searching databases of non-English-language publications, unpublished papers, and information published only in abstract form is an example of evidence-based guidance based on empirical research. Many publications on these topics exist,[17-19] and they form a cohesive and consistent body of evidence upon which recommendations can be made.

On the other hand, structural approaches designed to reduce variation in practice and assure consistency across EPCs have also been adopted. Examples are:

- Centralization at the SRC of activities where EPC proficiency and skill vary, such as searching clinical trial registries and the U.S. Food and Drug Administration (FDA) Web site.
- Adoption of strict policies regarding conflicts of interest.
- Introduction of an editorial review process that provides for an independent judgment of the adequacy of an EPC's response to public and peer review comments

Some of the most important structural components of the Effective Health Care Program are intended to ensure that patients' and clinicians' perspectives are heard by standardizing the governance of interactions with technical experts, stakeholders, and payers.

Principles—Conducting Comparative Effectiveness Reviews

In their charge, all workgroup participants were asked to make their guidance for conducting reviews consistent with the overarching principles of the Effective Health Care Program.[20] Principles for conducting reviews include:

- Approaching the evidence from a clinical, patient-centered perspective.
- Fully exploring the clinical logic underlying the rationale for a service.
- Casting a broad net with respect to types of evidence, placing a high value on effectiveness and applicability, in addition to internal validity.
- Presenting benefits and harms for different treatments and tests in a consistent way so that decisionmakers can fairly assess the important tradeoffs involved for different treatment or diagnostic strategies.

For example, to follow the principle of patient-centeredness, the Program encourages EPCs to use absolute measures whenever possible to promote better communication with patients and others who will use the reports. Similarly, the program has been aggressive in involving stakeholders at every step of the process to ensure public participation and transparency.[21]

The EPCs' approach to evidence synthesis incorporates important insights from clinical epidemiology, health technology assessment, outcomes research, and the science of decisionmaking.[22,23] These principles for conducting reviews reflect the EPC program's longstanding commitment to developing evidence reports that individuals and groups can use to

make decisions and that are relevant, timely, objective, and scientifically rigorous and to provide for public participation and transparency.

Clinical and Patient-Centered Perspective

Whoever the intended users are, a CER should focus on patients' concerns. As Black notes, "There is no inherent antithesis between patient-oriented medicine and evidence-based medicine; focus on what is perceived by the individual patient does not rule out a systematic search for evidence relevant to his treatment."[24] Patients' preferences and patient-centered care are fundamental principles of evidence-based medicine.[25] These principles mean that, regardless of who nominates a topic and who might use CERs, the reviews should address the circumstances and outcomes that are important to patients and consumers. Studies that measure health outcomes (events or conditions that the patient can feel and report on, such as quality of life, functional status, or fractures) are emphasized over studies of intermediate outcomes (such as changes in blood pressure levels or bone density). Reviews should also take into account the fact that, for many outcomes and decisions, variation in patients' values and preferences can and should influence decisions.[26] 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.[8] AHRQ has developed explicit processes for topic selection and refinement and for the development of key questions to ensure that CERs are patient centered and also meet the needs of other stakeholders.[21]

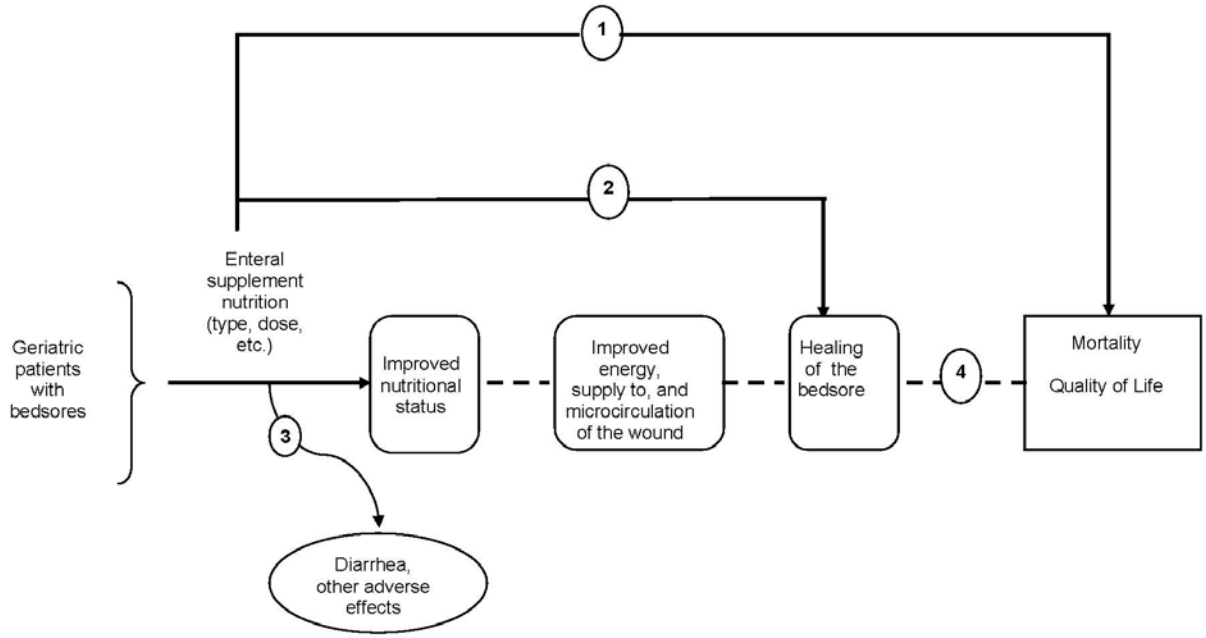
Clinical Logic and Analytic Frameworks

An evidence model is a critical element for fully exploring the clinical logic underlying the rationale for a service.[27] In the EPC program, the most commonly used evidence model is the "analytic framework."[28,29] The analytic framework portrays relevant clinical concepts and the clinical logic underlying beliefs about the mechanism by which interventions may improve health outcomes.[30] In particular, the analytic framework illustrates and clarifies the relationship between surrogate or intermediate outcome measures (such as cholesterol levels) and health outcomes (such as myocardial infarctions or strokes).[31] When properly constructed, it can provide an understanding of the context in which clinical decisions are made and illuminate disagreements about the clinical logic that underlie clinical controversies.

An analytic framework can also help clarify implicit assumptions about benefits from health care interventions, including assumptions about long-term effects on quality of life, morbidity, and mortality. These assumptions often remain obscure without a framework that can lead technical experts and manufacturers of drugs and devices to make explicit the reasoning behind their clinical theories linking surrogate outcomes, pathophysiology, and other intermediate factors to outcomes of interest to patients, clinicians, and other health care decisionmakers.

Figure 1 depicts an analytic framework for evaluating studies of a new enteral supplement to heal bedsores. Key questions are associated with the links (arrows) in the analytic frameworks. When available, evidence that directly links interventions to the most important health outcomes is more influential than evidence from other sources. In the figure, Arrow 1 corresponds to the question (Key Question 1): Does enteral supplementation improve mortality and quality of life?

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



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 supplements improves healing (link #2), EPCs may need to evaluate additional bodies of evidence. Specifically included would be evidence linking enteral supplementation to improved nutritional status and other evidence linking nutritional status to wound healing. Studies that measure health outcomes directly are given more weight, but the analytic framework makes clear what surrogate outcomes may represent them and what bodies of evidence link the surrogate outcomes to health outcomes.

Types of Evidence

Historically, evidence-based medicine has been associated with a hierarchy of evidence that ranks randomized trials higher than other types of evidence in all possible situations.[32,33] In recent years, broader use of systematic comparative effectiveness reviews has brought attention to the danger of over-reliance on randomized clinical trials and to suggestions for changing or expanding the hierarchy of evidence to take better account of evidence about adverse events and effectiveness in actual practice.[34-36]

AHRQ's EPC program from the outset has taken a broad view of eligible evidence.[1,37] AHRQ reviews published from 1997 through 2005 encompassed a wide variety of study designs, from randomized controlled trials (RCTs) to case reports. 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.[9-11,27,38,39]

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 high-quality, 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.

Discussions about the role of nonrandomized studies often focus on the limitations of RCTs and invoke the distinction between effectiveness and efficacy. Efficacy trials (explanatory trials) determine whether an intervention produces the expected result under ideal circumstances. Effectiveness studies use less stringent eligibility criteria, assess health outcomes, and have longer followup periods than most efficacy trials. Roughly speaking, effectiveness studies measure the degree of beneficial effect in "real-world" clinical settings.[40] The results of effectiveness studies are more applicable to the spectrum of patients who will use a drug, have a test, or undergo a procedure than results from highly selected populations in efficacy studies. Characteristics of efficacy trials that limit the applicability of their results include:

- Homogeneous populations. Trials may exclude patients from important subpopulations or those with relevant comorbidities.
- Small sample size.
- Limited duration.
- Focus on intermediate or surrogate outcomes.
- Selective focus on a limited number of intended or unintended effects.

In contrast, effectiveness studies aim to study patients who are likely to be offered the intervention in everyday practice. They also examine clinical strategies that are more representative of 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[41] and long-term outcomes that require longitudinal data collection to measure.

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. Effectiveness trials enable the investigator to obtain evidence about effectiveness while minimizing the risk of bias from confounding by indication and other threats to internal validity.[40,42-47] The ideal trial:

- Has good applicability to the patients, comparisons, setting, and outcomes important to patients and clinicians.
- Has a low risk of bias.
- Directly compares interventions.
- Reflects the complexity of interventions in practice.
- Includes all important intended and unintended effects, taking adherence and tolerability into account.

Often, RCTs are deficient in one or more of these respects. The decision to use other kinds of evidence—experimental or observational—should follow a critique of the applicability, risk of bias, directness, and completeness of the RCT evidence.[10] In addition to head-to-head effectiveness trials, types of evidence used in CERs include:

- Long-term head-to-head controlled trials focusing on a subset of relevant benefits or risks.
- Cohort, case-control, or before/after studies with broad applicability and comprehensive measurement of benefits and risks.
- Short-term head-to-head trials that use surrogate (efficacy) measures.
- Short-term head-to-head trials focusing on tolerability and side effects.
- Placebo-controlled trials demonstrating an important or unique benefit or harm of a particular drug.
- Before/after or time-series studies demonstrating an important or unique benefit or harm of a particular drug.
- Natural history (or conventionally treated history) studies that observe the outcomes of a cohort but do not compare the outcomes among different treatments.
- Case series and case reports.

In any particular review, any or all of these types of studies might be included or rendered irrelevant by stronger study types. Usually the reasons to include them overlap: RCTs may have poor applicability due to patient selection or inappropriate comparator or dosing of comparator; may not address all relevant intended effects; may not address all relevant unintended effects; or have few or only short-term head-to-head comparisons. Depending on the question, any of these types of studies might provide the best evidence to address gaps in the evidence from head-to-head effectiveness studies. Norris and colleagues offer further specific guidance on criteria for including observational studies in CERs in an upcoming chapter in this *Methods Guide*.

Balance of benefits and harms. CERs aim to present benefits and harms for different treatments and tests in a consistent way so that decisionmakers can fairly assess the important tradeoffs involved for different treatment or diagnostic strategies. The decisionmakers, not the reviewers, must weigh the benefits, harms, and costs of the alternatives. The reviewers, for their part, should seek to present the benefits and harms in a manner that helps with those decisions. The single most important feature of a good CER is that all important outcomes, rather than a selected subset of them, are described.

Expressing benefits in absolute terms (for example, a treatment prevents one event for every 100 treated patients) rather than in relative terms (for example, a treatment reduces events by 50 percent) can also help decisionmakers. Reviewers should highlight where evidence indicates that benefits, harms, and tradeoffs are different for distinct patient groups who, because of their personal characteristics, may be at higher or lower risk of particular adverse effects or may be more or less susceptible to complications of the underlying condition. Reviews should 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.

Future Development of the Methods Guide

Future chapters in this guide will look at:

- When and how to use observational studies.
- Assessing the applicability of studies.

- Assessing harms.
- Assessing the quality of studies.
- Finding evidence.
- Quantitative synthesis.
- Rating a body of evidence.

We have identified several gaps in the methodological literature that will be addressed through new guidance. We have also identified future research that is needed, including methodologies for the assessment of medical tests. Several groups are currently working on developing guidance for medical test assessment that will suggest a framework for the review of medical tests and will address issues such as when and how to use modeling, how to assess the quality of studies of medical tests, the relevance and consequences of the full range of patient outcomes on decisions to use a medical test, and the assessment of studies of genetic and prognostic tests.

For many of these issues, some variation in practice may persist because of differing opinions about the relative advantages of different approaches and a lack of sufficiently strong empirical evidence to dictate a single method. As further information accumulates, we expect to define more specific requirements related to these issues. We will continue to assess both the ability to implement our recommendations and the validity of the methods that we have adopted—both primary recommendations and secondary concepts introduced in the guidance—as we undertake comparative reviews on a wide assortment of topics. We anticipate the guidance will continue to evolve as we identify new issues and accumulate experience with new topic areas.

Author Affiliations

Oregon Health and Science University Evidence-based Practice Center, Portland, OR (MH, HB), Portland VA Medical Center, Portland, OR (MH).

This paper has also been published in edited form: Helfand M, Blashem H. AHRQ Series Paper 2: Principles for developing guidance: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2010;63:484–490.

References

1. Woolf SH. Manual for conducting systematic reviews. Agency for Health Care Policy and Research: 1996.
2. Agency for Healthcare Research and Quality. Suggesting a Topic for Effective Health Care Research. 2009. Available at: http://effectivehealthcare.ahrq.gov/documents/TopicFormRevExample.pdf. Accessed April 27, 2009.
3. Aschengrau A, Seage GR. Essentials of epidemiology in public health. Bartlett and Jones; 2003.
4. Mrkobrada M, Thiessen-Philbrook H, Haynes RB, et al. Need for quality improvement in renal systematic reviews. Clin J Am Soc Nephrol 2008 Jul;3(4):1102–14.
5. Shrier I, Boivin JF, Platt RW, et al. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? BMC Med Inf Decision Making 2008;8:19.
6. Egger M, Smith GD. Principles of and procedures for systematic reviews [book chapter]. In: Egger M, Smith GD, Altman DG, editors. Systematic Review in Health Care: Meta-analysis in Context. 2nd ed. London, England: BMJ Publishing Group; 2001. p. 23–42.

7. Moher D, Soeken K, Sampson M, et al. Assessing the quality of reports of systematic reviews in pediatric complementary and alternative medicine. BMC Pediatrics 2002;2(3).
8. Santaguida PL, Helfand M, Raina P. Challenges in systematic reviews that evaluate drug efficacy or effectiveness [review]. Ann Intern Med 2005 Jun 21;142(12 Pt 2):1066–72.
9. Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms [review]. Ann Intern Med 2005 Jun 21;142(12 Pt 2):1090–9.
10. Norris SL, Atkins D. Challenges in using nonrandomized studies in systematic reviews of treatment interventions [review]. Ann Intern Med 2005 June 21;142(12 Pt 2):1112–19.
11. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies [review]. Ann Intern Med 2005 Jun 21;142(12 pt 2):1048–55.
12. Hartling L, McAlister FA, Rowe BH, et al. Challenges in systematic reviews of therapeutic devices and procedures. Ann Intern Med 2005 Jun 21;142(12 Pt 2):1100–11.
13. Helfand M, Morton S, Guallar E, et al. A guide to this supplement. Ann Intern Med 2005 June 21, 2005;142(12 Pt 2):1033–34.
14. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. In: The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.
15. National Health Service Centre for Reviews and Dissemination. Undertaking systematic reviews of research on effectiveness (CRD Report 4, 2nd ed). York, UK: NHS Centre for Reviews and Dissemination. University of York; 2001 March. Report No. 4.
16. National Health Service Centre for Reviews and Dissemination. Review methods and resources. York, UK: NHS Centre for Reviews and Dissemination, The University of York; 2007 1-26-07.
17. Egger M, Zellweger-Zahner T, Schneider M, et al. Language bias in randomised controlled trials published in English and German. Lancet 1997 Aug 2;350(9074):326–9.
18. Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet 1996 Feb 10;347(8998):363–6.
19. Scherer RW, Dickersin K, Langenberg P. Full publication of results initially presented in abstracts. A meta-analysis. JAMA 1994 Jul 13;272(2):158–62.
20. Slutsky J, Atkins D, Chang S, et al. Comparing medical interventions: AHRQ and the effective health-care program [editorial]. J Clin Epidemiol 2008 Sep 30.
21. Whitlock EP, Lopez SA, Chang S, et al. Identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the Effective Health Care Program. J Clin Epidemiol [to be published].
22. Helfand M. Using evidence reports: progress and challenges in evidence-based decision making Health Aff 2005 Jan-Feb;24(1):123–7.
23. Drummond MF, Schwartz JS, Jönsson B, et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. Int J Technol Assess Health Care 2008;24(03):244–58.
24. Black D. POM + EBM = CPD? [editorial]. J Med Ethics 2000 Aug;26(4):229–230.
25. Guyatt GH, Montori VM, Devereaux PJ, et al. Patients at the centre: in our practice, and in our use of language [editorial]. Evidence-Based Med 2004;9(1):6–7.
26. Guyatt GH, Cook DJ, Haynes B. Evidence based medicine has come a long way [editorial]. BMJ 2004 Oct 30;329(7473):990–1.
27. Bravata DM, McDonald KM, Shojania KG, et al. Challenges in systematic reviews: synthesis of topics related to the delivery, organization, and financing of health care. Ann Intern Med 2005 June 21, 2005;142(12 Pt 2):1056–65.
28. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001 Apr;20(3 Suppl):21–35.
29. Whitlock EP, Orleans CT, Pender N, et al. Evaluating primary care behavioral counseling interventions: an evidence-based approach [review]. Am J Prev Med 2002 May;22(4):267–84.
30. Woolf SH, DiGuiseppi CG, Atkins D, et al. Developing evidence-based clinical practice guidelines: lessons learned by the US Preventive Services Task Force [review]. Ann Rev Public Health 1996;17:511–38.
31. Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. Ann Intern Med 1997 Dec 1;127(11):989–95.

32. Bigby M. Challenges to the hierarchy of evidence: does the emperor have no clothes? [article criticism]. Arch Dermatol 2001 Mar;137(Mar):345–6.
33. Devereaux PJ, Yusuf S. The evolution of the randomized controlled trial and its role in evidence-based decision making. J Intern Med 2003 Aug;254(2):105–13.
34. Shrier I, Boivin J-F, Steele RJ, et al. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. Am J Epidemiol 2007 Aug 21;166(10):1203–9.
35. Walach H, Falkenberg T, Fonnebo V, et al. Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol 2006;6(29).
36. Tucker JA, Roth DL. Extending the evidence hierarchy to enhance evidence-based practice for substance use disorders. Addiction 2006 Jul;101(7):918–32.
37. Atkins D, Fink K, Slutsky J. Better information for better health care: The Evidence-based Practice Center Program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005 June 21, 2005;142(12, Pt 2):1035–41.
38. Shekelle PG, Morton SC, Suttorp MJ, et al..Challenges in systematic reviews of complementary and alternative medicine topics. Ann Intern Med 2005 June 21, 2005;142(12 Pt 2):1042–7.
39. Pignone M, Saha S, Hoerger T, et al. Challenges in systematic reviews of economic analyses. Ann Intern Med 2005 June 21, 2005;142(12 Pt 2):1073–9.
40. Godwin M, Ruhland L, Casson I, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003;3(28).
41. Fullerton DSP, Atherly DS. Formularies, therapeutics, and outcomes: new opportunities. Med Care 2004 Apr;42(4 Suppl):III39–44.
42. Glasgow RE, Magid DJ, Beck A, et al. Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care 2005 Jun;43(6):551–7.
43. Kotaska A. Inappropriate use of randomised trials to evaluate complex phenomena: case study of vaginal breech delivery [review]. BMJ 2004 Oct 30;329(7473):1039–42.
44. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003 Sep 24;290(12):1624–32.
45. Medical Research Council. A framework for development and evaluation of RCTs for complex interventions to improve health. London, England: Medical Research Council; 2000
46. McAlister FA, Straus SE, Sackett DL. 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 1999 Nov 13;354(9191):1721–4.
47. Mosteller F. The promise of risk-based allocation trials in assessing new treatments [editorial]. Am J Public Health 1996 May;86(5):622–3.

Chapter 2. Identifying, Selecting, and Refining Topics

Evelyn P. Whitlock, Sarah A. Lopez, Stephanie Chang, Mark Helfand, Michelle Eder, Nicole Floyd

Key Points

The Agency for Healthcare Research and Quality's Effective Health Care (EHC) Program seeks to:

- Align its research topic selection with the overall goals of the program.
- Impartially and consistently apply predefined criteria to potential topics.
- Involve stakeholders to identify high-priority topics.
- Be transparent and accountable.
- Continually evaluate and improve processes.

A topic prioritization group representing stakeholder and scientific perspectives evaluates topic nominations for:

- Appropriateness (fit within the EHC Program).
- Importance.
- Potential for duplication of existing research.
- Feasibility (adequate type and volume of research for a new comparative effectiveness systematic review).
- Potential value and impact of a comparative effectiveness systematic review.

As the EHC Program develops, ongoing challenges include:

- Ensuring the program addresses truly unmet needs for synthesized research, since national and international efforts in this arena are uncoordinated.
- Engaging a range of stakeholders in program decisions while also achieving efficiency and timeliness.

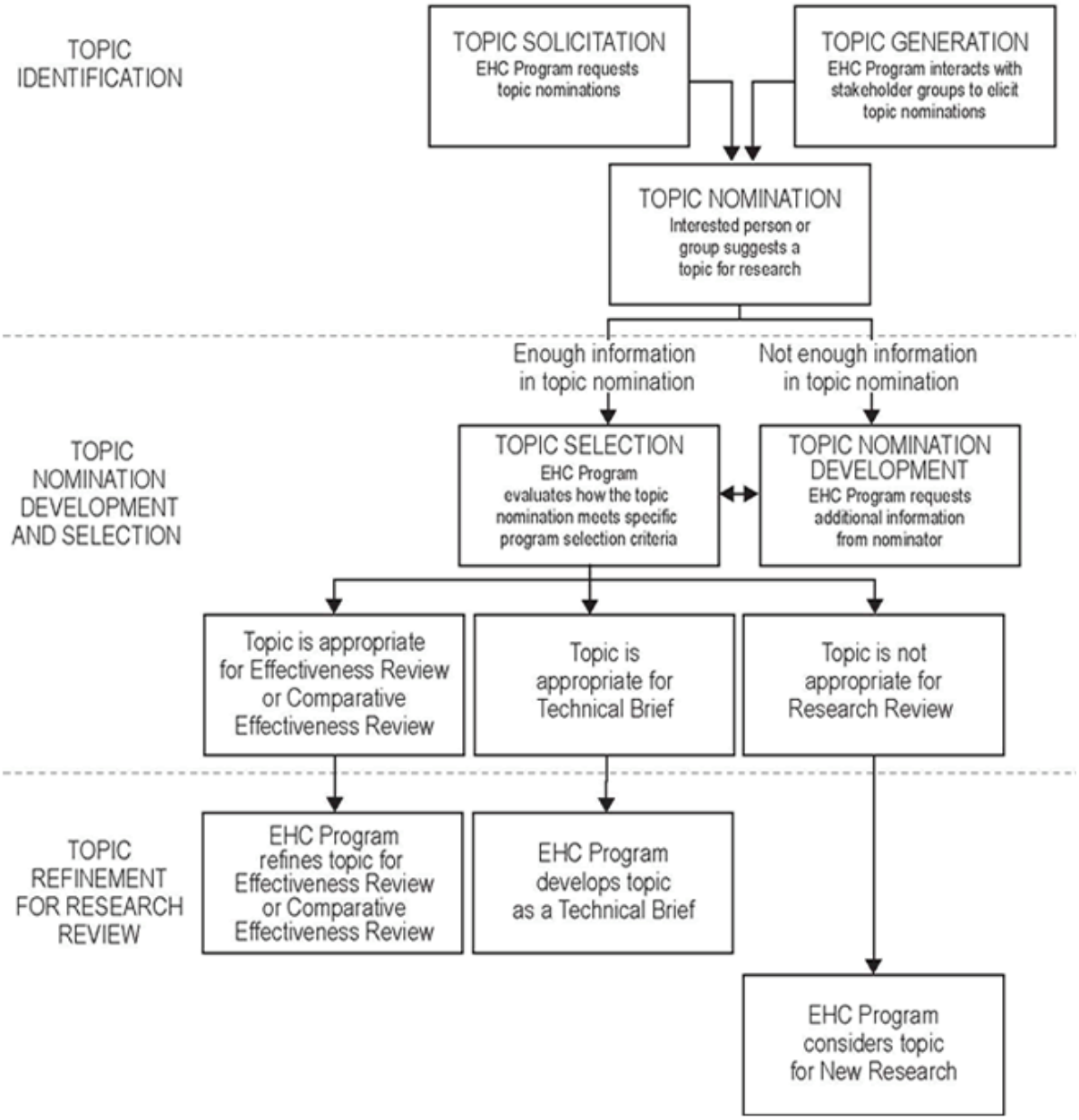
Introduction

Globally, people are struggling with the reality of limited resources to address the breadth of health and health care needs. Evidence has been recognized as the "new anchor for medical decisions,"[1] and many consider systematic reviews to be the best source of information for making clinical and health policy decisions.[2] These research products rigorously summarize existing research studies so that health and health care decisions by practitioners, policymakers, and patients are more evidence based. Yet, dollars for research—whether for systematic reviews, trials, or observational studies—are constrained, and are likely to be constrained in the future. Effective prioritization is clearly necessary in order to identify the most important topics for synthesized research investment that may help the U.S. health care system realize powerful and meaningful improvements in health status.

This paper discusses the identification, selection, and refinement of topics for comparative effectiveness systematic reviews within the Effective Health Care (EHC) Program of AHRQ, which has been described in more detail elsewhere.[3] In 2003, the U.S. Congress authorized AHRQ's Effective Health Care Program to conduct and support research on the

outcomes, comparative clinical effectiveness, and appropriateness of pharmaceuticals, devices, and health care services. This program utilizes the AHRQ Evidence-based Practice Center (EPC) Program, with 14 designated centers throughout North America that conduct comparative effectiveness systematic reviews, among other research products of the program. AHRQ has designated a Scientific Resource Center (SRC), currently housed at the Oregon EPC, to support the EHC Program as a whole. The SRC has specific responsibilities, including assisting AHRQ with all aspects of research topic development (Figure 1), providing scientific and technical support for systematic reviews and outcomes research, and collaborating with EHC stakeholder and program partners.

Figure 1. Effective Health Care (EHC) Program lifecycle of a topic nomination for research



It is not a simple process to select and develop good topics for research. Researchers' success depends in large part on their ability to identify meaningful questions, while funding

agencies continually seek to maximize the return on their investment by funding research on important, answerable questions relevant to significant portions of priority populations. Some have criticized how well funders have actually achieved these results.[4] However, there is little guidance for successfully developing a research program that generates the type of evidence necessary to improve the public's health.

Guiding Principles for Identifying and Selecting Topics

In order to derive guiding principles for selecting important comparative effectiveness systematic review topics, we considered what others have done when trying to select priority topics for any health care-related activity. Over the last 18 years, the Institute of Medicine (IOM) and selected others have explored priority-setting models and approaches.[5-10] Across a diverse set of international health- and health-care-related activities—including the development of guidelines by professional societies; clinical service and quality improvement priorities within health care organizations; and national health service guidance for health technologies, clinical practice, and public health—experts have tried to define clear-cut processes and criteria.[9,11-13] Although the majority of this existing work has not focused on specific priority setting for comparative effectiveness systematic reviews, the lessons learned from this process are relevant. These experts have found there is no obviously superior approach to setting priorities and little objective analysis to compare the relative strengths and shortcomings of various approaches.[10,14]

However, across these activities, the EHC Program has found five consistent themes for selecting the highest priority topics (Table 1). The first of these is to clearly identify the overall goals/strategic purpose of the activity in order to *align the goals for priority setting within the strategic purpose of the sponsoring program*. In the instance of the EHC Program, since no single entity can undertake activities to address all health or health care research needs, priority-setting decisions must flow from the overall mission and strategic purposes of the program.

Table 1. Effective Health Care (EHC) Program: Principles and processes for research topic selection

Principles for priority-setting in health-related programs	Applied principles for comparative effectiveness systematic review topic selection	Guidelines and processes used during comparative effectiveness systematic review topic selection
Align priority setting with the overall strategic purpose of the program	As mandated by the U.S. Congress, the EHC Program conducts research regarding "the outcomes, comparative clinical effectiveness, and appropriateness of healthcare items and services" on topics that are of broad interest and applicability, with an emphasis on topics of special importance to Medicare, Medicaid, and the State Children's Health Insurance Program (SCHIP).[1] Recent work by the Institute of Medicine (IOM) calls on us to focus these aims further by particularly considering how well potential research topics reflect the *clinical questions of patients and clinicians*, and whether selected topics truly represent *a potentially large impact on clinical or other outcomes that matter most to patients*.[2]	Under the direction of the U.S.Secretary of Health and Human Services, priority health conditions are identified to guide the focus of research (Table 5). These health conditions are being updated throughout the life of the program. For the EHC Program, robust research topics are those that represent an important decisional dilemma for consumers or for one or more participant groups in the U.S. health care system—including patients, clinicians, health system leaders, purchasers, payers, and policymakers—and that have a strong potential for significant improvements in health outcomes and/or reductions in unnecessary health-care-related burdens or costs. In aligning the EHC process with the desired outcomes for research topic selection, the overarching goal is to create a research agenda that is clearly stakeholder driven by first engaging with and then faithfully representing stakeholder interests in the products of the EHC Program.
Apply clear and consistent criteria for prioritization of potential program activities	To be ethically justifiable, prioritized topics must be relevant to the context of the program. This relevance is supported by specific rationales for prioritization that rest on reasons (evidence and principles) that could be agreed upon by "fair-minded" people.[3] A set of specific criteria has been adopted for use in prioritizing all nominated topics for systematic review (Table 4).	A topic prioritization group composed to represent scientific, stakeholder, and programmatic perspectives reviews, reasonably considers, and recommends disposition for all research topic nominations. Topic prioritization criteria applied by this group can be loosely grouped into a hierarchy of criteria to: First, determine the appropriateness of the topic for inclusion in the EHC Program. Second, establish the overall importance of a potential topic as representing a health or health care issue that matters. Third, determine the feasibility and desirability of conducting a new evidence synthesis. Fourth, estimate the potential value by considering the probable impact on health of commissioning a new evidence synthesis.

Table 1. Effective Health Care (EHC) Program: Principles and processes for research topic selection (continued)

Principles for priority-setting in health-related programs	Applied principles for comparative effectiveness systematic review topic selection	Guidelines and processes used during comparative effectiveness systematic review topic selection
Involve stakeholders	Engaging a range of stakeholders across various sectors in the United States (Table 3) increases the likelihood of identifying ideal EHC research topics. Ideal EHC research topics are those that can clearly lead to evidence-based practice and policies that support the public's health and that help better the Nation's health care system by reflecting the important needs of stakeholders. A major source of potential topics should come through regularly engaging stakeholders as active participants to generate topics. This enhanced involvement of stakeholders and more robust incorporation of their input will make the EHC Program research more relevant, with a higher propensity for effective dissemination and uptake.	As the constituencies of the EHC Program, stakeholders are key participants throughout the process (Figure 2). An EHC Program National Stakeholder Panel has been convened that represents leaders in various health and health-care-related sectors of the United States. A variety of means have been developed to engage outside experts and program partners at key points throughout the topic identification and development process. These include: An open forum, supplemented by ongoing regular engagement with key stakeholder groups, to generate topic nominations. Soliciting stakeholder consultation during topic refinement. Soliciting participation in the technical expert groups advising the EPCs conducting the systematic reviews in key question and research protocol refinement. Opportunities for public feedback during key question development. Stakeholder groups are also engaged in key aspects of report finalization and the creation of dissemination products, as described in future chapters.
Conduct program prioritization activities with adequate transparency to allow public accountability	As an ethical requirement, priority-setting decisions (and their rationales) must be publicly accessible. The IOM also emphasizes that topics for evidence syntheses that will underpin highly effective clinical services should be identified and prioritized using a system that aims to be "open, transparent, efficient, and timely" with sufficient input from key end users.[2]	Updates on program activities and priorities are available at http://www.effectivehealthcare.ahrq.gov/. The topic selection and refinement aspects of the EHC Program are meant to achieve a level of transparency that not only allows stakeholders to be a meaningful part of the process, but also tracks progress and decisions for specific nominations.

Table 1. Effective Health Care (EHC) Program: Principles and processes for research topic selection (continued)

Principles for priority-setting in health-related programs	Applied principles for comparative effectiveness systematic review topic selection	Guidelines and processes used during comparative effectiveness systematic review topic selection
Engage in ongoing self-evaluation/process improvement	Ethical principles require that there be an opportunity for challenge and revision in light of considerations raised by stakeholders. Similarly, some regulation of the process (voluntary or otherwise) to ensure its relevance, transparency, and responsiveness to appeals is required. The topic selection and refinement activities of the EHC Program will be continually reviewed to assess: How effectively outside experts and program partners are engaged in topic development. Whether the research products meet the needs of stakeholders. Whether the overall research portfolio represents a valuable set of critical evaluations for clinical and comparative effectiveness questions across a broad range of health and health care topics.	Processes are currently being finalized with input from the EHC Program National Stakeholder Panel.

1. 108th Congress. Medicare Prescription Drug, Improvement, and Modernization Act of 2003. Public Law 108–173. Section 1013.
2. Institute of Medicine. Knowing what works in health care: a roadmap for the nation. Washington: The National Academies Press; 2008.
3. Martin D, Singer P. A strategy to improve priority setting in health care institutions. Health Care Anal 2003;11:59–68.

The second principle is to *clearly define and apply criteria for prioritization among potential program activities*. Although a relatively consistent set of criteria has been utilized across health-related priority-setting activities in the United States, United Kingdom, and Canada (Table 2), specific criteria will vary with the overall goals and the purpose of any given activity. For example, to determine the national and regional estimates of health care utilization and expenditures, the Medical Expenditure Panel Survey (MEPS) prioritized data collected by considering the prevalence of medical conditions and also how accurately households could report on data related to these.[9] Similarly, to identify priority conditions for quality improvement research, the Veterans Health Administration's Quality Enhancement Research Initiative (QUERI) focused on prevalent diseases, but further prioritized prevalent diseases with evidence for both best practices and practice variation that could be improved to enhance quality.[9] Thus, for comparative effectiveness systematic review prioritization, additional criteria promulgated by the National Institute for Health and Clinical Excellence (NICE) have been considered when selecting topics for evidence-based guidance. These criteria have pointed out the importance of taking into account whether proposed topics are subject to influence by the program.[13]

Additional NICE criteria consider whether new evidence-based products could be produced in a timely manner and the risk of inappropriate treatment in the absence of evidence-based guidance.[13] This could also be considered as the opportunity cost associated with inaction.[5,13] The process of decisionmaking in health-related priority-setting activities is complex, is context dependent, and involves social processes; therefore, priority-setting processes should be guided by ethical principles, including careful attention to conflicts of interest.[14] A good priority-setting process that is fair and publicly accountable within a system that is capable of scrutiny, feedback, evaluation, and improvement is viewed as the best approach to gaining desirable outcomes.[14]

Table 2. Definitions of commonly used priority criteria for health-related topic selection

Criterion	Definition
Disease burden	Extent of disability, morbidity, or mortality imposed by a condition, including effects on patients, families, communities, and society overall.[1] Number of people/proportion of population affected; prevalence and burden of illness (quality-of-life years lost).[2] A condition associated with significant morbidity or mortality in the population as a whole or specific subgroups.[3]
Public or provider interest	Assessment to inform decisionmaking wanted by consumers, patients, clinicians, payers, and others.[1] Subject of interest to primary stakeholder.[2]
Controversy	Controversy or uncertainty around the topic and supporting data.[1] Potential to resolve ethical, legal, or social issues.[2]
Variation in care	Potential to reduce unexplained variations in prevention, diagnosis, or treatment; the current use is outside the parameters of clinical evidence.[1] Possibility of inappropriate variation in access or in clinical care in the absence of guidance.[3]
Cost	Economic cost associated with the condition, procedure, treatment, or technology related to the number of people needing care, unit cost of care, or indirect costs.[1] High costs of care (unit or aggregate); economic importance of technology.[2] An area of action where better evidence of cost effectiveness would be expected to lead to substantive cost efficiencies or might significantly impact on the National Health Service (for UK) or other societal resources (financial or other).[3]
Sufficient evidence	Adequate evidence in the available research literature to support an assessment.[1] Adequacy of data.[2] Substantive or developing body of research or related evidence.[3]
New evidence	New evidence with the potential to change conclusions from prior assessments.[1]
Potential impact	Potential to improve health outcomes (morbidity, mortality) and quality of life, improve decisionmaking for patient or provider.[1] No other assessment available; potential of assessment to impact health and economic outcomes of population.[2] Whether the guidance would promote the best possible improvement in public health or well-being and/or patient care. Whether the proposed guidance would address interventions or practices that could significantly improve quality of life (for patients or caregivers), reduce avoidable morbidity, reduce avoidable premature mortality, or reduce inequalities in health relative to current standard practice.[3]

1. Institute of Medicine. Knowing what works in health care: a roadmap for the nation. Washington: The National Academies Press; 2008.
2. Battista RN, Hodge MJ. Setting priorities and selecting topics for clinical practice guidelines. CMAJ 1995;153:1233–7.
3. National Institute for Health and Clinical Excellence. Guide to the topic selection process—interim process manual. London; November 15, 2006.

The third principle for priority setting addresses the need *to involve stakeholders in the identification and/or prioritization process*. Engaging stakeholders as key informants provides credibility and avoids prioritizing topics that have no relevance to real-world issues.

Organizations engaged in health-care-related priority setting indicate that stakeholders must be made familiar with and understand the criteria by which topics will be prioritized.[11] A recent report from the IOM on identifying highly effective evidence-based clinical services calls attention to the fact that different audiences have different needs from systematic reviews.[10] Health care payers may be most interested in the comparative effectiveness of a treatment or intervention. Regulatory agencies may be interested in questions of safety and effectiveness. Clinicians and patients may be particularly interested in the applicability of research to their specific populations. The priorities for research topics and the questions these topics should answer clearly vary by audience.

Fourth is the need for *transparency*. Because priority setting is actually an allocation of limited resources among many desirable but competing programs or people,[15] it is highly political and can be controversial. Some have asserted that priority setting in health care represents one of the most significant international health care policy questions of the 21st Century.[14] Battista and Hodge state that documentation of the process leading to a particular topic being selected (e.g., for a clinical practice guideline) should be explicit and made available to stakeholders.[5] The documentation should include the rationale that relates specific priority-setting decisions to priority-setting criteria, the evidence used when making these decisions, and any programmatic constraints that had a bearing on the process.[11] Transparency requires not only that documentation be kept, but also that program decisions and their rationales be actively communicated to stakeholders.

Fifth is the need for any prioritization approach to undertake *process evaluation and improvement* measures. Since priority setting at present is inherently a subjective process based on ideals (e.g., fairness) and decisions are made by considering clusters of factors rather than simple trade-offs,[14] there is a great need for ongoing process evaluation and improvement. As Battista and Hodge point out, process documentation forms the basis for process evaluation and improvement.[5]

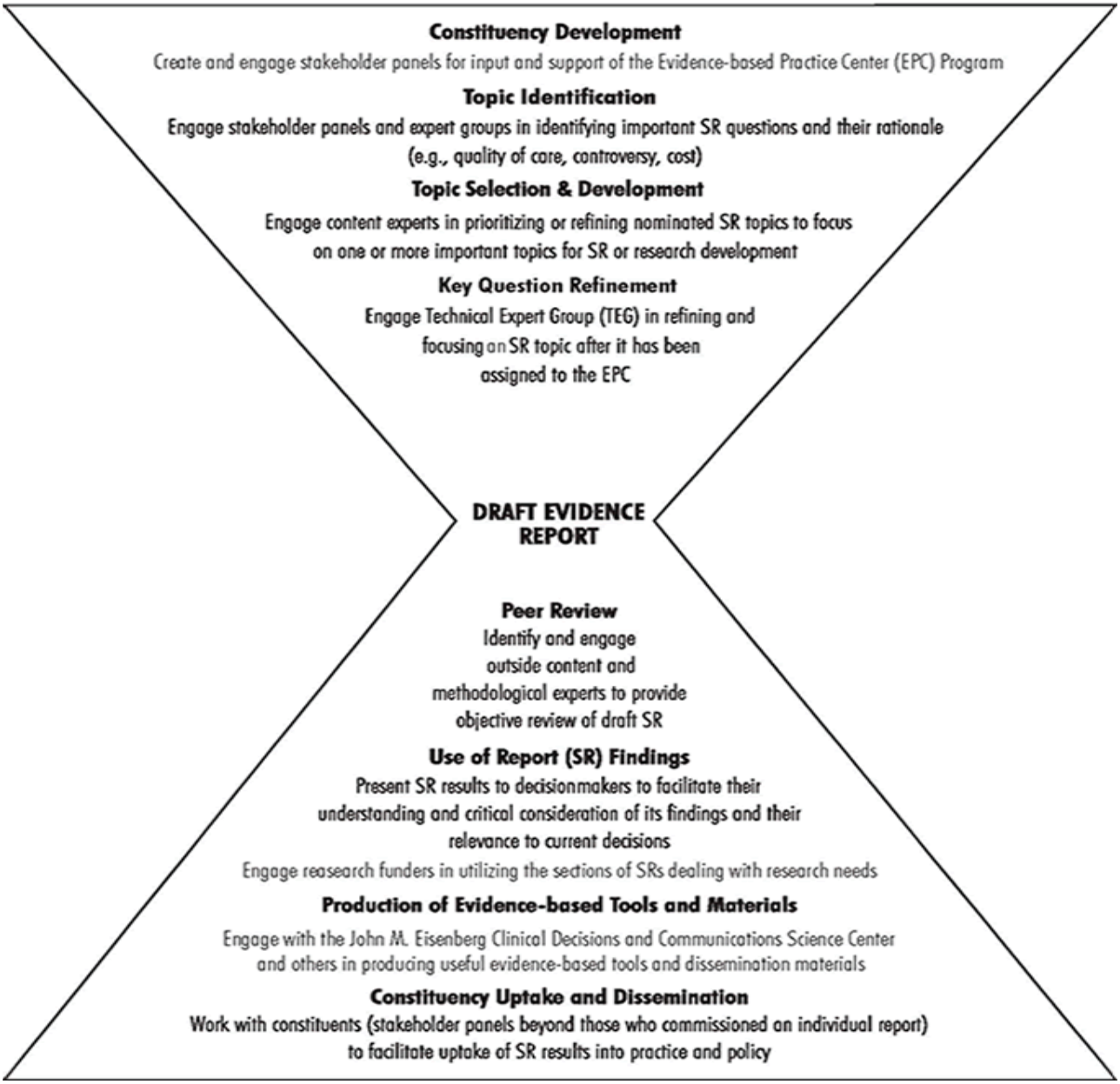
These general themes provide a good framework for selecting topics for comparative effectiveness systematic reviews. However, more specific additional criteria for clinical and comparative effectiveness research were recently articulated in a 2008 IOM report.[10] This report calls on us to consider how well potential comparative effectiveness research topics reflect the clinical questions of patients and clinicians and whether selected topics truly represent a potentially large impact on the clinical or other outcomes that matter most to patients. The IOM also emphasizes that topics for comparative effectiveness systematic reviews should be identified and prioritized using a system that aims to be “open, transparent, efficient, and timely,” with sufficient input from key end users.[10]

Processes for Identifying and Selecting Systematic Reviews

As illustrated in Figure 2, the current EHC Program processes are designed to allow the consistent, broadly focused development of a portfolio of relevant comparative effectiveness systematic reviews. These processes are focused on engaging stakeholders, particularly during topic identification, but throughout the processes of research development and dissemination within the EHC Program. This focus on stakeholders is more intense now than it was in the initial years of the EHC Program.

New and existing publicity avenues are being used to encourage nominations and engage in discussions with internal and external stakeholders interested in health care decisionmaking.

Figure 2. EHC Program Activities To Engage Stakeholders in Developing and Disseminating Systematic Reviews (SRs)



Although the EHC Program's initial mechanisms for topic identification included all of those recently cited by the IOM[10]—such as an open ongoing process for public engagement; topic solicitations; internal processes (e.g., engaging Federal agencies, such as the Centers for Medicare & Medicaid Services); and mandates—these approaches did not always produce products that met the needs of stakeholders. Nominations were often received through the Web site, but some of these nominations were insufficiently documented for consideration by the program. In addition, initial approaches did not always identify important topics that had not previously been systematically reviewed. Even when new, important systematic review topics were identified through topic nominations, these were not always developed into concise topics ideally suited for decisionmakers.

Thus, the EHC Program is currently implementing a revised system that has two important changes. First, the initial topic identification process involves more direct, focused conversations with stakeholders that represent the broad-based constituencies of the Program (Table 3). Stakeholders continue to be involved in other aspects of the program also, as described below. This direct interaction helps the EHC Program to better identify the populations, interventions, comparators, outcomes, timing, and settings of interest to the stakeholder, and to understand the current practice or health policy context underlying the need for synthesized research. A similar approach has been successfully undertaken by others.[16] Second, more explicit attempts are being made to reduce potential duplication through consulting experts and the literature to ensure that nominated topics have not already been adequately systematically reviewed. Unlike the case of primary research, where replication of existing research can be desirable, conducting duplicate systematic reviews is not clearly advantageous when existing reviews are current and of high quality.

Table 3. Stakeholder categories for the Effective Health Care Program

Clinicians
Consumers/patients, including consumer/patient organizations
Employers and business groups
Federal and State partners
Health care industry representatives
Payers, health plans, policymakers
Researchers

All fully articulated nominations are supported by issue briefs that provide data and contextual details addressing the EHC Program prioritization criteria (Table 4). Topic briefs are circulated before and presented during monthly or more frequent meetings of a topic prioritization group that represents stakeholder perspectives, scientific perspectives, and the programmatic authority vested in AHRQ. The topic prioritization group first considers objective information on the *appropriateness* of a topic and its fit within the mandate and priority conditions of the EHC Program. The priority conditions (Table 5) were determined through an open and transparent process and approved by the Secretary of Health and Human Services. The topic is then evaluated for its *importance* to the U.S. population and health care system. The available research basis on which a topic would build, including consideration of research activities already undertaken or underway by others, frames considerations of both the *feasibility* and *desirability* of a new systematic review for a nominated topic. Based on these objective data, the topic prioritization group engages in the more subjective discussions of the *potential and relative value* of commissioning a new systematic review for nominated topics. The group can request that final decisions regarding a topic nomination be deferred until further investigation is completed. Such investigations may involve outreach to nominators or other stakeholders, or further background research to determine answers to questions raised during presentation of the topic brief. At the end of the final topic prioritization discussion, the topic prioritization group can recommend that topics be sent for further refinement as a comparative effectiveness systematic review, be eliminated as outside the purview of the program, or be tabled due to other factors that affect their immediate priority. These recommendations are not binding, but are highly weighted in AHRQ's final decision as to which research topics are selected for comparative effectiveness systematic reviews.

Table 4. Selection criteria for Effective Health Care topics

Appropriateness	Represents a health care drug, intervention, device, or technology available (or soon to be available) in the United States. Relevant to enrollees in programs specified in Section 1013 of the Medicare Modernization Act of 2003 (Medicare, Medicaid, State Children's Health Insurance Program [SCHIP], other Federal health care programs). Represents one of the priority health conditions designated by the Department of Health and Human Services.
Importance	Represents a ***significant disease burden*** affecting a large proportion of the population or a priority population (e.g., children, elderly adults, low-income, rural/inner city, minorities, or other individuals with special health care or access issues). Is of ***high public interest***, affecting health care decisionmaking, outcomes, or costs for a large proportion of the U.S. population or for a priority population in particular. Was ***nominated/strongly supported by one or more stakeholder groups***. Represents ***important uncertainty*** for decisionmakers. Incorporates issues around both ***clinical benefits and potential clinical harms***. Represents ***important variation*** in clinical care or controversy in what constitutes appropriate clinical care. Represents ***high costs*** due to common use, high unit costs, or high associated costs to consumers, patients, health care systems, or payers.
Desirability of new research/ duplication	***Potential for redundancy*** (i.e., whether a proposed topic is already covered by an available or soon-to-be available high-quality systematic review by AHRQ or others)
Feasibility	***Effectively utilizes existing research and knowledge*** by considering: Adequacy (type and volume) of research for conducting a systematic review Newly available evidence (particularly for updates or new technologies)
Potential value	***Potential for significant health impact***: To improve health outcomes. To reduce significant variation in clinical practices known to be related to quality of care. To reduce unnecessary burden on those with health care problems. ***Potential for significant economic impact***: To reduce unnecessary or excessive costs. ***Potential for change***: Proposed topic exists within a clinical, consumer, or policymaking context that is amenable to evidence-based change. A product from the EHC program could be an appropriate vehicle for change. ***Potential risk from inaction:*** Unintended harms from lack of prioritization of a nominated topic *Addresses inequities, vulnerable populations* (including issues for patient subgroups) Addresses a topic that has ***clear implications for resolving important dilemmas in health and health care decisions*** made by one or more stakeholder groups.

Table 5. Priority conditions for the Effective Health Care Program

Arthritis and nontraumatic joint disorders.
Cancer.
Cardiovascular disease, including stroke and hypertension.
Dementia, including Alzheimer's Disease.
Depression and other mental health disorders.
Developmental delays, attention-deficit hyperactivity disorder, and autism.
Diabetes mellitus.
Functional limitations and disability.
Infectious diseases, including HIV/AIDS.
Obesity.
Peptic ulcer disease and dyspepsia.
Pregnancy, including preterm birth.
Pulmonary disease/asthma.
Substance abuse.

Principles and Processes for Refining Selected Topics

Once topics are selected for comparative effectiveness systematic review, they are further focused into research questions. This process is designed to ensure that the research review results in a product that meets the needs of stakeholders. Key questions should reflect the uncertainty that decisionmakers, patients, clinicians, and others may have about the topic. Key questions guide the entire systematic review process, from the formulation of comprehensive search strategies and the selection of admissible evidence to the types of data abstracted, synthesized, and reported in the final effectiveness report. Developing clear, unambiguous, and precise key questions is an early and essential step in the development of a meaningful and relevant systematic review.

For a fully formulated comparative effectiveness systematic review topic, key questions in their final form concretely specify the patient populations, interventions, comparators, outcome measures of interest, timing, and settings (PICOTS) to be addressed in the review.[17] Although the elements of the PICOTS construct are outlined in a general form at the topic identification phase, further focus and refinement of these parameters are generally required for a clear and transparent systematic review process (Tables 6 and 7). The processes to fully develop key questions are designed to carry forward the overall principles of the EHC Program of being relevant and timely, objective and scientifically rigorous, and transparent, with public participation.[3]

Table 6. PICOTS parameters for both topic nominations and key questions

PICOTS Parameters:[1]	
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 followup.
Setting	Primary, specialty, inpatient; co-interventions
Policy or Practice Context:	What are the current issues in health policy or clinical practice that define and frame the important questions to be answered?

1. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997;127:380–7.

Table 7. Issues that technical expert groups address during topic development

1. Focusing research questions for systematic review Who are the populations and clinical subgroups of interest? Why might clinical variation exist, especially if evidence-based guidelines are readily available? What specific patient characteristics may affect outcomes? Which interventions should be compared (leading to an understanding of why)? What is the potential impact of intervention on patients? What are the therapeutic aims of treatment? Which outcomes (intended and unintended effects) are relevant, including timing?
2. Clarifying clinical theories and beliefs underlying practice variation "…[E]very 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*)[1] 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 of 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.

1 Higgins JT, Green S, editors. Cochrane handbook for systematic reviews of interventions 4.2.6 [updated September 2006]. The Cochrane Library. Chichester, UK: John Wiley & Sons, Ltd; 2006.

The EHC Program's current approach to key question development is largely based on past experiences from AHRQ's Evidence-based Practice Center (EPC) Program and from other experts in systematic review. Since the inception of the EPC Program in 1997, AHRQ has emphasized the importance of input from key stakeholder informants, technical experts, and patients to elucidate the important concerns and clinical logic or reasoning underlying potential questions for systematic reviews.[18] A perfunctory set of questions or an incomplete problem formulation that outlines the general comparisons but does not specify the circumstances that are of most interest to decisionmakers clearly reduces the usability of the resulting review.[17-21] Formulating questions that address dilemmas in real-world situations, coupled with an understanding of the context around these dilemmas, prevents the production of irrelevant systematic reviews that can result from key questions that focus only on interests pertinent to researchers without much (if any) public input.[2]

The EHC Program has extended the original EPC concept of involving key stakeholder informants by developing additional mechanisms for public input. Key informants representing key stakeholder groups may be consulted as part of the topic selection process or, once selected, as part of the topic refinement process. The EHC Program also convenes a group of key stakeholder informants (including patients) and technical experts to provide additional input to

the EPC in finalizing key questions for the research review. The SRC, AHRQ, and the EPC conducting the research review work together with this group to refine the key questions for a given topic. Obtaining input from stakeholders on patients' preferences is essential to identifying pertinent clinical concerns that even expert health professionals may overlook.[22]

Incorporating a broad range of perspectives contributes to the objectivity and scientific rigor of a review by assisting EPC researchers in understanding the health care context, as well as clarifying the parameters of greatest interest when planning the research review (Table 6). These parameters are the basis for formulating good key questions and include focused determination of the most relevant populations, interventions, comparators, outcomes, timing, and setting (PICOTS).

In focusing on outcomes that matter most to patients, key questions need to identify the overarching, long-range goals of interventions. It is insufficient for key questions to focus only on what is assumed to be true or what is presently studied in the literature; they must include the populations, comparisons, and outcomes that are important to patients, providers, and policymakers using health information in their decisionmaking.

Furthermore, beliefs about the advantages or disadvantages of various alternative treatments are an important target for exploration. Many beliefs about the advantages and disadvantages of a treatment are based on direct evidence about health outcomes from long-term comparative trials. However, some 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 and stakeholders can bring attention to the issues that underlie uncertainty about the comparative effectiveness of alternative tests or therapies.

Stakeholders and other technical experts also provide important insight to direct the search for evidence that is most relevant to current practice. First, they can clarify specific populations/subpopulations or interventions of greatest clinical or policy interest. Second, interviewing those with knowledge of current clinical practices can identify areas in which studies differ in ways that may reduce their applicability.

Consistent with the principle of transparency and public participation, the EHC Program solicits public comments on proposed key questions before finalizing the scope of a new systematic review. These public comments are reviewed by AHRQ, the SRC, and the EPC, and all parties agree on changes to be made to the existing key questions to reflect this public input. Final key questions that reflect public input, as well as key stakeholder and expert input, are posted on the AHRQ EHC Web site after a review begins.

Through the processes outlined for topic identification, selection, and refinement, the EHC Program attempts to develop a considerable number of important topics for comparative effectiveness systematic reviews consistent with the principles that have been outlined above. Each topic must have appropriately focused key questions to adequately frame the systematic review while also faithfully incorporating public feedback and perspectives. The EHC processes have been developed to reduce the amount of bias that individual investigators working in isolation could potentially introduce into a topic for systematic review. However, given the complexities of the process, those involved must keep foremost in their minds the overall goal for EHC topic development: producing critically important research that positively impacts all levels of audiences' health and health care decisionmaking in order to improve the health of the public.

Challenges

Because of issues of timeliness and cost, the EHC Program cannot engage all types of stakeholders at each step for every topic. Therefore, one of the main challenges the Program faces as it moves forward is to ensure that the most important perspectives are engaged. The goal is to continue to develop a system that fairly represents the range of interests of all stakeholders across all aspects of the program (Figure 2), yet results in timely and clear reports that are useful to decisionmakers and other audiences. The process for topic identification and refinement is complicated by the large range of potential stakeholder perspectives for any given topic, by the wide-reaching clinical breadth of potential topics for the EHC Program, and by very short timeframes that are inherent in a program seeking to be publicly responsive and accountable. This tension between maintaining the relevance and rigor of research while being responsive to questions in a timely manner is an ongoing challenge.

A related challenge is gaining sufficient detail from nominators and stakeholders to allow topics to be adequately defined in order to be prioritized. The Web-based nomination system (http://effectivehealthcare.ahrq.gov) was revised recently, including definition of a minimum set of information that is necessary to understand a topic nomination sufficiently to develop it for explicit prioritization activities. This minimum set of information includes the populations, interventions, comparators, and outcomes of interest to the nominator, as well as the policy and/or clinical context. If any of these components is not clear in the nomination, the Program must have the ability to contact the nominator for more information. Since many Web-based nominations occur anonymously and since resource constraints prevent AHRQ from contacting every nominator to clarify all unclear topics, some good nominations may be missed simply because they are unclear.

Another challenging area is the relatively subjective nature of decisionmaking around topic prioritization and the sometimes highly political ramifications of these decisions. When one ventures into the realm of relative value or worth, considerations become less objective and more subject to bias. To address this challenge, the EHC Program has structured the topic prioritization process so that the same program criteria are considered for every potential topic in the same hierarchical order.

Objective evidence is considered and used as a basis for the more subjective aspects of the prioritization process. However, only process evaluation will allow determination of whether this approach helps in fairly selecting topics for research among viable and valuable candidates. Further experience in making this process and its results more transparent will undoubtedly raise unforeseen challenges as AHRQ seeks to balance the range of perspectives that are likely to be expressed, and to do so while minimizing conflicts of interest.

Prioritization of research is a necessity from a practical and a societal perception standpoint. There must be a commitment to target scarce research dollars and efforts to those areas where there will be the greatest impact and where there is a gap in needed research. There is a high level of interest in evidence-based policy and practice and the volume of uncoordinated effort internationally. Therefore, the EHC Program is working to more closely track the systematic review and policy-related activities of other programs, Federal agencies, and researchers. Enhanced coordination with others involved in setting topic priorities or in conducting analogous research is intended to reduce the opportunities for duplication. Such efforts would be greatly assisted by international registries of planned, in process, and completed comparative effectiveness and other systematic reviews.

Setting research priorities is still not a precise science. However, attempting to standardize and evaluate a structured process of setting research priorities for comparative effectiveness systematic reviews will further the goal of linking research to the actual needs of health care decisionmakers. It is necessary to find innovative and effective ways to increase the participation of health care decisionmakers in priority setting and the research process in order to bring a real-world perspective and findings that are increasingly relevant to the needs of decisionmakers.

Author Affiliations

Oregon Evidence-based Practice Center, Portland, OR, (EPW, MH, ME, NF). Kaiser Permanente Center for Health Research, Portland, OR, (EPW, ME). Oregon Health & Science University, Portland, OR, (SAL, NF). Agency for Healthcare Research and Quality, Rockville, MD, (SC). Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, (MH). Hospital and Specialty Medicine, Veterans Affairs Medical Center, Portland, OR, (MH).

This report has also been published in edited form: Whitlock EP, Lopez SA, Chang S, et al. AHRQ Series Paper 3: Identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2010;63:491–501.

References

1. Eddy DM. Evidence-based medicine: a unified approach. Health Aff (Millwood) 2005;24:9–17.
2. Laupacis A, Straus S. Systematic reviews: time to address clinical and policy relevance as well as methodological rigor. Ann Intern Med 2007;147:273–274.
3. Slutsky J, Atkins D, Chang S, et al. Comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2008 Sep 30. [Epub ahead of print]
4. Gross CP, Anderson GF, Powe NR. The relation between funding by the National Institutes of Health and the burden of disease. N Engl J Med 1999;340:1881–7.
5. Battista RN, Hodge MJ. Setting priorities and selecting topics for clinical practice guidelines. CMAJ 1995;153:1233–7.
6. Institute of Medicine. National priorities for the assessment of clinical conditions and medical technologies: report of a pilot study. Washington: The National Academy Press; 1990.
7. Institute of Medicine. Setting priorities for health technology assessment: a model process. Washington: The National Academy Press; 1992.
8. Institute of Medicine. Setting priorities for clinical practice guidelines. Washington: The National Academy Press; 1995.
9. Institute of Medicine. Priority areas for national action: transforming health care quality. Washington: The National Academy Press; 2003.
10. Institute of Medicine. Knowing what works in health care: a roadmap for the nation. Washington: The National Academies Press; 2008.
11. Gibson JL, Martin DK, Singer PA. Setting priorities in health care organizations: criteria, processes, and parameters of success. BMC Health Serv Res 2004;4:25.
12. Oxman AD, Schunemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 2. Priority setting. Health Res Policy Syst 2006;4:14.
13. National Institute for Health and Clinical Excellence. Guide to the topic selection process—interim process manual. London; November 15, 2006.
14. Martin D, Singer P. A strategy to improve priority setting in health care institutions. Health Care Anal 2003;11:59-68.

15. McKneally MF, Dickens BM, Meslin EM, et al. Bioethics for clinicians: 13. Resource allocation. CMAJ 1997;157:163–7.
16. Drug Effectiveness Review Project. Process. Available at: www.ohsu.edu/drugeffectiveness/. Accessed September 4, 2007.
17. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997;127:380–7.
18. Woolf SH, DiGuiseppi CG, Atkins D, et al. Developing evidence-based clinical practice guidelines: lessons learned by the US Preventive Services Task Force. Annu Rev Public Health 1996;17:511–38.
19. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005;142:1035–41.
20. Bravata DM, McDonald KM, Shojania KG, et al. Challenges in systematic reviews: synthesis of topics related to the delivery, organization, and financing of health care. Ann Intern Med 2005;142:1056–65.
21. Matchar DB, Westermann-Clark EV, McCrory DC, et al. Dissemination of Evidence-based Practice Center reports. Ann Intern Med 2005;142:1120–5.
22. Santaguida PL, Helfand M, Raina P. Challenges in systematic reviews that evaluate drug efficacy or effectiveness. Ann Intern Med 2005;142:1066–72.

Chapter 3. Finding Evidence for Comparing Medical Interventions

Rose Relevo, Howard Balshem

Key Points

- A librarian or other expert searcher should be involved in the development of the search.
- Sources of grey literature including regulatory data, clinical trial registries and conference abstracts should be searched in addition to bibliographic databases.
- Requests should be made to industry to request additional sources of unpublished data.
- For the main published literature search, more than one bibliographic database needs to be searched.
- Searches should be carefully documented and fully reported.

Introduction

While, this article both describes and advises on the process of literature searching in support of comparative effectiveness reviews (CERs) for the Effective Health Care program, it does not address searching for previously published systematic reviews, which is discussed in other articles in this series.[1,2]

Searches to support systematic reviews often require judgment calls about where to search, how to balance recall and precision, and when the point of diminishing returns has been reached. Searchers with more experience with complex search strategies are better equipped to make these decisions.[3] A number of reviews of the quality of systematic reviews suggest that those reviews that employed a librarian or other professional searcher had better reporting of and more complex search strategies.[4-6]

Table 1 describes the various search activities discussed in this paper and identifies who is responsible for performing each of these tasks. As is evident from the table, the EPC conducting the review is responsible for most of these activities. Because the EPC is involved in the development of the Key Questions, is familiar with the literature, and consults with experts regarding studies relevant to the topic, the EPC is in the best position to develop the required search strategies. However, some aspects of the search strategy benefit from centralization. Because grey literature searches (defined below) are by their nature highly variable, centralizing the grey literature search provides consistency across reports that would otherwise be difficult to attain. Similarly, centralizing the request to drug and device manufacturers for data on their products—what we call the Scientific Information Packet (SIP)—ensures that all requests to industry are conducted in the same manner; this also minimizes or eliminates contact between manufacturers and the EPC involved in writing the report.

Table 1. Centralized and disseminated tasks in the AHRQ Effective Health Care Program

Activity	Sources	Who does it
Key Questions and Analytic Framework	n/a	Evidence-Based Practice Center
Grey Literature Search	Clinical Trial Registries Regulatory Information Conference Proceedings	Evidence-Based Practice Center
Scientific Information Packets	Manufacturers of products under review	Scientific Resource Center
Main Literature Search	MEDLINE (plus in-process and other un-indexed citations) Cochrane Central Register of Controlled Trials	Evidence-Based Practice Center
Specialized Database Search	Variable (see Appendix B)	Evidence-Based Practice Center
Forward Citation Search	Scopus Web of Science Google Scholar	Evidence-Based Practice Center
Backwards Citations (Reading References)	Results of Main Literature Search	Evidence-Based Practice Center
Hand Search	Targeted Journals	Evidence-Based Practice Center
Corresponding with Researchers	Publication Authors	Evidence-Based Practice Center

Regulatory and Clinical Trials Searching

In addition to searching for studies that have been formally published (as described below), a comprehensive search will include a search of the grey literature.[7,8] Grey literature is defined as, "that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers".[9] Grey literature can include abstracts presented at conferences, unpublished trial data, government documents, or manufacturer information. Grey literature is, by definition, not systematically identified, stored, or indexed and therefore it can be difficult to locate.

The primary goal of the grey literature search is to identify and overcome publication and reporting bias.[10,11] Published literature does not always accurately represent trial results. Often, only articles with positive results are published, while those with "null" or negative results are not. And, even when studies are published, reporting can be biased in many other ways. Systematic reviews and meta-analysis based solely on published literature that report positive results will exaggerate any estimate of effectiveness. McAuley et al.[12] has shown an exaggerated estimate of 12 percent when grey literature is excluded, and Hopewell et al.[13] found a 9 percent exaggeration.

The usefulness of the grey literature naturally varies by topic, but it is particularly helpful in areas where there is little published evidence, where the field or intervention is new or changing,[14] when the topic is interdisciplinary,[15] and with alternative medicine.[16,17]

Despite these reasons to include grey literature, there are also potential problems. From a practical standpoint, grey literature is the least efficient body to search[18] and may not turn up more evidence to evaluate. Even if grey literature is located it may be of low quality or may not contain usable data.[19] Often unpublished studies are (or at least are perceived to be) of lower quality,[17,20] although there is limited evidence to support this.[13]

Because we have found them to be the most useful for identifying primary documents to compare with published results, the SRC routinely searches the following three types of grey literature for all CERs: regulatory data, clinical trial registries, and conference papers and abstracts.

Regulatory Data

The approval process for new drugs and devices involves submission to the Food and Drug Administration (FDA) of data that may not be published elsewhere. These approval documents—which can be found at Drugs@FDA.gov—may help identify publication bias even when complete methodological details of unpublished trials are not available.[21,22] This information is not available prior to a drug's approval and may be redacted. When they are available, reviewers can compare results of published and unpublished trials, identify inconsistencies, and often find additional data. In one meta-analysis, investigators found that published trials reported larger estimates for the efficacy of quinine than did FDA documents.[23] Similar discrepancies have been found by Turner[24] for the efficacy of antidepressants.

The SRC identifies for potential inclusion, all available medical and statistical reviews for all drugs under consideration, regardless of indication. This is partly because it is difficult to distinguish specific indications in the database, but also because the actual clinical data within the reviews may cover more than one indication and harms data are of importance regardless of indication. In addition to searching for regulatory documents from the FDA, the SRC also searches the Health Canada Drug Products Database[25] and the European Medicines Agency's European Public Assessment Reports.[26]

Trial Registries

Online trial registries such as ClinicalTrials.gov may include results of completed but unpublished clinical trials. In a prospective study of two systematic reviews, Savoie[27] found trial registries to be useful in identifying studies eligible for inclusion in systematic reviews; registries were more sensitive sources than were scanning references, hand searching, and personal communication. Trial registries can be helpful in identifying otherwise unreachable trials and in providing additional details of trials that have been published. Mathieu has found that elective outcome reporting is prevalent when trial registry information is compared with published results.[28] Even without results, knowledge that the trial exists can be helpful for reviewers because the principle investigator can be contacted for more information.[13] The FDA Amendments Act of 2007 mandates the expansion of ClinicalTrials.gov to include results of completed trials of approved drugs and devices. The results database now contains 2,279 entries, 1,958 of them from industry.[29] Although ClinicalTrials.gov contains trials completed and ongoing, we search only for completed trials, as those are the only trials that would potentially have data for inclusion in a systematic review. In addition to ClinicalTrials.gov, we routinely search the following trial registries, Current Controlled Trials,[30] Clinical Study Results,[31] and WHO International Clinical Trials Registry Platform.[32]

Abstracts and Conference Proceedings

Finally, abstracts and conference proceedings should be searched because those results often never end up as full publications,[33,34] or more formally published results often differ from the preliminary data presented in abstracts.[19,34] The SRC searches general databases of conference proceedings routinely and may search specific meetings as suggested by EPCs and key informants.

Scientific Information Packets: Requests to Industry

When interventions identified in key questions involve drugs or devices (or other products for which a manufacturer can be identified), it is important to supplement the literature search with a request to the manufacturer for a SIP. The SIP includes information about products available from the product label as well as information about published and unpublished trials or studies about the product. Requests for SIPs should not be confused with specific request to authors of publications about clarifications of data or to request additional information. These are ad hoc scientist-to-scientist communications and represent a different activity than the systematic request for SIPs from industry.

SIPs are important for two reasons. One is to overcome publication bias by identifying trials that remain unpublished. Manufacturers are not required to report results of studies of products marketed before 2008 to ClinicalTrials.gov, and so information on these studies may not be found when searching this data source. SIPs may also inform researchers about soon-to-be-published studies so that they can be included in the review without waiting for formal publication. A second reason for requesting SIPs is that they provide an explicit and transparent opportunity for drug and device manufactures to be actively involved in the CER and to provide data the manufacturer believes is important to the review of the topic. As noted above, to ensure consistency in the way SIPs are requested and to ensure transparency by eliminating contact between the EPC conducting the review and the manufacturers of products being reviewed, the Scientific Resource Center for the AHRQ Effective Health Care Program requests SIPs from manufacturers on behalf of the EPCs for all CERs and technical briefs.

Developing the Published Literature Search

The published literature search for a CER must begin with the concepts identified in the analytic framework and key questions that define the scale and scope of a project. The development of the key questions, the scope of the review, and the analytic framework is a formalized process undertaken by the systematic review team at an EPC.[2] Librarian involvement in the initial stages of the process, including reading the background materials that are prepared as the topic is developed, is an essential first step to understanding the key questions and crafting a pilot search. The searcher responsible for the main literature search is a member of the research team at the EPC performing the search. The analytic framework developed in the scoping explicitly describes both the relevant clinical concepts, as well as the logic underlying the mechanisms by which an intervention may improve health outcomes. Searchers should utilize the analytic framework to build queries for specific elements of the framework.

One thing to keep in mind while developing the search for a CER is that the retrieved results will be reviewed by hand with explicit inclusion and exclusion criteria dictated by the key questions and scope of the report. We recommend that the search be developed in tandem with these criteria.[10] Many aspects of the key question may not be adequately addressed in the search because index terms for the relevant concepts are poor or nonexistent.[35] While developing the search, if there are concepts that are difficult to articulate using search criteria alone, be sure to specify that these aspects need to be addressed in the inclusion and exclusion criteria.

The results of the pilot search can be used to help resolve questions about the boundaries of the key questions. Checking the indexing of known relevant articles provided by experts or found via reference lists can suggest additional terms and concepts that can be added to the strategy to improve its effectiveness.[35,36]

In the development of the main bibliographic search, we recommend the use of any validated hedges (filters) that exist for any of the concepts.[37-39] Hedges are predefined search strategies designed to retrieve citations based on criteria other than the subject of the article, such as study methodology or to identify papers dealing with harms.[39] Using hedges will save the work of developing the search from scratch and add a level of consistency to the Effective Health Care program's CERs. One set of hedges are the clinical queries that were developed by Haynes et al. for MEDLINE.[40] Additional filters are available from the Cochrane Collaboration,[41] the Scottish Intercollegiate Guidelines Network,[42] and the InterTASC Information Specialists' Sub-Group.[43] The Canadian Agency for Drugs & Technology in Health (CADTH) has developed a pragmatic critical appraisal tool for search filters to assist expert searchers working on systematic review teams to judge the methodological quality of a search filter.[44] For a comparison of filters designed to retrieve randomized controlled trials, see McKibbon et al.[39]

Additionally be sure to use advanced searching techniques as described in Sampson et al.'s 2008 Peer Review of Electronic Search Strategies.[45] This is a tool developed for peer review of expert searches that can also be useful as a check of the search strategy. Items to consider are:

- Spelling errors
- Line errors—when searches are combined using line numbers, be sure the numbers refer to the searches intended
- Boolean operators used appropriately
- Search strategy adapted as needed for multiple databases
- All appropriate subject headings are used, appropriate use of explosion
- Appropriate use of subheadings and floating subheadings
- Use of natural language terms in addition to controlled vocabulary terms
- Truncation and spelling variation as appropriate
- Appropriate use of limits such as language, years, etc.
- Field searching, publication type, author, etc.

Although many of the items on the list are self-explanatory, some need further clarification. Use of both natural language and indexing terms is essential for a comprehensive search.[37,46] Indexing is an important tool, but it often fails for any of the following reasons: lag time of indexing, inappropriate indexing, and lack of appropriate indexing terms or changes in indexing terms over time. Using only controlled vocabulary will miss any in-process citations in MEDLINE. As these represent the most recently published articles it is important to include natural language searching to retrieve them. When using natural language terms be sure to check for spelling errors, use truncation, and be aware of spelling variants, such as: anaemia, oesophagus, paralyse, etc.

Although the use of limits such as date ranges or age ranges may help improve the efficiency of the search, we don't recommend the use of the English language limit. Although the resources available to read or translate non-English language full text articles will vary, English language abstracts are usually available for reviewers to make an initial assessment of the study. Routinely limiting searches to English risks producing biased results.[47]

We recommend the use of a bibliographic management software package such as EndNote or RefWorks to keep track of the results.[10] We have no recommendation on specific software, however, Hernandez et al.[48] describes many currently available products. While many of these products have features that allow searches to be performed in databases such as

MEDLINE from within the software itself, we do not recommend the use of these features as they do not allow the complex searches needed for CERs.[49]

Strategies for Finding Observational Studies

CERs emphasize the use of randomized controlled trials when they are available, as this study design is least susceptible to bias and can produce high quality evidence. However, CERs include a broad range of types of evidence to confirm pertinent findings of trials and to assess gaps in the trial evidence.[50] A common use of observational studies is to compare results of trials with results in broader populations or in everyday practice.[51]

Searches for observational studies should always be included in reviews when harms and adverse effects are studied, or if the topic itself is unlikely to have been studied with randomized controlled trials.[52] For the most part, the decision on how to include observational studies will be made as the topic is being developed and is driven by the formulation of key questions and inclusion and exclusion criteria.[53] Unfortunately there is little empirical evidence on how best to approach a systematic search for observational studies.[54-56] In the absence of evidence the following is advice based on the consensus of the Cochrane Adverse Effect Methods Group[57] and other experts.[58,59]

Adverse Effects/Harms

A search for adverse effects should be more inclusive than a search for effectiveness.[53] While a search for studies about effectiveness would include only studies of the indication of interest, harms data should not be limited in this way; data about harms is of interest regardless of indication. The targeted search for adverse effects is best accomplished by combining the intervention search with terms to identify harms without limiting to any particular study type.[51,54]

Golder et al.[60] describes a number of approaches to search strategies for harms in both EMBASE and MEDLINE. In general, remember to use textwords, MeSH headings, as well as floating subheadings to identify adverse effects.[51] Because most hedges for adverse effects were designed within the context of a specific report, they may need to be adapted for new topics. For example a term such as “adverse drug reaction” would not be appropriate for nondrug interventions. Appendix A contains specific examples of these techniques and hedges.

Observational Studies in Other Situations

It can be challenging to search for observational studies because there are many designs and vocabulary is not used consistently.[56] Furlan et al.,[61] Fraser et al.,[58] and the SIGN group[62] have all explored hedges for retrieving observational studies. While they have not been validated outside of the reviews they were designed for, they offer a starting point for developing a strategy suited to the topic of the review and are described in detail in Appendix A.

While it is currently difficult to construct searches for observational studies, in the future, improved reporting and improved indexing may make it possible to develop standardized generic hedges that would be appropriate for systematic reviews. The STROBE statement[59,63] gives specific advice for the reporting of observational studies, which is a necessary first step to more accurate indexing and retrieval of observational studies.

Specialized Database Searching

While the Cochrane Central Register of Controlled Trials and MEDLINE are necessary for a thorough search of the literature, they are hardly sufficient.[64] Many topics of interest to the Effective Health Care program are interdisciplinary in nature and are concerned with more than strictly biomedical sciences. It is common, for example, to search databases such as CINAHL or PsycINFO for topics related to nursing or mental health, respectively. Failure to search multiple databases risks biasing the CER to the perspective of a single discipline and, because there is often little overlap between different databases,[46,65,66] has a high risk of missing studies that would affect the outcome of a systematic review. Sampson et al.[67] investigated the effect of such failure on meta-analyses and found that the intervention effect was increased by an average of 6 percent when only those studies found using MEDLINE were used.

When performing additional database searches, adapt search terms for each database. While keeping the conceptual structure of the original search, review the controlled vocabulary headings for each database to identify appropriate terms. Often headings that have similar scopes or definitions may vary slightly in the terminology used or differ in granularity from one database to another. Finally, keep in mind that search syntax will be different with every database, so be sure to review each database's unique syntax before performing the search.[68] Many of the more specialized databases do not have the advanced search interfaces needed to conduct complex searches, thus the searches need to be simplified. The loss in precision from the simplified search is often made up for by the fact that the databases contain a smaller number of citations, so the absolute number of citations needed to be screened—even with a simplified search—is often small.

Finally, it is always helpful to ask key informants if they know of any databases specific to the topic of interest. Consult Appendix B for a listing of possible databases of interest.

Using Key Articles

Consultation with experts will identify key articles, and these can be an important resource. If these key articles were not identified in the initial search, investigate why. By looking at the indexing terms applied to key articles, additional search terms can be identified.[35,36] Additionally, citation tracking—looking at both forward and backward citations of these key articles—can be invaluable for identifying studies.

Citation Tracking—Forward Citations

Citation tracking is an important way to identify articles because it relies on the author's choice to cite an article rather than keywords or indexing.[69] Therefore, citation tracking often identifies unique and highly relevant items. It can also be an efficient way of locating subsequent and tertiary articles stemming from a landmark trial, as these studies will all cite the original trial.

The Web of Science (which includes the Science Citation Index) is the original citation tracking service. In recent years, a number of other citation tracking databases have become available, including Google Scholar,[70] Scopus,[71] PubFocus,[72] and PubReMiner.[73] In addition, many publishers offer citation tracking within the pools of journals they publish.

While all citation tracking databases reveal who cited what, there is considerable variability in their coverage and search interfaces. Databases differ both in the number of

journals included as well as the number of years that are tracked, with Web of Science covering more years than the others.[74]

Recent comparisons of Scopus, Web of Science, and Google Scholar found that there were unique items located with each source[75,76] and that the amount of overlap varied considerably depending on the topic of interest.[74,77] Because the variation between databases is sensitive to the topic being researched, it is difficult to determine beforehand which database would be most fruitful based on content coverage alone. The decision of what database to use for citation tracking will likely be driven by more pragmatic differences between databases such as cost, availability, and search interfaces.

Web of Science and Scopus are both subscription-based services. If access is available to either of these databases, we recommend their use as they have the most developed search and export interfaces. Free citation tracking databases include: PubReMiner, PubFocus, and Google Scholar. Of these, we recommend Google Scholar for its broader coverage and superior interface. Google Scholar offers the ability to download citations into bibliographic management software as well as to link through to full-text with Google Scholar's "Scholar Preferences" settings.

Although many publishers offer citation tracking within the set of journals that they publish, we do not recommend their use because the citations are limited to results from that single publisher. Similarly, we do not recommend the "find citing articles" feature of OVID Medline, as that is restricted to journals available from Journals@OVID and does not represent all forward citations.

Reading References—Backward Citations

In addition to finding what articles have cited key studies, articles the key study has cited are a valuable resource. Sources of grey literature such as conference proceedings or poorly indexed journals relevant to the key questions are often discovered in this manner.

Reading the references of key articles is standard practice for systematic reviews[78,79] although this practice has not been systematically evaluated for effectiveness.[80] This step is often performed by the researchers tasked with reading the full text of studies and abstracting data. Since these people are often not the same people doing the literature searching, it is important to make sure that they communicate with each other during this process so that insights are not lost. We recommend that any articles that are identified through the reading of references be reviewed by the librarian conducting the search to examine why the original search strategy did not identify the article in question.

Often key articles are previous systematic reviews. The decision on when and how to use an existing review's search strategy and references is part of a larger question on how to utilize existing systematic reviews;[1] searchers should work closely with the review team to determine how to approach the use of previously published systematic reviews.

Related Articles Algorithms

Another way to use key articles is as a starting point for "related article" algorithms. Many databases offer a link to "related articles."[37] These links can be helpful in the preliminary, exploratory, and scoping stages of a search. However, we do not recommend them for the formal part of the search for a CER; it is difficult to be systematic about and report on these types of searches, and generally, they are impossible to reproduce.

Hand Searching Journals

Not all journals of interest will be indexed by the databases searched; often, abstracts, supplements, and conference proceedings are not indexed, even if the rest of the content of a journal is. Because many studies first appear (or only appear) in these nonindexed portions of a journal, hand searching journals can be an effective method for identifying trials.

We recommend that journals be hand searched if they are highly relevant to the topic of the report, but are not fully indexed[35,81,82] or not indexed at all by MEDLINE.[83] It is often the case that articles were missed by the initial search strategy because the journal the article is published in is poorly indexed. Asking key informants about specific journals or conferences related to the topic is another way to identify candidates for hand searching.[84,85]

Hand searching doesn't necessarily mean hand searching of the print journal (although that may be appropriate in some cases). Now that tables of contents and abstracts are often available electronically, hand searching can be done online by systematically reviewing the journal's content on an issue-by-issue basis. A more focused hand search may limit the number of years searched, or focus only on supplements or conference abstracts.

Corresponding With Researchers

During the course of preparing a CER it may be necessary to contact investigators and authors. Savoie[27] found that personal communication was a major source of identifying studies, especially when there are uncertainties surrounding a study's publication status. Direct contact with authors can often match these sources to full publications, confirm that there was no subsequent publication, identify unique published or soon-to-be-published sources, and clear up uncertainty surrounding duplicate publication.[86-91]

Email makes author correspondence quite easy. Gibson et al.[92] found that the response rate to email was higher than for postal mail. Aside from the usual Google search, email addresses can be identified by searching the author's institution's Web site. PubMed is also a good source of email addresses, as they are included in the author institution field shown in the abstract display.

Updating and Reporting the Search Strategy

While conducting the search be sure to take detailed notes. These will be useful for reporting as well as rerunning the search in the future. EPC program policy requires saving the main bibliographic searches to be rerun at the time the draft is sent for peer review. In addition, detailed notes about the full search strategy should be kept in order to accurately report the search strategy in the review. Transparency and reproducibility of the systematic review requires clear reporting;[93] critical appraisal is impossible if the search strategy is not thoroughly reported.[94]

Unfortunately, there is no consensus on how to report search strategies in systematic reviews. Sampson et al.[94] identified 11 instruments, either specific to search strategy reporting or more global reporting instruments that include elements for the search strategy. From these 11 instruments, they extracted the following elements:

- Database used
- Dates covered by the search
- Date search was conducted
- Statement of the search terms used

- Statement of any language restrictions
- Statement of nondatabase methods used
- Additional inclusion/exclusion criteria
- Presentation of the full electronic search strategy
- Statement of any publication status restrictions
- Platform or vendor for electronic database
- End date of the search
- List of excluded references
- Qualifications of the searcher
- Is the reported strategy repeatable?
- Number of references identified
- PRISMA-style flow diagram or other accounting for all references
- Evidence of effectiveness of the search strategy
- Statement of any methodological filters used
- Description of the sampling strategy

The PRISMA-style flow diagram refers to a chart that accounts for all citations identified from all sources as well as accounting for all citations that were later excluded and why.[95,96] See Appendix C for an annotated example.

Another element that falls outside of the basic mechanics of the search is evidence of the effectiveness of the search strategy.[94] The evidence of the effectiveness of the search strategy may be difficult to ascertain conclusively. However, reporting what techniques were used to check a strategy—such as expert review, use of previously published strategies or hedges, or testing against a group of known relevant articles (for example, from a previous review)—may be helpful.

With the lack of consensus on reporting, it is hardly surprising that current reporting of search strategies for systematic reviews is variable. In a recent review, Yoshii et al.[93] provided a good overview of studies of reporting of search strategies in systematic reviews; they also examined the reporting in Cochrane reviews. Reporting of search strategies is an area of systematic review methodology that can be improved, and the problems with poor reporting go beyond not being able to reproduce the search or build on it for updates. There is very little evidence on the effectiveness of various search strategies for CERs, and there is a need for primary research to identify the characteristics of valid searches.[94] Currently, it is difficult to do any research on this issue because reporting is so poor. Completely reported search strategies will build an evidence base from which research can be done on effective search strategies.

In the absence of reporting standards, we recommend working with the team writing the report to determine what to report in the review. Page limitations of journal publications may necessitate abbreviating the reporting in journal publications, but there is always room for complete reporting in the online appendices of the CER that are posted to the Effective Health Care Web site or included with the e-published version of the journal article.

Concluding Remarks

One of the most difficult aspects of conducting a comprehensive search is confidently knowing when to stop searching. Unfortunately, there is little guidance on how to determine that point. While Spoor et al.[97] suggests capture-mark-recapture statistical modeling to

retrospectively estimate the closeness to capturing the total body of literature, there is currently no tool that can easily be applied to searches for CERs. In the end, we rely on experienced searchers' judgments as to when the labor expended to search additional sources is likely to result in new and unique items or whether the search has reached the point of saturation. Like other decisions, such as the sensitivity of the search, the desire for comprehensiveness must be balanced with available resources.

Much of the methodology described here is not yet evidence based, but rather based on principles of expert searching and searcher experience. In order to develop more evidence-based methods we must first have an evidence base to work with. Poor reporting of search strategies in comparative effectiveness and other systematic reviews has hindered evaluations of the effectiveness of various techniques. Clear reporting of search strategies, therefore, is the first step needed to support further research on the effectiveness of various search techniques.

Within the AHRQ Effective Health Care Program, searching lacks the type of quality control that is found in many other steps in the process of conducting CERs, such as dual abstraction and internal peer review. The Scientific Resource Center, therefore, has initiated projects such as peer review of search strategies and improved structures for communication and dissemination of techniques intended to identify best practices that will help librarians share expertise across EPCs.

Author Affiliations

Scientific Resource Center, AHRQ Effective Health Care Program, Oregon Health & Science University, Portland, OR (RR, HB).

This paper has also been published in edited form: Relevo R, Balshem H. Finding evidence for comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2011 Nov;64(11):1168–1177. PMID: 21684115.

References

1. Whitlock EP, Lin JS, Chou R, et al. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008 May 20;148(10):776–82.
2. Whitlock EP, Lopez SA, Chang S, et al. Identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the Effective Health Care program. J Clin Epidemiol 2009 Jun 18.
3. Medical Library Association. Role of expert searching in health sciences libraries : Policy statement by the Medical Library Association adopted September 2003. J Med Libr Assoc 2005 Jan;93(1):42–4.
4. McGowan J, Sampson M. Systematic reviews need systematic searchers. J Med Libr Assoc 2005 Jan;93(1):74–80.
5. Golder S, Loke Y, McIntosh HM. Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. J Clin Epidemiol 2008 May;61(5):440–8.
6. Mokkink LB, Terwee CB, Stratford PW, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res 2009 Apr;18(3):313–33.
7. Alberani V, De Castro Pietrangeli P, et al. The use of grey literature in health sciences: a preliminary survey. Bull Med Libr Assoc 1990 Oct;78(4):358–63.
8. Illig J. Archiving "event knowledge": bringing "dark data" to light. J Med Libr Assoc 2008 Jul;96(3):189–91.
9. Grey Literature Network Service, editor. New frontiers in grey literature. Fourth International conference on Grey Literature; 1999 Oct 4-5;Washington, DC: GL'99 proceedings.

10. Conn VS, Valentine JC, Cooper HM, et al. Grey literature in meta-analyses. Nurs Res [Review] 2003 Jul-Aug;52(4):256–61.
11. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ [Review]. 1994 Nov 12;309(6964):1286–91.
12. McAuley L, Pham B, Tugwell P, et al. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000 Oct 7;356(9237):1228–31.
13. Hopewell S, McDonald S, Clarke Mike J, et al. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews. 2007; (2): Available at: http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/MR000010/frame.html.
14. Hartling L, McAlister FA, Rowe BH, et al. Challenges in systematic reviews of therapeutic devices and procedures. Ann Intern Med 2005 Jun 21;142(12 Pt 2):1100–11.
15. Helmer D, Savoie I, Green C, et al. Evidence-based practice: extending the search to find material for the systematic review. Bull Med Libr Assoc 2001 Oct;89(4):346–52.
16. Shekelle PG, Morton SC, Suttorp MJ, et al. Challenges in systematic reviews of complementary and alternative medicine topics. Ann Intern Med 2005 Jun 21;142(12 Pt 2):1042–47.
17. Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Asses 2003;7(1):1–76.
18. Cook AM, Finlay IG, Edwards AG, et al. Efficiency of searching the grey literature in palliative care. J Pain Symptom Manage 2001 Sep;22(3):797-801.
19. Fergusson D, Laupacis A, Salmi LR, et al. What should be included in meta-analyses? An exploration of methodological issues using the ISPOT meta-analyses. Int J Technol Assess Health Care 2000 Autumn;16(4):1109–19.
20. van Driel ML, De Sutter A, De Maeseneer J, et al. Searching for unpublished trials in Cochrane reviews may not be worth the effort. J Clin Epidemiol 2009 Aug;62(8):838–844 e3.
21. Bennett DA, Jull A. FDA: untapped source of unpublished trials. Lancet 2003;361:1402–3.
22. MacLean CH, Morton SC, Ofman JJ, et al. How useful are unpublished data from the Food and Drug Administration in meta-analysis? J Clin Epidemiol 2003 Jan;56(1):44–51.
23. Man-Son-Hing M, Wells G, Lau A. Quinine for nocturnal leg cramps: a meta-analysis including unpublished data. J Gen Intern Med 1998 Sep;13(9):600–6.
24. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008 Jan 17;358(3):252–60.
25. Health Canada Drug Products Database. Health Canada; Available at: http://webprod.hc-sc.gc.ca/dpd-bdpp/index-eng.jsp. Accessed September 21, 2010.
26. European Public Assessement Reports. European Medicines Agency; Available at: http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&murl=menus/medicines/medicines.jsp&mid=WC0b01ac058001d124. Accessed September 21, 2010.
27. Savoie I, Helmer D, Green CJ, et al. Beyond MEDLINE: reducing bias through extended systematic review search. Int J Technol Assess Health Care [Journal Article] 2003;19(1):168–78.
28. Mathieu S, Boutron I, Moher D, et al. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009 Sep 2;302(9):977–84.
29. U.S. National Institutes of Health. ClinicalTrials.gov. 2010; Available at: http://clinicaltrials.gov. Accessed September 21, 2010.
30. Current Controlled Trials. BioMed Central; Available at: http://www.controlled-trials.com/. Accessed September 21, 2010.
31. Clinical Study Results. Available at: www.clinicalstudyresults.org/home/. Accessed September 21, 2010.
32. WHO International Clinical Trials Registry Platform. World Health Organization; Available at: http://apps.who.int/trialsearch/. Accessed September 21, 2010.
33. von Elm E, Costanza MC, Walder B, et al. More insight into the fate of biomedical meeting abstracts: a systematic review. BMC Med Res Methodol 2003 Jul 10;3:12.
34. Toma M, McAlister FA, Bialy L, et al. Transition from meeting abstract to full-length journal article for randomized controlled trials. JAMA 2006 Mar 15;295(11):1281–7.

35. Matthews EJ, Edwards AG, Barker J, et al. Efficient literature searching in diffuse topics: lessons from a systematic review of research on communicating risk to patients in primary care. Health Libr Rev 1999 Jun;16(2):112–20.
36. Brettle AJ, Long AF. Comparison of bibliographic databases for information on the rehabilitation of people with severe mental illness. Bull Med Libr Assoc 2001 Oct;89(4):353–62.
37. O'Leary N, Tiernan E, Walsh D, et al. The pitfalls of a systematic MEDLINE review in palliative medicine: symptom assessment instruments. Am J Hosp Palliat Care 2007 Jun-Jul;24(3):181–4.
38. Glanville JM, Lefebvre C, Miles JN, et al. How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc 2006 Apr;94(2):130–6.
39. McKibbon KA, Wilczynski NL, Haynes RB, et al. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Info Libr J 2009 Sep;26(3):187–202.
40. Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994 Nov-Dec;1(6):447–58.
41. Cochrane Handbook for Systematic Reviews of Interventions : 6.4.11 Search filters 2008 updated September 2008; Version 5.0.1:Available at: http://www.cochrane-handbook.org. Accessed September 21, 2010.
42. Scottish Intercollegiate Guidelines Network (SIGN). Search Filters. Edingurgh 2009, updated August 3, 2009; Available at: http://www.sign.ac.uk/methodology/filters.html. Accessed September 21, 2010.
43. InterTASC Information Specialists' Sub-Group. Search Filter Resource 2009, updated July 2, 2009; Available at: www.york.ac.uk/inst/crd/intertasc/diag.htm. Accessed September 21, 2010.
44. Bak G, Mierzwinski-Urban M, Fitzsimmons H, et al. A pragmatic critical appraisal instrument for search filters: introducing the CADTH CAI. Health Info Libr J 2009 Sep;26(3):211–9.
45. Sampson M, McGowan J, Lefebvre C, et al. PRESS: Peer Review of Electronic Search Strategies. Ottowa: Canadian Agency for Drugs and Technologies in Health; 2008.
46. Conn VS, Isaramalai SA, Rath S, et al. Beyond MEDLINE for literature searches. J Nurs Scholarsh [Review] 2003;35(2):177–82.
47. Morrison A, Moulton K, Clark M, et al. English-language restriction when conducting systematic review-based meta-analyses: systematic review of published studies. Ottawa: Canadian Agency for Drugs and Technologies in Health 2009: Available at: www.mrw.interscience.wiley.com/cochrane/clcmr/articles/CMR-13119/frame.html.
48. Hernandez DA, El-Masri MM, Hernandez CA. Choosing and using citation and bibliographic database software (BDS). Diabetes Educ 2008 May-Jun;34(3):457–74.
49. Gomis M, Gall C, Brahmi FA. Web-based citation management compared to end note: options for medical sciences. Med Ref Serv Q 2008 Fall 2008;27(3):260–71.
50. White CM, Ip S, McPheeters M, et al. Using existing systematic reviews to replace de novo processes in conducting Comparative Effectiveness Reviews. Rockville, MD 2009 Available at: http://effectivehealthcare.ahrq.gov/repFiles/methodsguide/systematicreviewsreplacedenovo.pdf. Accessed September 21, 2010.
51. Loke YK, Price D, Herxheimer A, et al. Systematic reviews of adverse effects: framework for a structured approach. BMC Med Res Methodol 2007;7(32).
52. Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. Ann Intern Med [Review] 2005 Jun 21;142(12 Pt 2):1090–9.
53. Chou R, Aronson N, Atkins D, et al. Assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program. J Clin Epidemiol 2008 Sep 25.
54. Derry S, Kong Loke Y, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. BMC Med Res Methodol 2001;1(7).
55. Golder S, McIntosh HM, Duffy S, et al. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J 2006 Mar;23(1):3–12.
56. Wieland S, Dickersin K. Selective exposure reporting and Medline indexing limited the search sensitivity for observational studies of the adverse effects of oral contraceptives. J Clin Epidemiol 2005 Jun;58(6):560–7.

57. Loke YK, Price D, Herxheimer A. Appendix 6b. Including adverse effects. In: Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 425 [updated May 2005]. Chichester, UK: Cochrane Colloboration; Cochrane Adverse Effects Subgroup; 2007. pp. 194–5.
58. Fraser C, Murray A, Burr J. Identifying observational studies of surgical interventions in MEDLINE and EMBASE. BMC Med Res Methodol 2006 Aug 18;6(41).
59. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573–7.
60. Golder S, Loke Y, McIntosh HM. Room for improvement? A survey of the methods used in systematic reviews of adverse effects. BMC Med Res Methodol 2006;6(3).
61. Furlan AD, Irvin E, Bombardier C. Limited search strategies were effective in finding relevant nonrandomized studies. J Clin Epidemiol 2006 Dec;59(12):1303–11.
62. Scottish Intercollegiate Guidelines Network - Search Filters. Available at: http://www.sign.ac.uk/methodology/filters.html.
63. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. Ann Intern Med. [Web only] 2007;147:W163–W94.
64. Zheng MH, Zhang X, Ye Q, et al. Searching additional databases except PubMed are necessary for a systematic review. Stroke 2008 Aug;39(8):e139; author reply e40.
65. Suarez-Almazor ME, Belseck E, Homik J, et al. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Control Clin Trials 2000 Oct;21(5):476–87.
66. Betran AP, Say L, Gulmezoglu AM, et al. Effectiveness of different databases in identifying studies for systematic reviews: experience from the WHO systematic review of maternal morbidity and mortality. BMC Med Res Methodol 2005 Jan 28;5(1):6.
67. Sampson M, Barrowman NJ, Moher D, et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003 Oct;56(10):943–55.
68. DeLuca JB, Mullins MM, Lyles CM, et al. Developing a comprehensive search strategy for evidence based systematic reviews. Evid Based Libr Inf Pract 2008;3(1):3–32.
69. Kuper H, Nicholson A, Hemingway H. Searching for observational studies: what does citation tracking add to PubMed? A case study in depression and coronary heart disease. BMC Med Res Methodol 2006;6:4.
70. Google Scholar. Available at: http://scholar.google.com/. Accessed September 21, 2010.
71. Scopus. Elsevier; Available at: http://www.scopus.com/home.url. Accessed September 21, 2010.
72. PubFocus. Available at: http://pubfocus.com/. Accessed September 21, 2010.
73. PubMed PubReMiner. Available at: http://bioinfo.amc.uva.nl/human-genetics/pubreminer/. Accessed September 21, 2010.
74. Falagas ME, Pitsouni EI, Malietzis GA, et al. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J 2008 Feb;22(2):338–42.
75. Salisbury L. Web of Science and scopus : a comparative review of content and searching capabilities. The Charleston Advisor 2009 July;11(1):5–18.
76. Jasco P. As we may search—Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. Curr Sci 2005;89(9):1537–47.
77. Bakkalbasi N, Bauer K, Glover J, et al. Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomed Digit Libr 2006;3(7).
78. Jadad AR, McQuay HJ. Searching the literature. Be systematic in your searching [comment]. BMJ 1993 Jul 3;307(6895):66.
79. Gotzsche PC. Reference bias in reports of drug trials. Br Med J (Clin Res Ed) 1987 Sep 12;295(6599):654–6.
80. Armour T, Dingwall O, Sampson M. Contribution of checking reference lists to systematic reviews. Poster presentation at: XIII Cochrane Colloquium 2005.

81. Al Hajeri A, Al Sayyad J, Eisinga A. Handsearching the EMHJ for reports of randomized controlled trials by U.K. Cochrane Centre (Bahrain). East Mediterr Health J 2006;12 (Suppl 2):S253–S257.
82. Jadad AR, Moher D, Klassen TP. Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? Arch Pediatr Adolesc Med 1998 Aug;152(8):812–7.
83. Hopewell S, Clarke M, Lusher A, et al. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Stat Med 2002 Jun 15;21(11):1625–34.
84. Avenell A, Handoll HH, Grant AM. Lessons for search strategies from a systematic review, in The Cochrane Library, of nutritional supplementation trials in patients after hip fracture. Am J Clin Nutr 2001 Mar;73(3):505–10.
85. Armstrong R, Jackson N, Doyle J, et al. It's in your hands: the value of handsearching in conducting systematic reviews of public health interventions. J Public Health (Oxf) 2005 Dec;27(4):388–91.
86. Zarin DA, Ide NC, Tse T, et al. Issues in the registration of clinical trials. JAMA 2007 May 16;297(19):2112-2120.
87. Tramer MR, Reynolds DJ, Moore RA, et al. Impact of covert duplicate publication on meta-analysis: a case study. BMJ. 1997 Sep 13;315(7109):635–40.
88. Reveiz L, Cardona AF, Ospina EG, et al. An e-mail survey identified unpublished studies for systematic reviews. J Clin Epidemiol 2006 Jul;59(7):755–8.
89. Kelley GA, Kelley KS, Tran ZV. Retrieval of missing data for meta-analysis: a practical example. Int J Technol Assess Health Care 2004 Summer;20(3):296-299.
90. Peinemann F, McGauran N, Sauerland S, et al. Negative pressure wound therapy: potential publication bias caused by lack of access to unpublished study results data. BMC Med Res Methodol 2008;8:4.
91. Rennie D. Fair conduct and fair reporting of clinical trials. JAMA 1999 Nov 10;282(18):1766–8.
92. Gibson CA, Bailey BW, Carper MJ, et al. Author contacts for retrieval of data for a meta-analysis on exercise and diet restriction. Int J Technol Assess Health Care 2006 Spring;22(2):267–70.
93. Yoshii A, Plaut DA, McGraw KA, et al. Analysis of the reporting of search strategies in Cochrane systematic reviews. J Med Libr Assoc 2009;97(1):21–9.
94. Sampson M, McGowan J, Tetzlaff J, et al. No consensus exists on search reporting methods for systematic reviews. J Clin Epidemiol 2008 Aug;61(8):748–54.
95. Egger M, Juni P, Bartlett C, et al. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001 Apr 18;285(15):1996–9.
96. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 6(7):e1000100. July 21, 2009 [epub ahead of print].
97. Spoor P, Airey M, Bennett C, et al. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. BMJ. 1996 Aug 10;313(7053):342–3.

Chapter 3 Appendix A. Techniques for Observational Studies and/or Harms

Fraser 2006 Observational Studies – surgery			
MEDLINE (OVID)		EMBASE (OVID)	
Precision	Specificity	Precision	Specifity
Comparative studies/ Follow-up studies/ (preoperat$ or pre operat$).mp chang$.tw evaluat$.tw reviewed.tw prospective$.tw baseline.tw cohort.tw consecutive$.tw (compare$ or compara$).tw	Comparative studies/ Follow-up studies/ Time factors/ (preoperat$ or pre operat$).mp chang$.tw evaluat$.tw reviewed.tw prospective$.tw retrospective$.tw baseline.tw cohort.tw case series.tw	Controlled Study/ Treatment outcome/ Major clinical study/ (preoperat$ or pre operat$).mp chang$.tw evaluat$.tw reviewed.tw (compare$ or compara$).tw	Controlled Study/ Treatment outcome/ Major clinical study/ Clinical trial/ chang$.tw evaluat$.tw reviewed.tw baseline.tw (compare$ or compara$).tw

Furlan 2006 Observational Studies	
MEDLINE	EMBASE
Cohort studies/ comparative study/ follow-up studies/ prospective studies/ risk factors/ cohort.mp. compared.mp. groups.mp. multivariate.mp.	clinical article/ controlled study/ major clinical study/ prospective study/ cohort.mp. compared.mp. groups.mp. multivariate.mp.

Golder 2006 Adverse Effects		
search approach	MEDLINE	EMBASE
specified adverse effects	*Drug terms* AND Exp LIVER DISEASES/ci	*Drug terms* AND Exp LIVER DISEASE/si
subheadings linked to drug name	Exp *DRUG NAME*adverse events, po, to	Exp DRUG NAMEadverse events, to
floating subheadings	*Drug terms* AND (ae OR po OR to OR co OR de).fs.	*Drug terms* AND (ae OR to OR co).fs.
text word synonyms of “adverse effects” and related terms	*Drug terms* AND (safe OR safety OR side-effect$ OR undesirable effect$ OR treatment emergent OR tolerability OR toxicity OR adrs OR [adverse adj2 (effect or effects or reaction or reactions or event or events or outcome or outcomse)])	*Drug terms* AND (safe OR safety OR side-effect$ OR undesirable effect$ OR treatment emergent OR tolerability OR toxicity OR adrs OR [adverse adj2 (effect or effects or reaction or reactions or event or events or outcome or outcomse)])
indexing terms for “adverse effects”	*Drug terms* AND exp DRUG TOXICITY/	Drug terms AND (exp ADVERSE DRUG REACTION/ OR Exp Side-Effect/)

Loke 2007 – indexing terms (subheadings)	
MEDLINE	EMBASE
/adverse effects /poisoning /toxicity /chemically induced /contraindications /complications	/side effect /adverse drug reaction /drug toxicity /complication

Scottish Intercollegiate Guidelines Network (SIGN) Observational Studies		
MEDLINE	EMBASE	CINAHL
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	
10	10	
11	11	
12	12	
13	13	
	14	
	15	
	16	

References

Fraser C, Murray A, Burr J. Identifying observational studies of surgical interventions in MEDLINE and EMBASE. BMC Medical Research Methodology. 2006 Aug 18;6(41).

Furlan AD, Irvin E, Bombardier C. Limited search strategies were effective in finding relevant nonrandomized studies. Journal of Clinical Epidemiology. 2006 Dec;59(12):1303–11.

Golder S, McIntosh HM, Duffy S, Glanville J, Dissemination CfRa, Group. UCCSFD. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Information and Libraries Journal. 2006 Mar;23(1):3–12.

Loke YK, Price D, Herxheimer A, Group CAEM. Systematic reviews of adverse effects: framework for a structured approach. BMC Medical Research Methodology. 2007;7(32).

Scottish Intercollegiate Guidelines Network (SIGN). Search Filters. Edinburgh 2009; Available at: http://www.sign.ac.uk/methodology/filters.html. Accessed September 21, 2010.

Chapter 3 Appendix B. Specialized Databases

Please note that the topics listed are not the only topics indexed by that database, rather they are a subset of covered topics that are likely to be of interest to the Effective Health Care Program. References are to articles which discuss specific search strategies, present a general overview of the database, or discuss the use of these databases in systematic reviews. The URL's listed are for the database itself if it's a free resource, or a page describing the product if it's a subscription database. Please note that many of these databases are available from many vendors, and the choice of URL does not indicate a preference or endorsement of any particular vendor. If you are unsure about subscription databases, remember that free trials can often be arranged in order for you to evaluate its usefulness to your program.

Free Resources			
Database	URL	Topic Coverage	References
C2-SPECTR (Campbell Collaboration's Social, Psychological, Educational and Criminology Trials Register)	http://geb9101.gse.upenn.edu/	Trial Register for Social Sciences (similar to DARE)	Petrosio, 2000
ERIC (Education Resources Information Center)	http://www.eric.ed.gov/	Education, including the education of health care professionals as well as educational interventions for patients	Anon, 2006
IBIDS (International Bibliographic Information on Dietary Supplements)	http://ods.od.nih.gov/Health_Information/IBIDS.aspx	Dietary Supplements	Tomasulo, 2000
ICL (Index to Chiropractic Literature)	http://www.chiroindex.org/	Chiropractic	Aker, 1996
NAPS (New Abstracts and Papers in Sleep)	http://www.websciences.org/bibliosleep/naps/default.html	Sleep	
OTseeker (Occupational Therapy Systematic Evaluation of Evidence)	http://www.otseeker.com/	Occupational Therapy	Bennett, 2003 Bennett, 2006
PEDro (Physiotherapy Evidence Database)	http://www.pedro.org.au/	Physical Therapy	Sherrington, 2000 Moseley, 2002 Giglia, 2008 Fitzpatrick, 2008

PILOTS	http://www.ptsd.va.gov/ptsd_adv_search.asp	PTSD and Traumatic Stress	Banks, 1995 Kubany, 1995 Lerner, 2007
PopLine	http://www.popline.org	Population, Family Planning & Reproductive Health	Adebonojo, 1994
PubMed	http://www.ncbi.nlm.nih.gov/pubmed/	Biology and Health Sciences	
RDRB (Research and Development Resource Base)	http://www.rdrb.utoronto.ca/about.php	Medical Education	Anne, 1995
RehabData	http://www.naric.com/research/rehab/	Rehabilitation	Fitzpatrick, 2007
Social Care Online	http://www.scie-socialcareonline.org.uk/	Social Care including: Healthcare, Social Work and Mental Health	Gwynne-Smith, 2007
TOXNET	http://toxnet.nlm.nih.gov/	Toxicology Environmental Health Adverse Effects	Hochstein, 2007
TRIS (Transportation Research Information Service)	http://ntlsearch.bts.gov/tris/index.do	Transportation Research	Wang, 2001
WHO Global Health Library	http://www.who.int/ghl/medicus/en/	International biomedical topics. Global Index Medicus.	
Subscription Resources			
Database	URL	Topic Coverage	References
AgeLine	http://www.csa.com/factsheets/ageline-set-c.php	Aging, Health topics of interest to people over 50	Tomasulo, 2005
AMED (Allied and Complimentary Medicine Database)	http://www.ovid.com/site/catalog/DataBase/12.jsp	Complementary Medicine and Allied Health	Hoffecker, 2006 Pilkington, 2007
ASSIA (Applied Social Science Index and Abstracts)	http://www.csa.com/factsheets/assia-set-c.php	Applied Social Sciences including: Anxiety disorders, Geriatrics, Health, Nursing, Social Work and Substance abuse	LaGuardia, 2002
BNI (British Nursing Index)	http://www.bniplus.co.uk/about_bni.html	Nursing and Midwifery	Flemming 2007
ChildData	http://www.childdata.org.uk/	Child related topics including child health	

CINAHL (Cumulative Index to Nursing and Allied Health)	http://www.ebscohost.com/cinahl/	Nursing and Allied Health	Avenell, 2001 Betran, 2005 Brettle, 2001 Stevinson, 2004 Subirana, 2005 Walker-Dilks, 2008 Wong, 2006
CommunityWISE	http://www.oxmill.com/communitywise/	Community issues including community health	
EMBASE	http://www.embase.com/	Biomedical with and emphases on drugs an pharmaceuticals, more non-US coverage than MEDLINE	Avenell, 2001 Minozzi, 2000 Sampson, 2003 Suarez-Almozar, 2000
EMCare	http://www.elsevier.com/wps/find/bibliographicdatabasedescription.cws_home/708272/description#description	Nursing and allied health	Ulincy, 2006
Global Health	http://www.cabi.org/datapage.asp?iDocID=169	International Health	Fitzpatrick, 2006
HaPI (Health and Psychosocial Instruments)	http://www.ovid.com/site/catalog/DataBase/866.jsp	Health and psychosocial testing instruments	Arnold, 2006
IPA (International Pharmaceutical Abstracts)	http://www.csa.com/factsheets/ipa-set-c.php	Drugs and Pharmaceuticals	Fishman, 1996 Wolfe, 2002
MANTIS (Manual Alternative and Natural Therapy Index System)	http://www.healthindex.com/MANTIS.aspx	Osteopathy, Chiropractic and Alternative Medicine	Hoffecker, 2006 Murphy, 2003 Tomasulo, 2001
PsycINFO	http://www.apa.org/pubs/databases/psycinfo/index.aspx	Psychological literature	Eady, 2008 Pilkington, 2007 Stevinson, 2004
Sociological Abstracts	http://www.csa.com/factsheets/socioabs-set-c.php	Sociology including: Health and Medicine and the Law, Social psychology and Substance abuse and addiction	DeLuca, 2008
Social Services Abstracts	http://www.csa.com/factsheets/ssa-set-c.php	Social Services including: mental health services, gerontology and health policy	Taylor, 2007
Citation Tracking Databases			
Database	URL	Subscription Status	References
Google Scholar	http://scholar.google.com/	Free	Falagas, 2008 Jasco, 2005 Bakkalbasi, 2006
PubFocus	http://pubfocus.com/	Free	Plikus, 2006

PubReMiner	http://bioinfo.amc.uva.nl/human-genetics/pubreminer/	Free	
Scopus	http://info.scopus.com/	Subscription Required	Falagas, 2008 Salsbury, 2009 Jasco, 2005 Bakkalbasi, 2006
Web of Science	http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science	Subscription Required	Falagas, 2008 Salsbury, 2009 Jasco, 2005 Bakkalbasi, 2006

References

So Whatever Happened to ERIC? Searcher. 2006;14(2):10-18.

Adebonojo LG, Earl MF. POPLINE. A valuable supplement for health information. Database Magazine1994. pp. 112–15.

Aker PD, McDermaid C, Opitz BG, et al. Searching chiropractic literature: a comparison of three computerized databases. J Manipulative Physiol Ther 1996 Oct;19(8):518–24.

Anne T-V. Information needs of CME providers: Research and development resource base in continuing medical education. Journal of Continuing Education in the Health Professions 1995;15(2):117–21.

Arnold SJ, Bender VF, Brown SA. A Review and Comparison of Psychology-Related Electronic Resources. Journal of Electronic Resources in Medical Libraries 2006;3(3):61–80.

Avenell A, Handoll HH, Grant AM. Lessons for search strategies from a systematic review, in The Cochrane Library, of nutritional supplementation trials in patients after hip fracture. Am J Clin Nutr 2001 Mar;73(3):505–10.

Bakkalbasi N, Bauer K, Glover J, et al. Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomedical Digital Libraries 2006;3(7).

Banks JL. PILOTS (Published International Literature on Traumatic Stress) database. J Trauma Stress 1995 Jul;8(3):495–7.

Bennett S, Hoffmann T, McCluskey A, et al. Introducing OTseeker (Occupational Therapy Systematic Evaluation of Evidence): a new evidence database for occupational therapists. Am J Occup Ther 2003 Nov-Dec;57(6):635–8.

Bennett S, McKenna K, Tooth L, et al. Strong J. Searches and content of the OTseeker database: informing research priorities. Am J Occup Ther 2006 Sep-Oct;60(5):524–30.

Betran AP, Say L, Gulmezoglu AM, et al. Effectiveness of different databases in identifying studies for systematic reviews: experience from the WHO systematic review of maternal morbidity and mortality. BMC Medical Research Methodology. 2005 Jan 28;5(1):6.

Brettle AJ, Long AF. Comparison of bibliographic databases for information on the rehabilitation of people with severe mental illness. Bull Med Libr Assoc 2001 Oct;89(4):353–62.

DeLuca JB, Mullins MM, Lyles CM, et al. Developing a Comprehensive Search Strategy for Evidence Based Systematic Reviews. Evidence Based Library & Information Practice 2008;3(1):3–32.

Eady AM, Wilczynski NL, Haynes RB. PsycINFO search strategies identified methodologically sound therapy studies and review articles for use by clinicians and researchers. Journal of Clinical Epidemiology 2008 Jan;61(1):34–40.

Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J 2008 Feb;22(2):338–42.

Fishman DL, Stone VL, DiPaula BA. Where should the pharmacy researcher look first? Comparing International Pharmaceutical Abstracts and MEDLINE. Bull Med Libr Assoc 1996 Jul;84(3):402–8.

Fitzpatrick RB. Global health database. Med Ref Serv Q 2006 Summer;25(2):59–67.

Fitzpatrick RB. REHABDATA: a disability and rehabilitation information resource. Med Ref Serv Q 2007 Summer;26(2):55–64.

Fitzpatrick RB. PEDro: a physiotherapy evidence database. Med Ref Serv Q 2008 Summer;27(2):189–98.

Flemming K, Briggs M. Electronic searching to locate qualitative research: evaluation of three strategies. J Adv Nurs 2007 Jan;57(1):95–100.

Giglia E. PEDro: this well-known, unknown. Physiotherapy Evidence Database. Eur J Phys Rehabil Med 2008 Dec;44(4):477–80.

Gwynne-Smith D. The Development of Social Care Online. Legal Information Management. 2007 Spring;7(1):34–41.

Hochstein C, Arnesen S, Goshorn J. Environmental Health and Toxicology Resources of the United States National Library of Medicine. Medical Reference Services Quarterly 2007 Fall;26(3):21–45.

Hoffecker L, Reiter CM. A Review of Seven Complementary and Alternative Medicine Databases. Journal of Electronic Resources in Medical Libraries 2006;3(4):13–31.

Jasco P. As we may search—Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. Current Science 2005;89(9):1537–47.

Kubany ES. Searching the traumatic stress literature using PILOTS and PsycLIT. J Trauma Stress 1995 Jul;8(3):491–4.

LaGuardia C. DATABASE & DISC REVIEWS. Library Journal 2002: 166.

Lerner F, National Center for Post-Traumatic Stress D. PILOTS database user's guide. [White River Junction, VT]: Dept of Veterans Affairs, National Center for Posttraumatic Stress Disorder; 2007.

Minozzi S, Pistotti V, Forni M. Searching for rehabilitation articles on MEDLINE and EMBASE. An example with cross-over design. Arch Phys Med Rehabil 2000 Jun;81(6):720–2.

Moseley AM, Herbert RD, Sherrington C, et al. Evidence for physiotherapy practice: a survey of the Physiotherapy Evidence Database (PEDro). Aust J Physiother 2002;48(1):43–9.

Murphy LS, Reinsch S, Najm WI, et al. Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators. BMC Complement Altern Med 2003 Jul 7;3:3.

Petrosio A, Boruch R, Cath R, et al. The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR)" To Facilitate the Preparation and Maintenance of Systematic Reviews of Social and Educational Interventions. Evaluation and Research in Education 2000;14(3 & 4):206–19.

Pilkington K. Searching for CAM evidence: an evaluation of therapy-specific search strategies. J Altern Complement Med 2007 May;13(4):451–9.

Plikus MV, Zhang Z, Chuong CM. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinformatics 2006 Oct 2;7:424.

Salisbury L. Web of Science and Scopus : A Comparative Review of Content and Searching Capabilities. The Charleston Advisor 2009 July;11(1):5–18.

Sampson M, Barrowman NJ, Moher D, et al. Should meta-analysts search Embase in addition to Medline? [see comment]. Journal of Clinical Epidemiology [meta-analysis] 2003 Oct;56(10):943–55.

Sherrington C, Herbert RD, Maher CG, et al. PEDro. A database of randomized trials and systematic reviews in physiotherapy. Man Ther. 2000 Nov;5(4):223–6.

Stevinson C, Lawlor DA. Searching multiple databases for systematic reviews: added value or diminishing returns? Complement Ther Med 2004 Dec;12(4):228–32.

Suarez-Almazor ME, Belseck E, Homik J, et al. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Controlled Clinical Trials 2000 Oct;21(5):476–87.

Subirana M, Sola I, Garcia JM, et al. A nursing qualitative systematic review required MEDLINE and CINAHL for study identification. Journal of Clinical Epidemiology 2005 Jan;58(1):20–25.

Taylor B, Wylie E, Dempster M, et al. Systematically Retrieving Research: A Case Study Evaluating Seven Databases. Research on Social Work Practice 2007:697–706.

Tomasulo P. A new source of herbal information on the Web: the IBIDS database. Med Ref Serv Q 2000 Spring;19(1):53-57.

Tomasulo P. MANTIS—Manual, Alternative, and Natural Therapy Index System Database. Med Ref Serv Q 2001 Fall;20(3):45–55.

Tomasulo PA. AgeLine: free and valuable database from AARP. Med Ref Serv Q 2005 Fall;24(3):55–65.

Ulincy L. EMCare. Journal of the Medical Library Association 2006;94(3):357–60.

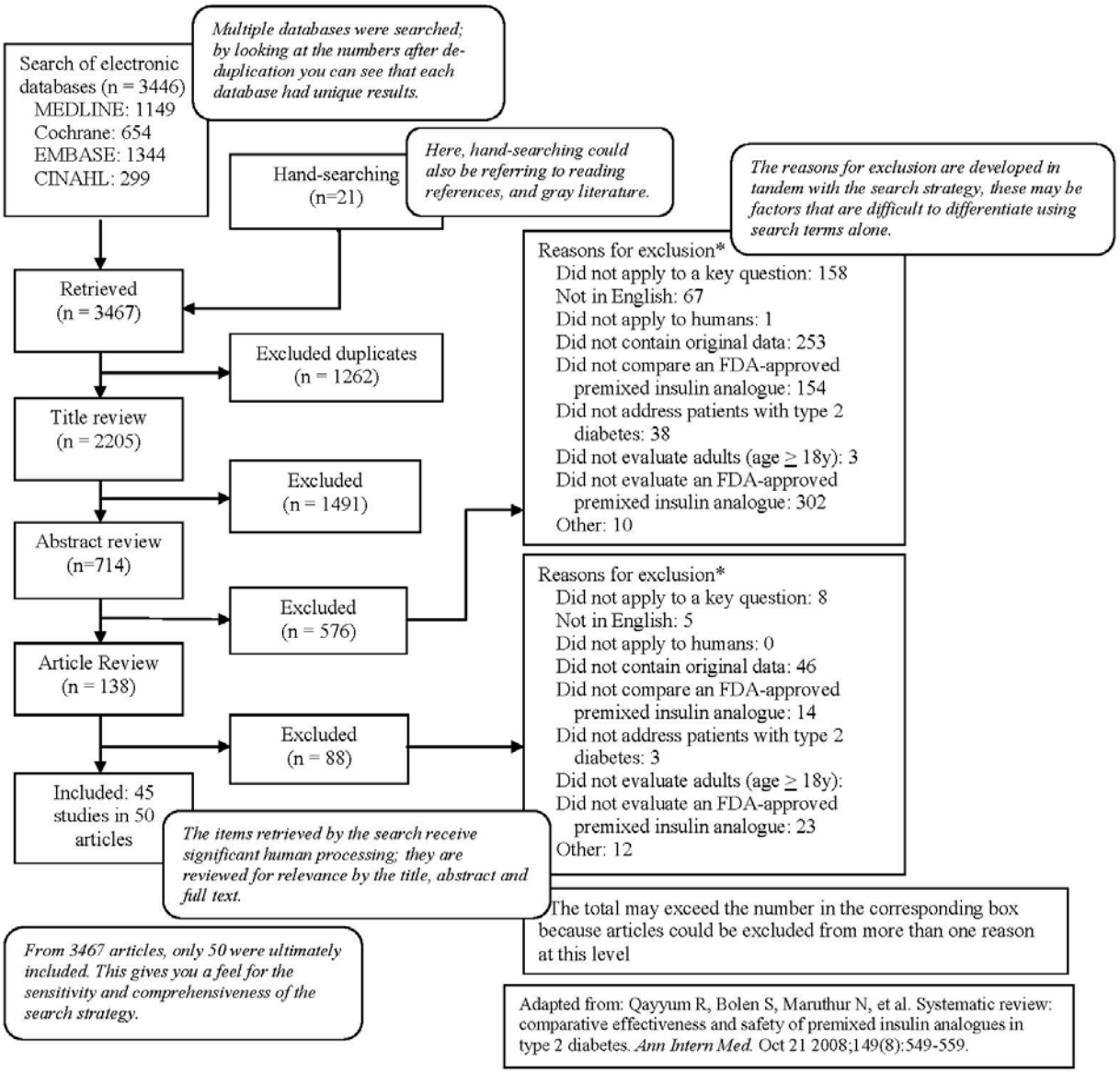
Walker-Dilks C, Wilczynski NL, Haynes RB. Cumulative Index to Nursing and Allied Health Literature search strategies for identifying methodologically sound causation and prognosis studies. Appl Nurs Res 2008 May;21(2):98–103.

Wang J. TRIS Online. Charleston Advisor 2001: 43–6.

Wolfe C. International Pharmaceutical Abstracts: what's new and what can IPA do for you? Am J Health Syst Pharm. 2002 Dec 1;59(23):2360–1.

Wong SS, Wilczynski NL, Haynes RB. Optimal CINAHL search strategies for identifying therapy studies and review articles. J Nurs Scholarsh 2006;38(2):194–9.

Chapter 3 Appendix C. PRISMA-Style Flow Diagram of Literature Search, Annotated



For more on the PRISMA flow diagram, see www.prisma-statement.org/statement.htm.

Chapter 4. Selecting Observational Studies for Comparing Medical Interventions

Susan Norris, David Atkins, Wendy Bruening, Steven Fox, Eric Johnson, Robert Kane, Sally C. Morton, Mark Oremus, Maria Ospina, Gurvaneet Randhawa, Karen Schoelles, Paul Shekelle, Meera Viswanathan

Key Points

- Systematic reviewers disagree about the ability of observational studies to answer questions about the benefits or intended effects of pharmacotherapeutic, device, or procedural interventions.
- This paper provides a framework for decisionmaking on the inclusion of observational studies to assess benefits and intended effects in comparative effectiveness reviews
- Comparative effectiveness reviewers should routinely assess the appropriateness of inclusion of observational studies for questions of benefit, and the rationale for inclusion or exclusion of such studies should be explicitly stated in reviews.

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

- Are there gaps in the evidence from randomized controlled trials?
- Will observational studies provide valid and useful information?

Introduction

While systematic reviewers disagree about the role of observational studies in answering questions about the benefits or intended effects of interventions, there is widespread agreement that observational studies, particularly those derived from large clinical and administrative databases, should be used routinely to identify and quantify potential adverse events.[1-3] Existing systematic reviews vary significantly in the use of observational studies for questions of efficacy or effectiveness of interventions.[4,5] This variation stems in part from concerns regarding the risk of bias in observational intervention studies, particularly the recognition that intended effects are more likely to be biased by preferential prescribing based on patients' prognosis.[6,7] In addition, the inclusion of data from observational studies increases the time and resources required to complete a comparative effectiveness review (CER) which is already a time- and resource-intensive endeavor.

We identified no conceptual framework for when to consider observational studies for inclusion in reviews of beneficial effects and we found no protocols on how to incorporate observational studies into the CER process for questions of benefit. While Cochrane reviews focus primarily on randomized trials, the Cochrane Handbook[8] notes that nonrandomized studies may be included in reviews to provide: (1) an explicit evaluation of their weaknesses; (2) evidence on interventions that cannot be randomized; or (3) evidence of effects that cannot be adequately studied in randomized trials. There is also a lack of consensus on how to assess the risk of bias in observational studies, although several groups have delineated the important domains, based on both empiric evidence and expert opinion.[9,10] Guidelines for reporting epidemiologic studies have been recently developed by an international collaboration and

adopted by many journals.[11] Although these criteria do not assess the risk of bias directly, they may assist systematic reviewers in thinking about bias in this type of observational study.

Our objective is to provide a conceptual framework for the inclusion of observational studies in CERs examining beneficial or intended effects of pharmacotherapeutic, device, or procedural interventions. CERs expand the scope of a typical systematic review, which focuses on the efficacy or 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.[12]

Since data from randomized controlled trials (RCTs) are often insufficient to address all aspects of a CER question on benefits, systematic reviewers should refrain from developing protocols that a priori rule out the use of observational studies when assessing the comparative effectiveness of interventions. Instead, when developing a CER protocol, investigators should examine the potential biases associated with including observational studies pertinent to the questions specified for the review. We outline an approach and various factors to consider in the decision to include or exclude observational studies in CERs. Rather than providing an exhaustive discussion of the potential sources of bias in observational studies, we present key issues relevant to the decision to include or exclude the body of evidence of observational studies.

Observational studies of interventions are defined herein as those where the investigators did not assign exposure; in other words, these are nonexperimental studies. Observational studies include cohort studies with or without a comparison group, cross-sectional studies, case series, case reports, registries, and case-control studies.

The Agency for Healthcare Research and Quality (AHRQ) convened a workgroup to address the role of observational studies in CERs. The workgroup used a consensus process to arrive at our recommendations. This process is detailed in another paper in this series. [12]

Decision Framework

In considering whether to use observational studies in CERs for addressing beneficial effects, systematic reviewers should answer two questions:

1. Are There Gaps in the RCT Evidence for the Review Questions Under Consideration?

Data from RCTs may be insufficient to address a review question about benefit for a number of reasons.[13] RCTs may be inappropriate due to patient values or preferences; the intervention may be hazardous; or randomization may decrease benefit if the intervention effect depends in part on subjects' active participation based on their beliefs and preferences. RCTs may be unnecessary in interventions with obvious benefit, such as the treatment of susceptible organisms with penicillin or where the alternative to treatment of a new and otherwise fatal disease is a high likelihood of death. RCTs may be difficult to implement due to entrenched clinical practice or to active consumer pressure for access to a treatment, problems with recruitment when a drug is already marketed, the need for long-term followup to detect either benefits or harms, or difficulty randomizing feasible intervention units. In situations where RCT data are impractical, infeasible, or incomplete, observational studies may provide valid and useful data to help address CER questions.

Gaps in the RCT evidence available to answer review questions can be identified at a number of points in the review. First, gaps may be identified when refining the questions for the review and may be explicitly outlined in the original review protocol or work plan. Second, existing reviews on related topics or consultation with clinical experts may also identify important gaps in the RCT evidence at the protocol stage of a CER. Third, gaps may also be identified during the initial search of titles and abstracts, where, for example, the review team finds that all the RCTs involve short-term outcomes or that RCTs lack information about a key outcome of interest. A fourth point at which gaps in RCT data are frequently identified occurs after detailed review of the available RCT data.

The criteria in Table 1 can be used at any of these points in the review process to determine whether RCT data are sufficient to address a CER question about benefit or the balance of benefits and harms. These criteria closely resemble those criteria used by the GRADE group[14] and by AHRQ Evidence-based Practice Centers (EPC) to assess the quality of a body of evidence.[15]

Table 1. Criteria for assessing whether a body of evidence from RCT data is sufficient to address a question of benefits or the balance of benefits and harms

Criteria	Definition	Considerations
Risk of bias (internal validity)	The degree to which the observed effect may be attributed to factors other than the intervention under review; potential bias should be minimized and confounding adjusted for, 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 degree to which reported effect sizes from included studies appear to have the same direction of effect.	Inconsistency may be due to heterogeneity across PICOTS or the etiology may not be apparent.
Directness	Whether the RCT evidence links the interventions directly to health outcomes. Indirect evidence can encompasses intermediate or surrogate outcomes, or refers to the situation when two or more bodies of evidence are needed to compare interventions.	The important outcomes are usually health outcomes such as coronary events or mortality, but the available data are often surrogate, intermediate, or physiologic outcomes.
Precision	The degree of certainty surrounding an effect estimate for a given outcome. Includes sample size, number of studies, and heterogeneity within or across studies.	Greater levels of precision may be needed if the estimates of the effect size of benefits and harms are closely balanced or if either is near a threshold that decision makers might use to make a recommendation.
Outcome reporting bias	The extent to which authors of RCTs appear to have reported all outcomes examined and there is no strong evidence for publication bias (at the study level).	The presence of outcome reporting bias can be difficult to determine, but may be inferred when important outcomes or contributors to a composite outcome are missing, or when small studies demonstrate skewed treatment effects (as in an asymmetric funnel plot).
Applicability	The extent to which the data from RCTs 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.

Key: CER=comparative effectiveness review; PICOTS=population, intervention, comparator, outcomes, setting; RCTs=randomized controlled trials

This table is adapted from the work of Owens and colleagues[15] and the work of the Methods Guide for Effectiveness and Comparative Effectiveness Reviews: Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program.[16]

Table 2 lists situations where observational studies were considered at various stages of the CER, along with examples. One very compelling situation for considering observational studies in a CER for a question of benefit occurs when all RCTs can be classified as efficacy studies and the need for inclusion of observational studies is apparent at the outset (Table 2, example 1).[17] Although efficacy trials are not synonymous with poor applicability to clinical populations of interest to the CER questions, such RCTs often recruit selected populations that are not representative of the population affected by the condition of interest, may involve intensively administered interventions, and may not adequately examine longer-term, patient-centered outcomes.[18] Thus when all RCTs identified for a CER have selected or narrow populations, the applicability of these data to more general populations is likely poor and apparent at the outset. High-quality observational studies can help address these gaps.

Table 2. Examples of the use of observational studies in comparative effectiveness reviews

Example 1. Need to include observational studies is clear at the onset of the review In a review of antipsychotic medications[17] short-term efficacy trials evaluated a relatively narrow spectrum of patients with schizophrenia, raising a number of 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 drugs have better persistency and sustained effectiveness for longer-term followup (e.g., 6 months to 2 years)? Given this multitude of questions not addressed by RCTs, these review authors determined that they would examine and include observational studies from the outset of the review.
Example 2. Expert input raises questions about applicability to clinical populations A review of percutaneous coronary intervention (PCI) versus coronary artery bypass (CABG) for coronary disease identified 23 RCTs conducted from 1987 to 2002.[19] 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 versus 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.
Example 3. Trial data are sufficient The clinical question of antioxidant supplementation to prevent heart disease has been studied in numerous large clinical trials, including among 20,536 elevated-risk subjects participating in the Heart Protection Study.[20] 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.
Example 4. Paucity of trial data and inadequacy of available evidence In a recently completed EPC report (AHRQ Report #148) on heparin to treat burn injury[21] 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.
Example 5. Important outcomes are not captured in RCTs More than 50 RCTs of triptans focused on the speed and degree of migraine pain relief related to a few isolated episodes of headache.[22] 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 data that a particular triptan improved work productivity—information that was not available in RCTs.

Table 2. Examples of the use of observational studies in comparative effectiveness reviews (continued)

Example 6. Potential selection bias: 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[23] 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 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. Thus, at baseline, 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.

Key: EPC=Evidence-based Practice Center of the Agency for Healthcare Research and Quality; RCT=randomized controlled trial

In other cases, content experts and decision makers may raise concerns about whether trial results are applicable to the full spectrum of patients with the condition of interest (Table 2, example 2).[19] Later in the review process, a thorough review of the characteristics of the available RCTs may reveal whether the interventions or patient populations are representative of those found in current practice.[24] Guidance on the assessment of study characteristics for applicability to populations and settings of clinical interest is found in another paper in this series.[16]

Identifying gaps with initial consideration of the review questions or after discussion with content experts, may lead the team to perform their initial searches very broadly, to identify both RCT and observational study evidence in the same search. On the other hand, reviewers may choose to do these searches sequentially and search for observational studies only after reviewing in detail all the identified RCTs. Whether reviewers choose one strategy or the other, the important point is that there is an explicit assessment of whether there are gaps in the RCT evidence, and if so, there is explicit consideration of the potential usefulness of observational studies to help fill these gaps. If RCT data are sufficient to answer the key questions about benefit or the balance of benefits and harms, reviewers do not need to consider observational study designs. In Table 2, example 3, reviewers found conclusive RCT data, and they therefore did not assess observational studies of antioxidant supplementation.[20] It is expected that in most CERs, however, gaps will be present and observational studies should be considered for inclusion.

In Table 2, example 4,[21] the review authors identified very few RCTs in a preliminary search and after input from experts, and therefore planned to consider including observational studies prior to running the primary search and detailed review of the trials. A paucity of RCT evidence is common, particularly for many surgical and diagnostic procedures, and for therapeutic devices.

Failure of RCTs to include all important outcomes is common. In Table 2, example 5, a large number of head-to-head efficacy trials were available, but they provided insufficient evidence to assess two important long-term outcomes.[22]

2. Will Observational Studies Provide Valid and Useful Information To Address Key Questions?

To answer this question, reviewers need to perform three steps, while explicitly presenting decisions on inclusion and exclusion of observational studies and carefully describing the rationale for those decisions.

a. Refocus the review questions on gaps in the RCT evidence. Specifying the PICOTS (population, intervention, comparator, outcome, timing, and study design) characteristics for gaps in the RCT evidence guides subsequent steps in assessing whether observational studies will be helpful. This step does not likely involve a substantive change in the review questions, which ideally were framed a priori in a review protocol, but rather a change in focus such that the (RCT) gap questions are clear to the reviewer and reader.

b. Assess the risk of bias of observational studies to answer the gap review questions. The suitability of observational studies for assessing intervention effectiveness in CERs depends on the potential for bias. In deciding whether to include observational studies in a CER, the assessment of potential for bias is based on an appraisal of the body of observational studies as a whole, and is not based on the characteristics and internal validity of the individual observational studies. Detailed examination of the potential for bias in a subset of the relevant observational studies may, however, inform the global assessment of the body of observational studies.

Work by Glasziou and colleagues suggests a procedure for implementing this advice: Before looking at individual observational studies, consider whether the clinical context and natural history of disease would make observational studies unsuitable.[25] Specifically, Glasziou and colleagues considered various clinical examples to identify conditions in which observational studies were likely or unlikely to provide valid and meaningful answers to questions about efficacy. They found that fluctuating or intermittent conditions are much more difficult to assess with observational studies. 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. Observational studies of interventions for diseases with stable or steadily progressing courses, however, may be useful. For example, individuals afflicted with amyotrophic lateral sclerosis steadily decline in function and spontaneous recovery is virtually unknown and a cohort study that compared group responses to an intervention over time may demonstrate meaningful effects.

Poor-quality evidence from observational studies should not be used or relied on, even if it appears to address gaps in the trial evidence. Internal validity is always central to answering a review question. Observational studies with low risk of bias, however, may provide more useful data than RCTs with respect to applicability to populations of interest.

Five main biases can affect intervention research: selection, performance, detection, attrition, and selective outcomes reporting bias.[8] Thoughtful consideration of the potential for these biases in the body of relevant observational studies will help to determine the suitability of these studies for inclusion in a CER. In some clinical circumstances the likelihood of one or more of these biases affecting studies is so high that observational studies can be excluded as a group prior to detailed review of the body of observational evidence.

The primary distinguishing factors between RCTs and observational studies is the potential for selection bias, which must be carefully considered to determine if observational studies as a group are suitable for inclusion or exclusion in a CER for questions of benefit or the balance of benefits and harms. Selection bias refers to systematic differences among the groups being compared that arise from patient or physician selection of treatments, or the association of treatment assignments with demographic, clinical, or social characteristics that relate to outcome. The result of selection bias is that differences among the compared groups in prognosis, likelihood of adherence to treatment regimes, responsiveness to treatment, susceptibility to

adverse effects, and the use of cointerventions can obscure or overestimate the effects of the intervention being examined.[26]

To make decisions about the severity of selection bias when considering the suitability of observational studies for examination of benefits in CERs, reviewers should examine the specific type and cause. When different diagnoses, severity of illness, or comorbid conditions are important reasons for physicians to assign different treatments, selection bias is called "confounding by indication" (Table 2, example 6).[23] Confounding by indication is a common problem in pharmacoepidemiological studies comparing beneficial effects of interventions because physicians often assign treatment based on their expectations of beneficial effects.

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 followup. Moreover, the control group is 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 methodological rigor by excluding established users of the drug and following only patients with new drug use,[27] although determining who is a new user from administrative claims data can be challenging.

Systematic reviewers should look carefully for how investigators defined new users. 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 make it unlikely that apparent new users are actually former users returned from an extended drug holiday.

It is also useful to determine 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 be prone to bias when patients who fail one drug in a class switch to a different drug in the same class. The least biased observational studies require all patients in the cohort to be new users of the entire class of drugs related to the review question.

Performance bias refers to systematic differences in the care provided to participants in the comparison groups other than the intervention under investigation.[26] Because retrospective observational studies are virtually never double-blinded, treatment groups may differ in their expectations, information, and enthusiasm. 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 cointerventions (provision of unintended additional care to either comparison group) occur more often in observational studies and are much more likely to go undetected than in RCTs. Thus with complex or multi-component interventions, it may not be possible to separate out the effect of the intervention from other factors affecting outcomes. In such situations, observational studies may not be suitable for inclusion in a CER.

Attrition and detection bias usually require assessment at the individual study level: their consideration a priori will not likely lead to exclusion of the body of observational studies. Rather, the assessment and impact of these biases is addressed first at the individual study level and then synthesized across the body of evidence. 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. The issues raised by attrition bias in observational studies are similar to those in RCTs.

Systematic differences in outcomes assessment among the comparison groups (detection bias)[26] can be effectively countered in observational studies with well-designed registries, for example. Thus observational studies will not likely be excluded as a group because of concerns about this type of bias. Detection bias is important in cohort studies in which outcomes in comparison groups may be assessed at different time points by nonblinded assessors, using different measurement techniques, quality control, and outcome definitions. This 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.

Selective outcome reporting is defined as the selection of a subset of the original variables recorded on the basis of the results, for inclusion in the study publications.[28] The main concern is that statistically nonsignificant results might be selectively withheld from publication. Selective outcome reporting can occur in a number of ways, including selective omission of outcomes from reports, selective choice of data for an outcome, selective reporting of analyses using the same data, selective reporting of subsets of the data, and selective underreporting of data.[26] There are data to suggest that selective outcome reporting is common in RCTs[29-31] although data are sparse on reporting practices in observational studies.[32]

We do not consider an assessment of magnitude of effect a criterion for including or excluding the body of observational studies. Magnitude of benefits (or harms) and the various types of bias are, however, all used in the assessment of the strength of a body of evidence of observational studies according to well-accepted approaches.[33] In the GRADE schema, the quality of a body of observational studies is downrated (with respect to RCTs) unless the effect size is large, as the observed effect may be due to biases and random variation rather than the effect of the intervention.[33]

c. Assess whether observational studies address the review questions. Even when RCT data are insufficient and the risk of bias does not preclude the inclusion of observational studies, such studies will only be suitable for filling in the gaps if they provide additional evidence that is relevant to the review question, including the specific PICOTS characteristics of interest. For example, high-quality observational studies that focus on outcome measures such as persistency or adherence to therapy will be relevant to a CER, as such data from RCTs may be obtained from highly selected subjects (e.g., after a run-in period), with closely monitored and intensely implemented interventions.

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 will add useful information. Procedure registries may have higher internal validity than other types of observational studies because the data are typically collected prospectively according to a protocol and the date of the procedure serves as an inception date. The inception date allows investigators to measure characteristics that may have influenced the choice of procedure (e.g., ventricular assist devices) and control potential confounding. The inception date also allows investigators to capture the benefits and harms that occurred after a procedure. For example, INTERMACS® is a national registry in the United States that enrolls patients who have received ventricular assist devices for end-stage heart failure and follows them for quality of life endpoints and the incidence of rehospitalization (http://www.intermacs.org). The INTERMACS registry has the support of Federal decisionmakers, including the U.S. Food

and Drug Administration and the Center for Medicare and Medicaid Services. Registries in which enrollment has been defined by procedures may be more valid for comparative effectiveness research than registries in which enrollment has been defined by disease onset because disease-based registries aren't designed in relation to an intervention's inception date.

As a further 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. A potential advantage of this type of study is that long-term benefits, tolerability, and harms 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 short time period (e.g., 6 weeks), have much lower withdrawal rates than the broader population of interest in a CER.

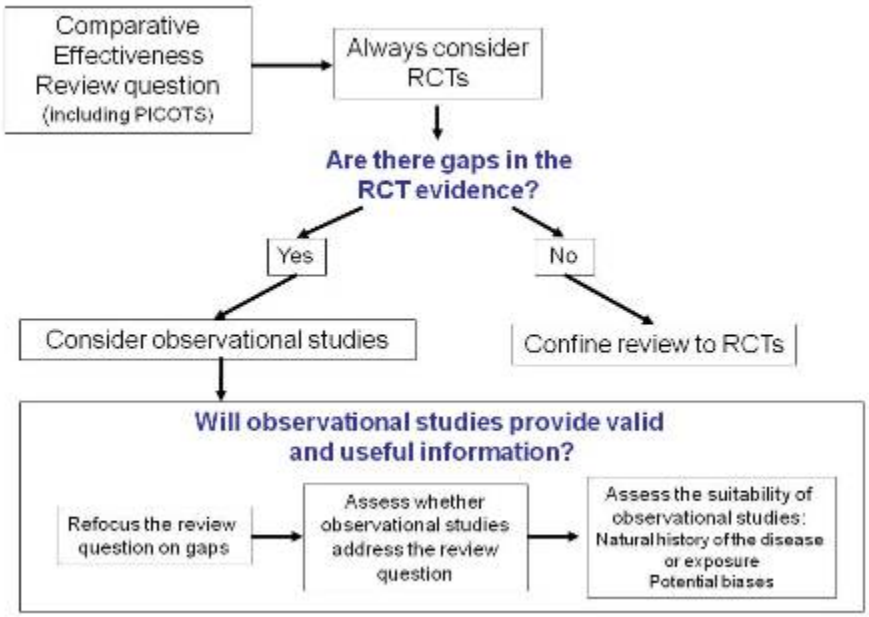
Many data sources for observational studies are suited to long-term followup but are limited in the type of outcomes that can be measured. For example, databases that combine data from hospitalization databases, vital registries, claims data, and laboratory, pharmacy, and clinical records through deterministic or probabilistic data linkage 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 and health services utilization (e.g., emergency room visits, hospital admissions, discontinuation of a drug, initiation of a drug associated with treatment of an adverse effect, or a surgical procedure). With few exceptions, however, administrative and clinical databases lack data on quality of life, severity of symptoms, and function. In future, electronic health records may enable the retrieval of rich clinical, observational data.

Some study designs are more suitable for examining treatment effects in patients who have diseases that have an unpredictable natural history. For example, valid data on the beneficial effects of an intervention in a fluctuating condition may be gained from prospective, interrupted time-series studies with an active control group, where data were collected at regular intervals according to a protocol developed a priori. In prospective observational studies, all precautions against bias that can be taken should be—for example, even if it is not possible to mask treatment assignment from patients and clinicians, outcome assessors may be blinded.

Discussion

The conceptual framework for making decisions as to whether observational studies should be included in CERs needs to be implemented in an explicit and efficient manner. CER work groups can implement the approach recommended herein (see Figure 1) in a variety of ways, but the following steps may be a useful guide. In the CER work plan or protocol, reviewers start with a clearly defined review question with respect to PICOTS, followed by a preliminary search for relevant trials and systematic reviews, and consultation with topic experts. Well-known or large RCTs should be examined in detail at this stage. If these studies address all important aspects of the review questions, then observational studies may not need to be included. Since this rarely occurs, reviewers need to justify any decision to exclude observational studies in this or subsequent steps. In addition, reviewers should outline in the review protocol the approach to considering the inclusion of observational studies.

Figure 1. Flow diagram for consideration of observational studies for comparative effectiveness questions concerning benefit



Key: PICOTS=population, intervention, comparator, outcomes, timing, study design; RCTs=randomized, controlled trial.

If during this preliminary review, data from RCTs do not appear to be sufficient to answer the review questions concerning benefit, then reviewers should proceed to assess the potential risk of bias in a body of observational studies used to answer gap questions. This assessment will focus particularly on issues of the natural history of the condition under study and selection and performance bias. Potential biases that vary across individual observational studies (such as detection and attrition bias) are not considered in this global assessment of observational studies, but rather are assessed at the individual study level if observational studies are included in the CER.

If observational studies are likely to provide valid data on important outcomes, the CER team then proceeds with a systematic search for these studies. If reviewers have knowledge of gaps in RCT data early in the review process and observational studies are deemed likely to be useful, then the review team may choose to search for trials and observational studies concurrently. Ideally, sensitive and specific search strategies will be developed in the future to identify observational studies with designs that are considered most appropriate to address a review question, or to identify other markers of relevant, high-quality observational studies in bibliographic database searches.

As observational studies are examined and reviewers become further informed on the clinical topic, the risk of bias in observational studies can be further understood. It may be decided that the risk is excessive with any or all types of observational studies, at which time the team abandons their further consideration. If assessment of the risk of bias suggests that the observational evidence may be valid, the team identifies and synthesizes those data. The decision to include or exclude observational studies must be thoughtfully presented in the results section. Quality assessment of both RCTs and included observational studies is performed, with strengths and limitations delineated.

We suggest that observational studies should be considered for questions of benefit in CERs just as for harms. The same basic principle of research synthesis applies to considerations of all types of review questions and evidence: minimize bias at all steps in CER development. Invalid results (i.e., those that cannot be attributed in all likelihood to the intervention) from any study design should not be included or should be labeled as such. Different study designs may be optimal for different types of review questions, and study designs must be assessed for risk of bias with respect to the specific review question. Risk of bias is just as important a consideration in using observational studies for harms as for benefits or intended effects.

Conclusions

It is unusual to find sufficient evidence from RCTs to answer all key questions about benefits or the balance of benefits and harms, therefore the default approach for CERs should be to consider observational studies for questions of benefit or intended effects of interventions. There is no a priori reason to exclude observational studies for questions of benefit. Rather, observational studies should be evaluated using the same criteria used to evaluate the inclusion of RCT data, namely whether the observational study results address a key question and whether the observational data are likely to be valid. We promote an explicit approach within the context of each specific review question. In future there should be a formal evaluation of our proposed approach, examining its reliability, sensitivity (i.e., not missing important, valid observational studies), specificity (i.e., not exploring studies that do not provide valid data), and feasibility while optimizing use of systematic review resources.

Acknowledgements

The authors gratefully acknowledge the technical contributions of Nancy Brown, M.L.S., Marcie Merritt, Edwin Reid, M.S., and Jill Rose.

Author Affiliations

Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR (SN). VA Quality Enhancement Research Initiative (QUERI), Washington, DC. (DA). ECRI Institute, Plymouth Meeting, PA, (WB, KS). Agency for Healthcare Research and Quality, Rockville, MD, (SF, GR). The Center for Health Research, Kaiser Permanente Northwest, and Oregon Evidence-based Practice Center, Portland, OR, (EJ). Minnesota Evidence-Based Practice Center, Minneapolis, MN, (RK), RTI International, Triangle Park, NC, (SGM, MV). Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON (MO). University of Alberta Evidence-Based Practice Centre, Edmonton, AB, (MO). Southern California Evidence-Based Practice Center, RAND Corporation, Los Angeles, CA, (PS).

This paper has also been published in edited form: Norris S, Atkins D, Bruening W, et al. Observational studies in systematic reviews of comparative effectiveness. J Clin Epidemiol 2010;63: in press.

References

1. Laupacis A, Paterson JM, Mamdani M, et al. Gaps in the evaluation and monitoring of new pharmaceuticals: proposal for a different approach. Can Med Assoc J 2003;169:1167–70.
2. Etminan M, Gill S, Fitzgerald M, et al. Challenges and opportunities for pharmacoepidemiology in drug-therapy decision making. J Clin Pharmacol 2006;46:6–9.
3. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2010 May;63(5):502–12.
4. Moja LP, Telaro E, D'Amico R, et al. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 2005;330:1053–57.
5. Norris SL, Atkins D. Challenges in using nonrandomized studies in systematic reviews of treatment interventions. Ann Intern Med 2005;142:1112–19.
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323–37.
7. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004;363:1728–31.
8. Higgins JP, Green S, eds. Cochrane handbook for systematic reviews of interventions Chichester, UK: John Wiley & Sons, Ltd; 2006.
9. Deeks JJ, Dinnes J, D'Amico R, et al. International stroke trial collaborative group and European carotid surgery trial collaborative group. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii–x, 1–173.
10. West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). Rockville, MD: Agency for Healthcare Research and Quality. April 2002. AHRQ Publication No. 02-E016.
11. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370:1453–7.
12. Helfand M, Balshem H. AHRQ series, paper 2: principles for developing guidance: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2010;63:484–90.
13. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215–8.
14. GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490–8.
15. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol 2010 May;63(5):513–23.
16. Atkins, D, Chang, S, Gartlehner, G, et al. Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol, under review.
17. McDonagh M, Peterson K, Carson S, et al. Drug class review: atypical antipsychotic drugs, Final report update 2. In: Helfand M, ed. Drug Effectiveness Review Project. Portland, OR: Oregon Evidence-based Practice Center; 2008.
18. Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040–8.
19. Bravata DM, McDonald KM, Gienger AL, et al. Comparative effectiveness of percutaneous coronary interventions and coronary artery bypass grafting for coronary artery disease. Rockville, MD: Agency for Healthcare Research and Quality; 2007. AHRQ Publication No. 08-EHC002-EF.
20. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7–22.

21. Oremus M, Hanson M, Whitlock R, et al. The uses of heparin to treat burn injury. Evidence Report/Technology Assessment No. 148. (Prepared by the McMaster University Evidence-based Practice Center, under Contract No. 290-02-0020). Rockville, MD: Agency for Healthcare Research and Quality; 2006. AHRQ Publication No. 07-E004.
22. Helfand M, Peterson K. Drug class review on the triptans: Drug Effectiveness Review Project. Portland, OR: Oregon Evidence-based Practice Center; 2003.
23. Go AS, Yang J, Gurwitz JH, et al. Comparative effectiveness of different beta-adrenergic antagonists on mortality among adults with heart failure in clinical practice. Arch Intern Med 2008;168:2415–21.
24. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet 2005 Jan 1-7;365(9453):82–93.
25. Glasziou P, Chalmers I, Rawlins M, et al. When are randomised trials unnecessary? Picking signal from noise [see comment]. BMJ 2007;334:349–51.
26. Reeves BC, Deeks JJ, Higgins JP, et al. Chapter 13: Including nonrandomized studies. In: Higgins JP and Green S, eds. Cochrane Handbook for Systematic Reviews. Chichester, UK: Wiley; 2008.
27. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003;158:915–20.
28. Hutten JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. J R Stat Soc Ser C 2000;49:359–70.
29. Chan AW, Hrobjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457–65.
30. Chan AW, Krleza-Jeric K, Schmid I, et al. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. Can Med Assoc J 2004;171:735–40.
31. Furukawa TA, Watanabe N, Omori IM, et al. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. JAMA 2007;297:468–70.
32. Peters J, Mengersen K. Selective reporting of adjusted estimates in observational epidemiology studies: reasons and implications for meta-analyses. Eval Health Prof 2008;31:370–89.
33. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.

Chapter 5. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions

Meera Viswanathan, Mohammed T. Ansari, Nancy D. Berkman, Stephanie Chang, Lisa Hartling, Melissa McPheeters, P. Lina Santaguida, Tatyana Shamliyan, Kavita Singh, Alexander Tsertsvadze, Jonathan R. Treadwell

Key Points

- The task of assessing the risk of bias of individual studies is part of assessing the strength of a body of evidence. In preparation for evaluating the overall strength of evidence, reviewers should separate criteria for assessing risk of bias of individual studies from those that assess precision, directness, and applicability.
- EPCs may choose to use the terms "assessment of risk of bias" or "quality assessment." EPCs should define clearly the term used in their systematic review (SR) and comparative effectiveness review (CER) protocols and describe the constructs included as part of the assessment.
- We recommend that AHRQ reviews:
  - Opt for tools that are specifically designed for use in systematic reviews; have demonstrated acceptable validity and reliability; specifically address items related to methodological quality (internal validity) and preferably are based on empirical evidence of bias; where available, are specific to the study designs being evaluated; and avoid the presentation of risk of bias assessment as a composite score.
  - Do not use study design labels (e.g., RCT, cohort, case-control) as a proxy for assessment of risk of bias of individual studies.
  - Explicitly evaluate risk of selection, performance, attrition, detection, and selective outcome reporting biases.
  - Allow for separate risk of bias ratings by outcome to account for outcome-specific variations in detection bias and selective outcome reporting bias. Categories of outcomes, such as harms and benefits, may have different sources of bias.
  - Select items from recommended criteria for each included study design, as appropriate for the topics.
  - Evaluate validity and reliability of outcome measures as a component of detection bias and fidelity to the protocol as a component of performance bias.
  - Generally speaking, exclude precision and applicability when assessing the risk of bias because these are assessed in other domains when evaluating the strength of a body of evidence.
  - Assess risk of bias based on study design and conduct rather than reporting. Poorly reported studies may be judged as unclear risk of bias.
  - Not rely solely on poor reporting, industry funding, or disclosed conflict of interest, to rate a study as high risk of bias, although reviewers should report these issues transparently.

- Conduct sensitivity analyses, when appropriate, for the body of evidence to evaluate whether poor reporting, industry funding. or disclosed conflict of interest may be associated with the studies' results. Industry funding or other conflict of interest may raise the risk of bias in design, analysis, and reporting. Reviewers suspecting high risk of bias because of industry funding should pay attention to the risk of selective outcome reporting.
- Define decision rules for assessing the risk of bias category for each outcome from an individual study to improve transparency and reproducibility.
- Conduct dual assessment of risk of bias.

Introduction

This document updates the existing Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) Methods Guide for Effectiveness and Comparative Effectiveness Reviews on assessing the risk of bias of individual studies. As with other AHRQ methodological guidance, our intent is to present standards that can be applied consistently across EPCs and topics, promote transparency in processes, and account for methodological changes in the systematic review process. These standards are based on available empirical evidence, theoretical principles, or workgroup consensus: as greater evidence accumulates in this methodological area, our standards will continue to evolve. When possible, our recommended standards offer flexibility to account for the wide range of AHRQ EPC review topics and included study designs.

Some EPC reviews may rely on an assessment of high risk of bias to serve as a threshold between included and excluded studies; in addition, EPC reviews use risk-of-bias assessments in grading the strength of the body of evidence. Assessment of risk of bias as unclear, high, medium, or low may also guide other steps in the review process, such as study inclusion for qualitative and quantitative synthesis, and interpretation of heterogeneous findings.

This guidance document begins by defining terms as appropriate for the EPC program, explores the potential overlap in various constructs used in different steps of the systematic review, and offers recommendations on the inclusion and exclusion of constructs that may apply to multiple steps of the systematic review process. We note that this guidance applies to reviews—such as AHRQ-funded reviews—that separately assess the risk of bias of outcomes from individual studies, the strength of the body of evidence, and applicability of the findings. This guidance applies to comparative effectiveness reviews that require interventions with comparators and systematic reviews that may include noncomparative studies. A key construct, however, is that risk-of-bias assessments judge whether the design and conduct of the study compromised the believability of the link between exposure and outcome. This guidance may not be relevant for reviews that combine evaluations of risk of bias or quality of individual studies with applicability.

Later sections of this guidance document provide guidance on the stages involved in assessing risk of bias and design-specific minimum criteria to evaluate risk of bias. We discuss and recommend tools and conclude with guidance on summarizing risk of bias.

Terminology and Constructs

Differences in Terminology

Risk of bias, defined as the risk of "a systematic error or deviation from the truth, in results or inferences,"[1] is interchangeable with internal validity, defined as "the extent to which the design and conduct of a study are likely to have prevented bias"[2] or "the extent to which the results of a study are correct for the circumstances being studied."[3] Despite the central role of the assessment of the believability of individual studies in conducting systematic reviews, the specific term used has varied considerably across review groups. A common alternative to "risk of bias" is "quality assessment," but the meaning of the term *quality* varies, depending on the source of the guidance. One source defines quality 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."[4] The Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE) uses the term quality to refer to *an individual study* and judgments based about the strength of the *body of evidence* (quality of evidence).[5] The U.S. Preventive Services Task Force (USPSTF) equates quality with internal validity and classifies individual studies first according to a hierarchy of study design and then by individual criteria that vary by type of study.[6] In contrast, the Cochrane collaboration argues for wider use of the phrase "risk of bias" instead of "quality," reasoning that "an emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research (although does not overcome the problem of having to rely on reports to assess the underlying research)."[1]

Because of inconsistency and potential misunderstanding in the use of the tem "quality," this guidance will refer to risk of bias. We understand risk of bias to refer to the extent to which a single study's design and conduct protect against all bias in the estimate of effect using the more precise terminology "assessment of risk of bias." Thus, assessing the risk of bias of a study can be thought of as assessing the risk that the study results reflect bias in study design or execution in addition to the true effect of the intervention or exposure under study.

Guidance on Terminology

This guidance uses risk of bias as the preferred terminology. Nonetheless, we recognize the competing demands for flexibility across reviews to account for specific clinical contexts and consistency within review teams and across EPCs. We advocate transparency of planned methodological approach and documentation of decisions and therefore recommend that EPCs define the term selected in their SR and Comparative Effectiveness Review (CER) protocols and describe the constructs included in the assessment.

Differences in Constructs Included in Risk-of-Bias Assessment

Across prior guidance documents and instruments, the types of constructs included in risk of bias or quality assessments have included one or more of the following issues: (1) conduct of the study/internal validity, (2) random error, (3) external validity or applicability, (4) completeness of reporting, (5) selective outcome reporting, (6) choice of outcome measures, (7) study design, (8) fidelity of the intervention, and (9) conflict of interest in the conduct of the study.

The lack of agreement on what constructs to include in risk-of-bias assessment stems from two sources. First, no strong empirical evidence supports one approach over another; this

gap leads to a proliferation of approaches based on the practices of different academic disciplines and the needs of different clinical topics. Second, in the absence of updated guidance on risk-of-bias assessment that accounts for how new guidance on related components of systematic reviews (such as selection of evidence,[7] assessment of applicability,[8] or grading the strength of evidence[5,9-17]) relate to, overlap with, or are distinct from risk-of-bias assessment of individual studies, some review groups continue to use quality practices that have served well in the past.

In the absence of strong empirical evidence, methodological decisions in this guidance document rely on epidemiological principles.[1] Thus, this guidance document presents a conservative path forward. Systematic reviewers have the responsibility to evaluate potential sources of bias and error if these concerns could plausibly influence study results; we include these concerns even if no empirical evidence exists that they influence study results.

Guidance on Constructs To Include or Exclude From Risk-of-Bias Assessment

The constructs selected in the assessment of risk of bias may differ because of the academic orientation of the reviewers, guidelines by sponsoring organizations, and clinical topic. In AHRQ-sponsored reviews, recent guidance and requirements for systematic reviews have reduced the variability in other related steps of the systematic review process and, therefore, allow for greater consistency in risk-of-bias assessment as well. Some constructs that EPCs may have considered part of risk of bias (or quality) assessment in the past now overlap with or fall within the domains of other systematic review tasks. Table 1 illustrates which constructs to include for each systematic review task when systematic reviews separately assess the risk of bias of individual studies, the strength of the body of evidence, and applicability of the findings for individual studies. We note that the GRADE approach to grading the strength of evidence incorporates applicability within strength of evidence assessments,[12] and the AHRQ-EPC approach does not, but the distinction between concepts relevant for risk of bias and applicability are relevant to both systems.[9]

Table 1. Inclusion and exclusion of constructs for risk-of-bias assessment, applicability, and strength of evidence

Construct	Included in appraisal of individual study risk of bias?	Included in assessing applicability of studies and the body of evidence?	Included in grading strength of the body of evidence?
Risk of bias (from selection bias and confounding, attrition, performance, detection, reporting, and other biases)	Yes	No	Yes (required domain of risk of bias)
Precision	Only when no quantitative pooling or presentation is possible	No	Yes (required domain of precision)
Applicability/external validity	Only when components of applicability influence risk of bias (e.g., duration of follow-up varies across intervention arms)	Yes	Depends on the SOE system. GRADE includes applicability as part of directness, AHRQ-EPC does not (with the exception of rating surrogate outcomes as indirect evidence)

Table 1. Inclusion and exclusion of constructs for risk-of-bias assessment, applicability, and strength of evidence (continued)

Construct	Included in appraisal of individual study risk of bias?	Included in assessing applicability of studies and the body of evidence?	Included in grading strength of the body of evidence?
Poor or inadequate reporting	Yes, studies may be rated as having unclear risk of bias	No	No
Selective outcome reporting	Yes, only when judgments can be made about the impact of differences between outcomes listed in full protocol and published materials	No	Yes
Outcome measures	Yes (potential for outcome measurement bias, specifically validity, reliability, variation across study arms)	Yes (applicability of outcomes measures)	Yes (directness of outcome measures)
Study design	Assessment should evaluate the relevant sources of risk of bias by study design rather than rate the study risk of bias by design labels alone	No	Yes (overall risk of bias is rated separately for randomized and nonrandomized studies)
Fidelity to protocol	Yes	Yes	No
Conflict of interest from sponsor bias	Indirectly (sponsor bias may influence one or more sources of bias)	Indirectly (sponsor bias may limit applicability)	Indirectly (sponsor bias may influence domains of risk of bias, directness, and publication bias)

Abbreviations: GRADE=Grading of Recommendations Assessment, Development and Evaluation; SOE=strength of evidence.

Types of Bias Included in Assessment of Risk of Bias

Numerous, often discipline-specific, taxonomies exist for classifying the different phenomena that introduce bias in studies.[18] For example, although some use the terms confounding and selection bias interchangeably, others see a very clear structural difference between the two and the manner in which they should be handled when detected.[19] What constitutes performance and detection bias in one scheme may be classified under the broader category of information bias in another[1,20] Irrespective of the different classification schemes, the end result identifies associations that are either spurious or related to a variable other than intervention/exposure. We use the taxonomy suggested by Higgins et al. in the Cochrane Handbook as a common, comprehensive, and well-disseminated approach (Table 2).[1] Subsequent sections of this guidance refer to this taxonomy of biases.

Table 2. Taxonomy of core biases in the Cochrane Handbook[1]

Types of bias related to conduct of the study (including analysis and reporting)	Definition	Risk of bias assessment criteria
Selection bias and confounding*	Systematic differences between baseline characteristics of the groups that arise from self-selection of treatments, physician-directed selection of treatments, or association of treatment assignments with demographic, clinical, or social characteristics. Includes Berkson's bias, nonresponse bias, incidence-prevalence bias, volunteer/self-selection bias, healthy worker bias, and confounding by indication/contraindication (when patient prognostic characteristics, such as disease severity or comorbidity, influence both treatment source and outcomes).	Randomization, allocation concealment, sequence generation, control for confounders in cohort studies, and case matching in case-control studies
Performance bias	Systematic differences in the care provided to participants and protocol deviation. Examples include contamination of the control group with the exposure or intervention, unbalanced provision of additional interventions or co-interventions, difference in co-interventions, and inadequate blinding of providers and participants.	Fidelity to protocol, unintended interventions or co-interventions
Attrition bias	Systematic differences in the loss of participants from the study and how they were accounted for in the results (e.g., incomplete follow-up, differential attrition). Those who drop out of the study or who are lost to follow-up may be systematically different from those who remain in the study. Attrition bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations.	Completeness of outcome data, intention-to-treat analysis with appropriate imputations for missing data, and completeness of follow-up
Detection bias	Systematic differences in outcomes assessment among groups being compared, including systematic misclassification of the exposure or intervention, covariates, or outcomes because of variable definitions and timings, diagnostic thresholds, recall from memory, inadequate assessor blinding, and faulty measurement techniques. Erroneous statistical analysis might also affect the validity of effect estimates.	Blinding of outcome assessors, especially with subjective outcome assessments, bias in inferential statistics, valid and reliable measures
Reporting bias	Systematic differences between reported and unreported findings (e.g., differential reporting of outcomes or harms, incomplete reporting of study findings, potential for bias in reporting through source of funding).	Selective outcome reporting evaluation by comparing study report and (a) protocol or (b) outcomes prespecified in methods

*One approach defines selection bias as the bias that occurs when selection is conditioned on common effects of exposures and outcomes and confounding as the bias that occurs when exposure and outcome have a common cause.[19] According to another classification scheme, selection bias is differential selection affected by exposure/intervention in the study, while confounding is differential selection that occurs before exposure and disease.[20]

A brief review of *Cochrane Handbook of Systematic Reviews,*[1] *Systems to Rate the Strength of Scientific Evidence,*[21] and *Evaluation of Non-randomized Studies*[22] shows empirical evidence for detection bias, attrition bias, and reporting bias.

Risk of Bias and Precision

One key distinction between risk of bias and quality assessment is in the treatment of precision. As noted earlier, one definition of quality subsumes freedom from nonsystematic bias

or random error.[4] Tools relying on this definition of quality have included the evaluation of sample size and power to evaluate the impact of random error on the precision of estimates.[23]

Both GRADE[24] and AHRQ guidance on evaluating the strength of evidence[9] separate the evaluation of precision from that of risk of bias. Systematic reviews now routinely evaluate precision (through consideration of the confidence intervals around a summary effect size from pooled estimates) when grading the strength of the body of evidence.[9] Under such circumstances, the evaluation of degree to which studies were designed to allow a precise enough estimate would constitute double-counting limitations to the evidence from a single source. We recommend that AHRQ reviews exclude evaluation of the ability of the study to obtain a precise estimate when assessing the risk of bias for outcomes that can be pooled in meta-analysis or presented quantitatively for single-study bodies of evidence. When outcomes cannot be pooled (as with highly heterogeneous bodies of evidence) or presented quantitatively, assessing the extent to which individual studies are designed to obtain precise estimates in addition to (but separately from) risk of bias may be appropriate.

Risk of Bias and Applicability

Many commonly used quality assessment tools evaluate external validity in addition to internal validity (risk of bias). A review of tools to rate observational studies identified 14 “best” tools. Each evaluated core elements of internal validity and included questions on representativeness of the sample (a component of applicability).[22] Guidance for the EPC program on how to address applicability (also known as external validity, generalizability, or relevance) recommends that EPCs provide a summary report of the applicability of the body of evidence separately from their judgment of the applicability of individual studies.[8] This guidance notes that although individual studies may not be representative of the population of interest, consistent findings across studies with individually limited generalizability may suggest broad applicability of the results.

We recommend that AHRQ reviews generally exclude considerations of applicability in risk-of-bias assessments of individual studies. We note, however, that some study features may be relevant to both risk of bias and applicability. Duration of follow-up is one such example: if duration of followup is different across comparison groups within a study, this difference could be a source of bias; the absolute duration of follow-up for the study would be relevant to the clinical context of interest and therefore the applicability of the study. Likewise study population may be considered within both risk of bias and applicability: if the populations are systematically different between comparison groups within a study (e.g., important baseline imbalances) this may be a source of bias; the population selected for the focus of the study (e.g., inclusion and exclusion criteria) would be a consideration of applicability. Reviewers need to clearly separate study features that may be potential sources of bias from those that are concerned with applicability outside of the individual study context.

Risk of Bias and Poor or Inadequate Reporting

In theory, internal validity focuses on design and conduct of a study. In practice, assessing the internal validity of a study requires adequate reporting of the study, unless additional information is obtained by reaching out to investigators. Although new standards on reporting seek to improve reporting of study design and conduct,[25-29] EPC review teams continue to need a practical approach to dealing with poor or inadequate reporting. The Cochrane risk of bias tool judges the risk of bias to be uncertain when information is inadequate. EPC reviews

have varied in their treatment of reporting of study design and conduct; for example, some have elected to rate poorly *reported* studies as studies with high risk of bias. In general, we recommend that assessment of risk of bias focus primarily on the design and conduct of studies and not on the quality of reporting. EPCs may choose to select an "unclear risk of bias" category for studies with missing or poorly reported information on which to base risk of bias judgments. When studies include meta-analyses, we recommend that quantitative estimates of effect account, through sensitivity analyses, for the impact of including studies with high or unclear risk of bias.

Risk of Bias and Conflict of Interest From Sponsor Bias

Many studies examining the issue of financial conflict of interest have found that sponsor participation in data collection, analysis, and interpretation of findings can threaten the internal validity and applicability of primary studies and systematic reviews.[30,31] The pathways by which sponsor participation can influence the validity of the results are manifold. They include the following:

1. selection of designs and hypotheses—for example, choosing noninferiority rather than superiority approaches,[32] picking comparison drugs and doses,[32] choosing outcomes,[31] or using composite endpoints (e.g., mortality and quality of life) without presenting data on individual endpoints;[33]
2. selective outcome reporting—for example, reporting relative risk reduction rather than absolute risk reduction or "cherry-picking" from multiple endpoints;[32]
3. differences in internal validity of studies and adequacy of reporting;[34]
4. biased presentation of results;[33] and
5. publication bias.[35]

EPCs can evaluate these pathways if and only if the relationship between the sponsor(s) and the author(s) is clearly documented; in some instances, such documentation may not be sufficient to judge the likelihood of conflict of interest (for example, authors may receive speaking fees from a third party that did not support the study in question).

Editors have grown increasingly concerned about the practice of ghost authoring (i.e., primary authors or substantial contributors are not identified) or guest authoring (i.e., one or more identified authors are not substantial contributors)[36] sponsored studies, a practice that makes the actual contribution of the sponsor very difficult to discern.[37,38]

All these concerns may lead to the conclusion that sponsorship from industry (i.e., for-profit entities) should be included as an explicit consideration for assessment of risk of bias. We concur that sponsorship of studies should be considered in critically appraising the evidence but caution against equating industry sponsorship with high risk of bias for three reasons. First, sponsor bias is not limited to industry; nonprofit and government-sponsored studies may also be guest- or ghost-authored. Moreover, the researchers may have various financial or intellectual conflicts of interest by virtue of, for example, accepting speaking fees from many sources.[39] Second, financial conflict is not the only source of conflict of interest: other potential conflicts include personal, professional, or religious beliefs, desire for academic recognition, and so on.[30] Third, the multiple pathways by which sponsorship may influence studies are not all solely within the domain of assessment of risk of bias: several of these pathways fall under the purview of other systematic review tasks. For instance, concerns about the choice of designs, hypotheses, and outcomes relate as much or more to applicability than other aspects of reviews. Reviewers

can and should consider the likely influence of sponsor bias on selective outcome reporting, but when these judgments may be limited by lack of access to full protocols, the assessment of selective outcome reporting may be more easily judged for the body of evidence than for individual studies.

The biased presentation or "spin" on results, although of concern to the lay reader, if limited to the discussion and conclusion section of studies, should have no bearing on systematic review conclusions because systematic reviews do not rely on interpretation of data by study authors.

Internal validity and completeness of reporting constitute, then, the primary pathway by which sponsors may influence the validity of study results that is entirely within the domain of assessment of risk of bias. We acknowledge that this pathway may not be the most important source of sponsor influence: as standards for conduct and reporting of studies become widespread and journals require that they be met, differences in internal validity and reporting between industry-funded studies and other studies will likely attenuate. In balancing these considerations with the primary responsibility of the systematic reviewer—objective and transparent synthesis and reporting of the evidence—we make three recommendations: (1) at a minimum, EPCs should routinely report the source of each study's funding; (2) EPCs should consider issues of selective outcome reporting at the individual study level and for the body of evidence; and (3) EPCs should conduct sensitivity analyses for the body of evidence when they have reason to suspect that the source of funding or disclosed conflict of interest is influencing studies' results.[32] One limitation of relying on sensitivity analyses to demonstrate evidence of risk of bias for industry-funded studies when sponsor bias is suspected (rather than assuming higher risk for industry-funded studies) is that newer studies may appear to be biased when compared to older studies, because of changes in journal reporting standards.

Risk of Bias and Selective Outcome Reporting

Selective outcome reporting refers to the selection of a subset of analyses for publication based on results[40] and has major implications for both the risk of bias of individual studies and the strength of the body of evidence. Comparisons of the full protocol to published or unpublished results can help to flag studies that selectively report outcomes. In the absence of access to full protocols,[9,17] Guyatt et al. note as follows:

> Selective reporting is present if authors acknowledge pre-specified outcomes that they fail to report or report outcomes incompletely such that they cannot be included in a meta-analysis. One should suspect reporting bias if the study report fails to include results for a key outcome that one would expect to see in such a study or if composite outcomes are presented without the individual component outcomes.[17,p 409]

Methods continue to be developed for identifying and judging the risk of bias when results deviate from protocols in the timing or measure of the outcome. No guidance currently exists on how to evaluate the risk of selective outcome reporting in older studies with no published protocols or whether to downgrade all evidence from a study where comparisons between protocols and results show clear evidence of selective outcome reporting for some outcomes.

Even when access to protocols is available, the evaluation of selective outcome reporting may be required again at the level of the body of evidence. Selective outcome reporting across several studies for a body of evidence may result in downgrading the body of evidence.[17]

Previous research has established the link between industry funding and publication bias, a form of reporting bias in which the decision to selectively publish the entire study is based on results.[41] Publication bias may be a pervasive problem in some bodies of evidence and should be evaluated when grading the body of evidence. New research is emerging on selective outcome reporting in industry-funded studies.[42] As methods on identifying and weighing the likely effect of selective outcome reporting continue to be developed, this guidance will also require updating. Our current recommendation is to consider the risk of selective outcome reporting for individual studies and the body of evidence, particularly when a suspicion exists that forces such as sponsor bias may influence the reporting of outcomes.

Risk of Bias and Outcome Measures

The use of valid and reliable outcome measures reduces the likelihood of detection bias. For example, studies relying on self-report measures may be rated as having a higher risk of bias than studies with clinically observed outcomes. In addition, differential assessment of outcome measures by study arm (e.g., electronic medical records for control arm versus questionnaires for intervention arm) constitute a source of measurement bias and should, therefore, be included in assessment of risk of bias. We recommend that assessment risk of bias of individual studies include the evaluation of the validity and reliability of outcome measures, and their variation across study arms.

Recent guidance on the evaluation of applicability by Atkins and colleagues states the importance of considering the relevance of outcome measures for judging applicability (or external validity) of the evidence.[43] For instance, studies that focus on short-term outcomes and fail to report long-term outcomes may be judged as having poor applicability or not being directly relevant to the clinical question for the larger population. The choice of specific outcome measures is a consideration when judging applicability and directness rather than risk of bias; their validity and reliability, on the other hand, is a component of risk of bias, as noted above.

Risk of Bias and Study Design

Some designs possess inherent features (such as randomization and control arms) that reduce the risk of bias and increase the potential for causal inference, particularly when considering benefit of the intervention. Other study designs have specific and inherent risks of biases that cannot be minimized. The clinical question will dictate which study designs are suitable to answer a specific question. EPCs consider these design-specific sources of bias at two points in the systematic review process: (1) when evaluating whether to admit observational studies into the review and (2) when evaluating individual studies for design-specific risks of bias. Norris et al. note that the default strategy in systematic reviews should be to *consider* including observational studies for evidence of benefit and the decision rests on the answer to two questions: (1) are there gaps in the trial evidence for the review questions under consideration? and (2) will observational studies provide valid and useful information to address key questions?[7] In considering whether observational studies provide valid and useful information for benefit, EPCs will need to consider the likelihood that observational studies will generally have more numerous and more serious sources of bias than trials. Once an EPC makes

the decision to include observational studies, then the review team needs to evaluate each study based on the risks of bias specific to that design.

Both AHRQ and GRADE approaches to evaluating the strength of evidence include study design and conduct (risk of bias) of individual studies as components needed to evaluate body of evidence. The inherent limitations present in observational designs (e.g., absence of randomization) are factored in when grading the strength of evidence, EPCs generally give evidence derived from observational studies a low starting grade and evidence from randomized controlled trials a high grade. They can then upgrade or downgrade the observational and randomized evidence based on the strength of evidence domains (i.e., risk of bias of individual studies, directness, consistency, precision, and additional domains if applicable).[9]

Because systematic reviews evaluate design-specific sources of bias in selecting studies for inclusion in the review and then use study design as a component of risk of bias in judging the strength of evidence, we recommend that EPCs do not use study design labels as a proxy for assessment of risk of bias of individual studies. In other words, EPCs should not downgrade the risk of bias of *individual* studies on the basis solely of study design because doing so would penalize studies again (i.e., at the level of individual studies and the body of evidence). This approach accounts for the fact that a study can be performed with the highest quality *for that study design* but still have some (if not serious) potential risk of bias.[1] This approach also acknowledges that quality varies, perhaps widely, within designs and that some study designs do have inherent limitations that can never be fully overcome when considering the validity of their results for benefits. For observational studies, an important consideration is to make a list of possible biases based on the topic and specific design and then evaluate their potential importance for each study.

This approach does not, however, address the fact that no grading system presently accounts for variations in potential risk of bias from different types of observational studies. Under current systems of grading strength of evidence, reviews that consider including observational study designs with highly varying risks of bias (e.g., case reports and data from large registries) for the same clinical question would evaluate all such observational designs together in strength of evidence grades. Under such circumstances, our guidance is to consider the question of value to the review with regard to each study design type: "Will [case reports/case series/case control studies, etc.] provide valid and useful information to address key questions?" Depending on the clinical question, the sources of bias from a particular study design may be so large as to constitute an unacceptably high risk of bias. For instance, EPCs may judge information on benefits from case series of interventions as having a very high risk of bias. In such instances, we recommend that EPCs exclude such designs from the review rather than include the study and then apply a common rating of high risk of bias across all studies with that design without consideration of individual variations in study performance.

In summary, this approach allows EPCs to deal with variations in included studies by study design, for instance by rating outcomes for benefit from individual randomized controlled trials (RCTs), or observational studies, as low, medium, high, or unclear risk of bias. It then defers the issue of study design limitations to assessment of the strength of evidence.

Risk of Bias and Fidelity to the Intervention Protocol

Failure of the study to maintain fidelity to the intervention protocol can influence performance bias; it is, therefore, a component of assessment of risk of bias. We note, however, that the interpretation of fidelity may differ by clinical topic. For instance, some behavioral

interventions include "fluid" interventions; these involve interventions for which the protocol explicitly allows for modification based on patient needs; such fluidity does not mean the interventions are implemented incorrectly. When interventions implement protocols that have minimal concordance with practice, the discrepancy may be considered an issue of applicability. This lack of concordance with practice does not, however, constitute risk of bias. We also note that when studies implement an intervention with previously established efficacy in varied settings but are unwilling or unable to maintain fidelity to the original intervention protocol, this deviation may influence the risk of bias of the study and the applicability of the intervention overall. We recommend that EPCs account for the specific clinical considerations in determining and applying criteria about fidelity for assessment of risk of bias. Our recommendation is consistent with the Institute of Medicine guidelines on systematic reviews.[44]

Stages in Assessing the Risk of Bias of Studies

International reporting standards require documentation of various stages in a comparative effectiveness review.[45-47] We lay out recommended approaches to assessment of risk of bias in five steps: protocol development, pilot testing and training, assessment of risk of bias, interpretation, and reporting. Table 3 describes the stages and specific steps in assessing the risk of bias of individual studies that contribute to transparency through careful documentation of decisions.

Table 3. Stages in assessing the risk of bias of individual studies

Stages in risk-of-bias assessment	Specific steps
1. Develop protocol	• Specify terms (i.e., quality assessment or risk of bias) and included concepts • Explain the inclusion of specific risk-of-bias criteria • Select and justify choice of specific risk-of-bias rating tool(s) • Include tools for assessment of risk of bias that justify research-specific risk-of-bias standards and operational definitions of risk-of-bias criteria • Explain how individual risk-of-bias criteria will be summarized to obtain low, moderate, high, or unclear risk of bias for individual outcomes and justify any use of scales (numerical scores leading to categories of risk of bias) • Explain how inconsistencies between pairs of risk of bias reviewers will be resolved • Explain how the synthesis of the evidence will incorporate assessment of risk of bias (including whether studies with high or unclear risk of bias will be used in synthesis of the evidence)
2. Pilot test and train	• Determine composition of the review team. A minimum of two reviewers must rate the risk of bias of each study, with a third reviewer to serve as arbiter of conflicts • Train reviewers • Pilot test assessment of risk of bias tools using a small subset of studies that represent the range of risk of bias in the evidence base • Identify issues and revise tools or training as needed
3. Perform assessment of risk of bias of individual studies	• Determine study design of each (individual) study • Make judgments about each risk of bias criterion, using the preselected appropriate criteria for that study design and for each predetermined outcome • Make judgments about overall risk of bias for each included outcome of the individual study, considering study conduct, and categorize as low, moderate, high, or unknown risk of bias within study design; document the reasons for judgment and process for finalizing judgment • Resolve differences in judgment and record final rating for each outcome

Table 3. Stages in assessing the risk of bias of individual studies (continued)

Stages in risk-of-bias assessment	Specific steps
4. Use assessment of risk of bias in synthesis of evidence	• Conduct preplanned analyses • Consider additional required analyses • Incorporate assessment of risk of bias in quantitative/qualitative synthesis, keeping study design categories separate
5. Report assessment of risk of bias process and limitations	• Cite reports on validation of the selected tool(s), the assessment of risk of bias process (summarizing from the protocol), and limitations to the process • Describe actions to improve assessment of risk-of-bias reliability if applicable

The plan for assessment of risk of bias should be included within the protocol for the entire review. As prerequisites to developing the plan for assessment of risk of bias, EPCs must identify the important intermediate and final outcomes that need assessment of risk of bias and other study descriptors or study data elements that are required for the assessment of risk of bias in the systematic review protocol. Protocols must justify what risk-of-bias criteria will be evaluated and how the reviewers will incorporate risk of bias of individual studies in the synthesis of evidence.

The assessment must include a minimum of two reviewers per study with a third to serve as arbitrator. EPCs should anticipate having to review and revise assessment of risk of bias forms and instructions in response to problems arising in training and pilot testing.

Assessment of risk of bias should be consistent with the analysis plans in registered protocols of the reviews. Published reviews must include risk-of-bias criteria and should describe the selected tools and their reliability and validity when such information is available. EPC reviews should report all criteria used for each evaluated outcome. The synthesis of the evidence should reflect the *a priori* analytic plan for incorporating risk of bias of individual studies in qualitative or quantitative analyses. EPCs should report the outcomes of all preplanned analyses that included risk-of-bias criteria regardless of statistical significance or the direction of the effect. Published reviews should also include justifications of all *post hoc* decisions to limit synthesis of included studies to a subset with common methodological or reporting attributes.

Design-Specific Criteria To Assess Risk of Bias

We present design-specific criteria to assess risk of bias for five common study designs: RCTs, cohort (prospective, retrospective, and nonconcurrent), case-control (including nested case-control), case series, and cross-sectional (Table 4).[48] Table 4 draws on other instruments,[1,49] was modified based on workgroup consensus and peer review, and is not intended to serve as a one-size-fits-all instrument. Rather, it is intended to remind reviewers of common sources of bias for some common types of study designs. A critical task that reviewers need to incorporate within each review is the careful identification and recording of likely sources of bias for each topic and each included design. Reviewers may select specific criteria or combinations of criteria relevant to the topic. For instance, blinding of outcome assessors may not be possible for surgical interventions but the inability to blind outcome assessors does not obviate the risk of bias from lack of blinding. Reviewers should be alert to the use of self-reported or subjective outcome measures or poor controls for differential treatment in such studies that could elevate the risk of bias further.[1,50]

Table 4. Design-specific criteria to assess for risk of bias for benefits

Risk of bias	Criterion	RCTs	CCTs or cohort	Case-control	Case series	Cross-sectional
Selection bias	Was the allocation sequence generated adequately (e.g., random number table, computer-generated randomization)?	x				
	Was the allocation of treatment adequately concealed (e.g., pharmacy- controlled randomization or use of sequentially numbered sealed envelopes)?	x				
	Were participants analyzed within the groups they were originally assigned to?	x	x			
	Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?		x			x
	Were cases and controls selected appropriately (e.g., appropriate diagnostic criteria or definitions, equal application of exclusion criteria to case and controls, sampling not influenced by exposure status)			x		
	Did the strategy for recruiting participants into the study differ across study groups?		x			
	Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?	x	x	x*	x	x
Performance bias	Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?	x	x	x	x	x
	Did the study maintain fidelity to the intervention protocol?	x	x	x	x	
Attrition bias	If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?	x	x	x	x	x
Detection bias	In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome the same for cases and controls?	x	x	x		
	Were the outcome assessors blinded to the intervention or exposure status of participants?	x	x	x	x	x
	Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?		x	x	x	x
Reporting bias	Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?	x	x	x	x	x

*Cases and controls should be similar in all factors known to be associated with the disease of interest, but they should not be so uniform as to be matched for the exposure of interest.

Another example of a criterion that requires topic-specific evaluation is prespecification of outcomes. Depending on the topic, prespecification of outcomes is entirely appropriate and expected, regardless of study design. For other topics, data from observational studies may offer the first opportunity to identify unexpected outcomes that may need confirmation from RCTs. For review topics in search of evidence on rare long-term outcomes, requiring prespecification would be inappropriate. A third example of a criterion requiring topic-specific evaluation is the expected attrition rate. Differential or overall attrition because of nonresponse, dropping out, loss to follow-up, and exclusion of participants can introduce bias when missing outcome data are related to both exposure/treatment and outcome. Reviewers of topics that focus on short-term clinical outcomes may select a low expected attrition rate. We also note that with attrition rate in particular, no empirical standard exists across all topics for demarcating a high risk of bias from a lower risk of bias; these standards are often set within clinical topics. The list of recommended criteria does not represent comprehensive sources of bias for other study designs. For instance, case series studies with repeated time measures may require a question asking whether the study accounted for regression to the mean. Some concepts included in Table 4, particularly intention-to-treat, have been interpreted in a variety of ways. The *Cochrane Handbook of Systematic Reviews* offers a more detailed treatment of intention to treat.[1]

Tools for Assessing Risk of Bias

EPCs can use one of two general approaches to assessing risk of bias in systematic reviews. One method is often referred to as a *components approach*. This involves assessing individual items that are deemed by the systematic reviewers to reflect the methodological risk of bias, or other relevant considerations, in the body of literature under study. For example, one commonly assessed component in RCTs is allocation concealment.[51] Reviewers assess whether the randomization sequence was concealed from key personnel and participants involved in a study before randomization; they then rate the component as adequate, inadequate, or unclear. The rating for each component is reported separately. The second common approach is to use a *composite approach* that combines different components related to risk of bias or reporting into a single overall score.

Many tools have emerged over the past 20 years to assess risk of bias. Some tools are specific to different study designs, whereas others can be used across a range of designs. Some have been developed to reflect nuances specific to a clinical area or field of research. Because many AHRQ systematic reviews typically address multiple research questions, they may require the use of several risk of bias tools or the selection of various different components to address all the study designs included.

- Currently there is no consensus on the best approach or preferred tool for assessing risk of bias, because the components associated with risk of bias are in contention. As such, there are a large number of tools available, and their marked variations and relative merits can be problematic for systematic reviewers. We advocate the following general principles when selecting a tool, or approach, to assessing risk of bias in systematic reviews. EPCs should opt for tools that:were specifically designed for use in systematic reviews;
- have demonstrated acceptable validity and reliability, or show transparency in how assessments are made by providing explicit support for each assessment;
- specifically address items related to risk of bias (internal validity), and preferably are based on empirical evidence of bias;

- where available, are specific to the study designs being evaluated; and
- avoid the presentation of risk-of-bias assessment as a composite score, that is, an overall numeric rating of study risk of bias across items, for example 11 from 15 items.

Although there is much overlap across different tools, there is no single universal tool that addresses all the varied contexts for assessment of risk of bias. Appendix A details a select list of tools that have been shown to be reliable or valid, are widely used, or have been recommended for use in systematic reviews that compared risk-of-bias assessment instruments.[21,22,52-54] We do not discuss tools that have been developed to guide and assess the reporting of studies. These reporting guidelines assess different constructs than what is commonly understood as risk of bias (internal validity). A list of reporting guidelines for different study designs is available through the EQUATOR network at www.equator-network.org.

Assessing the Risk of Bias for Harms

Although the assessment of harms is almost always included as an outcome in intervention studies, the manner of capturing and reporting harms is significantly different than the outcomes of benefit. Harms are defined as the "totality of possible adverse consequences of any intervention, therapy or medical test; they are the direct opposite of benefits, against which they must be compared."[55] For a detailed explanation of terms associated with harms please refer to the AHRQ Methods guide on harms.[56] Systematic reviews of intervention studies need to consider the balance between the harms and benefits of the treatment. Empirical evidence across diverse medical fields indicates that reporting of safety information—including milder harms—receives much less attention than the positive efficacy outcomes.[57,58] Thus, an evaluation of the benefits alone is likely to bias conclusions about the net efficacy or effectiveness of the intervention. Although reviewers recognize the importance of harms outcomes, harms are generally ignored in risk-of-bias assessment checklists. Several recent reviews[21,52-54] of risk-of-bias checklists and instruments do not identify harms as a key criterion within the checklists. We infer that many of the current risk-of-bias scales and checklists have assumed that harms are simply another study "outcome" and that taking this view suggests that the developers assume that no differences exist between harms and benefits in terms of risk-of-bias assessment.

For some aspects of risk-of-bias assessment, this approach may be reasonable. For example, consider an RCT evaluating the outcomes of a new drug therapy relative to those of a placebo control group; improper randomization would increase the risk of bias for measuring both outcomes of benefit and harm. However, unlike outcomes of benefit, harms and other unintended events are unpredictable and methods or instruments used to capture all possible adverse events can be problematic. This implies that that there is a potential for risk of bias for harms outcomes that is distinct from biases applicable to outcomes of benefit.

Because the type, timing, and severity of some harms are not anticipated—especially for rare events—many studies do not specify exact protocols to actively capture events. Standardized instruments used to systematically collect information on harms are often not included in the study methods. Study investigators may assume that patients will know when an adverse event has occurred, accurately recall the details of the event, and then "spontaneously" report this at the next outcome assessment. Thus, harms are often measured using passive methods that are poorly detailed, resulting in potential for selective outcome reporting, misclassification, and failure to capture significant events. Although some types of harms can be

anticipated (e.g., pharmacokinetics of a drug intervention may identify body systems likely to be affected) that include both common (e.g., headache) and rare conditions (e.g., stroke), harms may also occur in body systems that are not necessarily linked to the intervention from a biologic or epidemiologic perspective. In such instances, an important issue is establishing an association between the event and the intervention. The primary study may have established a separate committee to evaluate association between the harm and the putative treatment; as such blinding is not possible in such evaluations. Similarly, evaluating the potential for selective outcome reporting bias is complex when considering harms; some events may be unpredictable or they occur so infrequently relative to other milder effects that they are not typically reported. Given the possible or even probable unevenness in evaluating harms and benefits in most intervention studies, we recommend that EPCs assess the risk of bias of the study separately for benefits and for harms (see Appendix A for suggested tools and approaches).

Summarizing the Risk of Bias of a Study

For any outcomes undergoing assessment of strength of evidence, reviewers must consider all of the items together after completing evaluations of the assessment of risk of bias items for a given study. Then reviewers place risk of bias in a given study for each outcome into a summary category: low, medium or high.[9] Reviewers may conclude unclear risk of bias from poorly reported studies. This section describes methods for achieving that categorization and discusses guidelines for reporting this information. A study's risk of bias category can be different for different outcomes, which means that review teams should record the different outcome-specific categories as necessary. This situation can arise from, for instance, variation in the completeness of data, differential blinding of outcome assessors, or other outcome-specific items. Summarizing risk of bias for each patient-centered outcome within a study is recommended for synthesis of evidence across the studies and evaluating strength of evidence.[1] We do not recommend summarizing risk of bias across several outcomes for a given study because such global assessments across outcomes would involve subjective author judgments about relative importance of patient-centered outcomes and other factors for decision making.

Categories for Outcome-Specific Risk of Bias

An overall rating of low, medium, high, or unclear risk of bias should be made for the most clinically important outcomes as defined in the review protocol. As is true for scoring individual criteria or items, EPCs should do this overall rating within the study design. Observational studies and RCTs should be evaluated separately using recommended domains (Table 4). EPCs should adopt a dual reviewer approach to this step as well. Finally, given that these assessments involve subjective considerations, reviewers must clearly describe their rationale and explicit definitions for all ratings.

A study categorized as low risk of bias implies confidence on the part of the reviewer that results represent the true treatment effects (study results are considered valid). The study reporting is adequate to judge that no major or minor sources of bias are likely to influence results. A study rated as medium risk of bias implies some confidence that the results represent true treatment effect. The study is susceptible to some bias but the problems are not sufficient to invalidate the results (i.e., no flaw is likely to cause major bias).[59] A study categorized as h*igh* risk of bias implies low confidence that results represent true treatment effect. The study has significant flaws that imply biases of various types that may invalidate its results; these may arise *from serious errors in conduct, analysis, or reporting, large amounts of missing information, or*

discrepancies in reporting. A study categorized as "unclear" risk of bias is missing information, making it difficult to assess limitations and potential problems.

Methods and Considerations for Summarizing Risk of Bias

Some outcomes within a systematic review will receive ratings of the strength of evidence. One core component of the strength of a body of evidence for a given outcome is the overall risk of bias of the outcome data in all studies reporting that outcome.[9] This overall risk of bias is dictated by the risk of bias of the individual studies.

Incomplete reporting is an unavoidable challenge in summarizing the risk of bias of individual studies. To categorize the study, the reviewer must simultaneously consider (1) the known strengths, (2) the known weaknesses, and (3) the unknown attributes. A preponderance of unknown attributes may result in the study being categorized as unclear risk of bias; this might occur, for example, when EPC reviewers cannot determine whether the study was prospective or when investigators did not report the proportion of enrollees who provided data. In some cases, however, the unknown attributes are relatively minor; in these cases, EPC reviewers might still deem them of low risk of bias.

One way to assign a category is to make a simple "holistic" judgment; that is, a judgment based on an overall perception of risk of bias rather than an evaluation of all components of bias. Unfortunately, this approach is not transparent and is likely not to be reproducible. The main problem is inconsistent bases for judgment: if the studies were reexamined, the same reviewer might alter the category assignments. Reviewers may also be influenced, consciously or unconsciously, by other unstated aspects of the studies, such as the prestige of the journal or the identity of the authors. EPCs can and should explain how their reviewers made these judgments, but the fact remains that these approaches can suffer from substantial subjectivity. This transparency in terms of providing explicit support for each of the judgments or assessments made is a key feature of the Risk of Bias tool developed by The Cochrane Collaboration. Detailed and explicit support for each assessment not only ensures complete transparency, but allows the reader to (re)evaluate each assessment.

Instead, we recommend that, in aiming for transparency and reproducibility, EPC reviewers use a set of specific rules for assigning a category. These rules can take the form of declarative statements. For instance, in reviews of topics requiring randomization and blinding, one may make a declarative statement such as "adequately randomized and blinded studies are good; adequately randomized but unblinded studies are fair; inadequately randomized and unblinded studies are poor." EPCs could also lay out more complicated rules that reflect the items in the chosen instrument, but the key is transparency. Obviously, many other items could be incorporated into these rules, but, again, the key is transparency. Notice that such declarative statements implicitly assign weights to the different items. In any case, the authors must justify how synthesis of evidence incorporated risk-of-bias criteria or overall rank of risk of bias.

Within rule-based assignment, one option is to use the domains of risk of bias and then the items within those domains as a basis for the rules. For example, studies that met the majority of the items for all domains are good; studies that met the majority of the items for some (previously specified number) of the domains are fair; all other studies are poor. This process relies on an accurate assignment of items into domains. The basic requirement is adequate explanation of the method used.

The use of a quantitative scale is another way to employ a transparent set of rules. For a scale, the weights of different items are explicit rather than implicit. But any weighting system,

whether qualitative or quantitative, must be recognized as subjective and arbitrary, and different reviewers may choose to use different weighting methods. Using transparent rules does not remove the subjectivity inherent in assigning the risk of bias category. Subjectivity remains in the choice of different rules, or rules that assigning items to domains, and if the latter, what proportion of items must be met to earn a given rating. Consequently, reviewers should avoid attributing unwarranted precision (such as a score of 3.42) to ratings or creating subcategories or ambiguous language such as "in the middle of the fair range."

The approaches outlined above reveal two competing concerns: being transparent, and not being too formulaic. Transparency is important so that users can understand how categories were assigned, and also have some assurance that the same process was used for all of the studies. There is a danger, however, in being too formulaic and insensitive to the specific clinical context of the review. For example, if an outcome is unaffected by blinding, then the unconsidered use of a blinding "rule" (e.g., studies must be blinded to be categorized as low risk of bias) would be inappropriate for that outcome. Thus, we recommend careful consideration of the clinical context as reviewers strive for good transparency.

Previous research has demonstrated that empirical evidence of bias differed across individual domains rather than overall risk of bias.[60] Meta-epidemiological studies have demonstrated that treatment effects did not differ across overall categories of high versus low-risk of bias but did differ by criteria such as masking of treatment status or valid statistical methods.[60-62] Reviewers may use meta-analyses to the association between risk of bias domains and treatment effect with subgroup analyses or meta-regression.[61-63]

Conclusion

Assessment of risk of bias is a key step in conducting systematic reviews that informs many other steps and decisions made within the review. It also plays an important role in the final assessment of the strength of the evidence. The centrality of assessment of risk of bias to the entire systematic review task requires that assessment processes be based on sound empirical evidence when possible or on theoretical principles. In assessing the risk of bias of studies, EPCs should specify constructs and risks of bias specific to the content area, use at least two independent reviewers with a defined process for consensus and standards for transparency, and clearly document and justify all processes and decisions.

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Kathleen N. Lohr, Ph.D.; Mark Helfand, M.D., M.P.H.; Jeffrey C. Andrews, M.D.; and Loraine Monroe, EPC Publications Specialist. We also wish to acknowledge the thoughtful contributions of Susan Norris, M.D., M.Sc., M.P.H., our Associate Editor.

Author Affiliations

RTI International–University of North Carolina at Chapel Hill Evidence-based Practice Center, Research Triangle Park, NC (MV, NDB). University of Ottawa Evidence-based Practice Center, Ottawa, Ontario, Canada (MTA, KS, AT). Agency for Healthcare Research and Quality, Rockville, MD (SC). University of Alberta Evidence-based Practice Center, Edmonton, Alberta, Canada (LH). Vanderbilt University Evidence-based Practice Center, Nashville, TN (MM). McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada (PLS).

Minnesota University Evidence-based Practice Center, Minneapolis, MN (TS). ECRI Institute Evidence-based Practice Center, Plymouth Meeting, PA (TRD).

References

1. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. In: Higgins JPT, Green S, eds. The Cochrane Collaboration; 2011.
2. Cochrane Collaboration Glossary Version 4.2.5. 2005. Available at: http://www.cochrane.org/sites/default/files/uploads/glossary.pdf. http://effectivehealthcare.ahrq.gov/. Accessed January 2011.
3. Juni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. In: Egger M, Davey SG, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. 2001/07/07 ed. London: BMJ Books; 2001. p. 87–108.
4. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 2004;16(1):9–18. PMID: 15020556.
5. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011 Apr;64(4):401–6. PMID: 21208779.
6. U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF. Available at: http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed July 2008.
7. Norris SL, Atkins D, Bruening W, et al. Observational studies in systemic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2011 Nov;64(11):1178–86. PMID: 21636246.
8. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004 Jun 19;328(7454):1490. PMID: 15205295.
9. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol 2010;63(5):513–23.
10. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011 Apr;64(4):395–400. PMID: 21194891.
11. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence-imprecision. J Clin Epidemiol 2011 Dec;64(12):1283–93. PMID: 21839614.
12. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol 2011 Dec;64(12):1303–10. PMID: 21802903.
13. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. J Clin Epidemiol 2011 Dec;64(12):1294–302. PMID: 21803546.
14. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence-publication bias. J Clin Epidemiol 2011 Dec;64(12):1277–82. PMID: 21802904.
15. Guyatt GH, Oxman AD, Schunemann HJ, et al. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011 Apr;64(4):380–2. PMID: 21185693.
16. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011 Dec;64(12):1311–6. PMID: 21802902.
17. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011 Apr;64(4):407–15. PMID: 21247734.
18. Delgado-Rodriguez M, Llorca J. Bias. J Epidemiol Community Health 2004 Aug;58(8):635–41. PMID: 15252064.
19. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004 Sep;15(5):615–25. PMID: 15308962.
20. Validity in Epidemiologic Studies. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008. p. 418–55, 9129–47.
21. West SL, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality; 2002.

22. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii–x, 1–173. PMID: 14499048.
23. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston: Houghton Mifflin Company; 1979.
24. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924–6. PMID: 18436948.
25. Little J, Higgins JP, Ioannidis JP, et al. Strengthening the reporting of genetic association studies (STREGA): an extension of the strengthening the reporting of observational studies in epidemiology (STROBE) statement. J Clin Epidemiol 2009 Jun;62(6):597–608 e4. PMID: 19217256.
26. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001 Apr 14;357(9263):1191–4. PMID: 11323066.
27. Knottnerus A, Tugwell P. STROBE—a checklist to Strengthen the Reporting of Observational Studies in Epidemiology. J Clin Epidemiol 2008 Apr;61(4):323. PMID: 18313555.
28. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003 Nov;56(11):1118–28. PMID: 14615003.
29. Davidoff F, Batalden P, Stevens D, et al. Publication guidelines for improvement studies in health care: evolution of the SQUIRE Project. Ann Intern Med 2008 Nov 4;149(9):670–6. PMID: 18981488.
30. Bekelman JE, Li Y, Gross CP. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. JAMA 2003 Jan 22-29;289(4):454–65. PMID: 12533125.
31. Newcastle-Ottawa Quality Assessment Scale: Cohort studies. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed January 2011.
32. Smith R. Medical journals are an extension of the marketing arm of pharmaceutical companies. PLoS Med 2005 May;2(5):e138. PMID: 15916457.
33. Julian DG. What is right and what is wrong about evidence-based medicine? J Cardiovasc Electrophysiol 2003 Sep;14(9 Suppl):S2–S5. PMID: 12950509.
34. Jorgensen AW, Maric KL, Tendal B, et al. Industry-supported meta-analyses compared with meta-analyses with non-profit or no support: differences in methodological quality and conclusions. BMC Med Res Methodol 2008;8:60. PMID: 18782430.
35. Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med 2008 Sep 23;5(9):e191. PMID: 18816163.
36. American Medical Writers Association. AMWA ethics FAQs, publication practices of particular concern to medical communicators. 2009. Available at: http://www.amwa.org/default.asp?Mode=DirectoryDisplay&DirectoryUseAbsoluteOnSearch=True&id=466. Accessed June 2, 2011.
37. Ross JS, Hill KP, Egilman DS, et al. Guest authorship and ghostwriting in publications related to rofecoxib: a case study of industry documents from rofecoxib litigation. JAMA 2008 Apr 16;299(15):1800–12. PMID: 18413874.
38. DeAngelis CD, Fontanarosa PB. Impugning the integrity of medical science: the adverse effects of industry influence. JAMA 2008 Apr 16;299(15):1833–5. PMID: 18413880.
39. Hirsch LJ. Conflicts of interest, authorship, and disclosures in industry-related scientific publications: the tort bar and editorial oversight of medical journals. Mayo Clin Proc 2009 Sep;84(9):811–21. PMID: 19720779.
40. Kirkham JJ, Dwan KM, Altman DG, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010;340:c365. PMID: 20156912.
41. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990 Mar 9;263(10):1385–9. PMID: 2406472.
42. Vedula SS, Bero L, Scherer RW, et al. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. N Engl J Med 2009 Nov 12;361(20):1963–71. PMID: 19907043.
43. Atkins D, Chang S, Gartlehner G, et al. Assessing the Applicability of Studies When Comparing Medical Interventions. Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC019-EF. Available at: http://effectivehealthcare.ahrq.gov/. Accessed January 2011.

44. Institute of Medicine. Finding what works in health care: standards for systematic reviews. Available at: http://www.nap.edu/openbook.php?record_id=13059&page=R1. Accessed June 2, 2011.
45. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009 Oct;62(10):1013–20. PMID: 19230606.
46. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol 2009 Oct;62(10):1006–12. PMID: 19631508.
47. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700. PMID: 19622552.
48. Hartling L, Bond K, Harvey K, et al. Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures. Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023. Rockville, MD: Agency for Healthcare Research and Quality: June 2009. AHRQ Publication No. 11-EHC007-EF.
49. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol 2011 Sep 28;PMID: 21959223.
50. Egger M, Smith DH. Under the meta-scope: potentiatls and limitations of meta-analysis. Evidence based resource in anaesthesia and analgesia. BMJ Publication 2000.
51. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Obstet Gynecol 2010 May;115(5):1063–70. PMID: 20410783.
52. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008 Feb;88(2):156–75. PMID: 18073267.
53. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36(3):677–8. PMID: 17470488.
54. Whiting P, Rutjes AW, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005 Jan;58(1):1–12. PMID: 15649665.
55. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004 Nov 16;141(10):781–8. PMID: 15545678.
56. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010 May;63(5):502–12. PMID: 18823754.
57. Ioannidis JP, Lau J. Improving safety reporting from randomised trials. Drug Saf 2002;25(2):77–84. PMID: 11888350.
58. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001 Jan 24–31;285(4):437–43. PMID: 11242428.
59. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001 Apr;20(3 Suppl):21–35. PMID: 11306229.
60. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002 Jun 12;287(22):2973–82. PMID: 12052127.
61. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557–60. PMID: 12958120.
62. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006 Dec;59(12):1249–56. PMID: 17098567.
63. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2011 Nov;64(11):1187–97. PMID: 21477993.

Chapter 5 Appendix A. Tools To Assess Risk of Bias of Individual Outcomes

This appendix provides a brief overview of tools to evaluate randomized controlled trials (RCTs), nonrandomized studies, and harms. This information does not represent a comprehensive systematic synthesis of tools available but provides details for a select list of tools that have been shown to be reliable or valid, are widely used, or have been recommended for use in systematic reviews that compared risk of bias assessment instruments.[1-5] For most tools, the preliminary step in assessing whether a chosen tool is applicable to the specific study is to categorize the study design. We recommend the use of tools such as that developed by Hartling et al. to categorize study designs.[6]

Randomized Controlled Trials

A large number of tools have been developed to assess risk of bias in RCTs. In 2008, Armijo Olivo et al.[1] published a systematic review identifying scales designed to assess the risk of bias of RCTs. They identified 21 scales but found that the majority were not "rigorously developed or tested for validity and reliability."

Armijo Olivo et al. found that the Jadad scale demonstrated the strongest evidence in terms of validity and reliability. The Jadad scale demonstrates face, content, criterion, and construct validity. One limitation regarding the assessment of criterion or concurrent validity for all risk of bias tools is that it depends on a gold standard that does not exist for these tools. Hence, reports of construct validity need to be interpreted in light of the tool used as the reference standard for comparisons. Armijo Olivo et al. found that the Jadad scale was most commonly cited in the medical literature. The Jadad scale was the most commonly used tool in systematic reviews produced by The Cochrane Collaboration until recently, and it is still the most commonly used tool to assess risk of bias of RCTs in AHRQ evidence reports. The Jadad scale addresses three domains (i.e., randomization, blinding, and handling of withdrawals and drop-outs), but does not address adequacy of allocation concealment. The tool includes five questions which take approximately 10 minutes to apply to an individual trial. Although the Jadad scale was developed in the context of pain research it has been tested and used widely in other fields. Although the Jadad scale is the most commonly used tool to assess risk of bias of RCTs, concerns regarding its appropriateness have recently emerged.[7-9] Specifically, there is some evidence that the tool reflects quality of reporting rather than risk of bias.[10]

Armijo Olivo et al. highlighted two other tools that were developed using rigorous methods and tested for validity and reliability. Verhagen et al. developed the Delphi List to assess RCTs in general (i.e., not specific to a clinical area or field of study). It has demonstrated good face, content, and concurrent validity and has been tested for reliability. It includes the following items: inclusion/exclusion criteria of study population defined; randomization; allocation concealment; baseline comparability of study groups; blinding of investigator, subjects, and care providers; reporting of point estimates and variability for primary outcomes; and intention-to-treat analysis.[11]

Yates et al. developed a tool to assess the risk of bias of RCTs of cognitive behavioral therapy for chronic pain. The tool has two parts, one related to the treatment (five items) and the second related to study design and methods (eight items with multiple parts). The latter part of the tool includes questions on the following domains: reporting of inclusion/exclusion criteria;

reporting of attrition; adequate description of the sample; steps to minimize bias (i.e., randomization, allocation, measurement, treatment expectations); outcomes justified, valid, and reliable; length of follow-up (i.e., sustainability of treatment effects); adequacy of statistical analyses; comparability or adequacy of control group. It has shown face, content, and construct validity and good inter-rater reliability.[12] The tool has not been widely used.

In 2005, The Cochrane Collaboration convened a group to address several concerns in the assessment of trial risk of bias. One concern was the growing number of tools being used and inconsistent approaches to risk of bias assessment across different systematic reviews. Participants also recognized that many of the tools being used were not based on empirical evidence showing that the items they included were related to biased results. Moreover, many tools combined elements examining methodological conduct with items related to reporting.

From this work a new tool for randomized trials emerged—the Risk of Bias tool.[6] This tool was released after publication of the review by Armijo Olivo et al. described above. The Risk of Bias tool includes seven domains for which empirical evidence demonstrates associations with biased estimates of effect. The domains are sequence generation; allocation concealment; blinding of participants and personnel; blinding of outcome assessment; missing outcome data; selective outcome reporting; and other sources of bias. The final domain, "other sources of bias," includes design specific risks of bias, baseline imbalance, blocked randomization in unblinded trials, differential diagnostic activity, and other potential biases.[13] The Cochrane Handbook[13] provides guidance on assessing the different domains including "other sources of bias." The Handbook emphasizes that topics within the other domain should focus on issues related to bias and not imprecision, heterogeneity, or other quality measures that are unrelated to bias. Further, these items will vary across different reviews and should be identified and prespecified when developing the review protocol.

Although the Risk of Bias tool is now the recommended method for assessing risk of bias of RCTs in systematic reviews conducted through The Cochrane Collaboration, the tool has not undergone extensive validity or reliability testing. However, one of the unique and critical features of the Risk of Bias tool is its transparency. That is, users are instructed to document explicit support for each assessment alongside the assessment. The developers of the tool argue that this transparency is more important than demonstrations of "reliability" and "validity," because complete transparency is ensured and each assessment can readily be (re)evaluated by the reader.

Nonrandomized Studies

Several systematic reviews have been conducted to identify, assess, and make recommendations regarding risk of bias assessment tools for use in nonrandomized studies (including nonrandomized experimental studies and observational studies). West et al.[5] identified 12 tools for use in observational studies and recommended 6 of these for use in systematic reviews. Deeks et al.[4] identified 14 "best tools" from among 182 and recommended 6 for use in reviews. Of interest is that the two reports identified only three tools in common: Downs and Black,[14] Reisch,[15] and Zaza.[16] These three tools are applicable to a range of study designs; only two were developed for use in systematic reviews.[14,16]

One recent and comprehensive systematic review of risk of bias assessment tools for observational studies identified 86 tools.[2] The tools varied in their development and their purpose: only 15 percent were developed specifically for use in systematic reviews; 36 percent were developed for general critical appraisal and 34 percent were developed for "single use in a

specific context." The authors chose not to make recommendations regarding which specific tools to use; however, they broadly advised that reviewers select tools that

- contain a small number of components or domains;
- are as specific as possible with regard to study design and the topic under study;
- are developed using rigorous methods, evidence-based, and valid and reliable; and
- are simple checklists rather than scales when possible.

The Cochrane Collaboration provides recommendations on use of tools for nonrandomized studies. They acknowledge the abundance of tools available but, like Sanderson et al., make no recommendation regarding a single instrument.[2] They recommend following the domains in the Risk of Bias tool, particularly for prospective studies. A working group within the Cochrane Collaboration is currently modifying the Risk of Bias tool for use in nonrandomized studies.

The Cochrane Handbook highlights two other tools for use in nonrandomized studies: the Downs and Black[14] and Newcastle Ottawa Scale.[17] They implicitly recommend the Newcastle Ottawa Scale over the Downs and Black because the Downs and Black is time-consuming to apply, requires considerable epidemiology expertise, and has been found difficult to apply to case-control studies.[17]

The Newcastle Ottawa Scale is frequently used in systematic reviews for articles about studies with this type of design. It contains separate questions for cohort and case-control studies. It was developed based on threats to validity in nonrandomized studies; these specifically include selection of participants (generalizability or applicability), comparability of study groups, methods for outcome assessment (cohort studies) or ascertainment of exposure (case-control studies), and adequacy of follow-up. The developers have reported face and content validity for this instrument, and they revised it based on experience using the tool in systematic reviews.[17] It has also been tested for inter-rater reliability.[18,19] Examination of its criterion validity and intra-rater reliability is underway and plans are being developed to examine its construct validity.

Other recently developed checklists address the quality of observational, nontherapeutic studies of incidence of diseases or risk factors for chronic diseases[20] or observational studies of interventions or exposures.[21] The checklists have been developed based on a comprehensive literature review,[22] are based on predefined flaws in internal validity, and discriminate reporting from conduct of the studies. These tools are continuing inter-rater reliability tests.

Instruments and Tools To Evaluate Quality of Harms Assessment

No systematic reviews evaluating tools to assess the potential for biases associated with harms were found. However, three tools/checklists were identified and two of these recognize that some biases may arise when capturing and reporting harms that are distinct from the outcomes of benefit and therefore require separate assessment.

One checklist developed by the Cochrane Collaboration offers some guidance, and leaves the final choice up to the reviewer to select items from a list that is stratified by the study design.[13] It assumes that these questions (see Table A-1) can be added to those criteria already detailed in the Cochrane Risk of Bias tool.

Table A-1. Recommendations for elements of assessing quality of the evidence when collecting and reporting harms, by study design

Study design	Quality considerations
RCTs	On study conduct: • Are definitions of reported adverse effects given? • Were the methods used for monitoring adverse effects reported, such as use of prospective or routine monitoring; spontaneous reporting; patient checklist, questionnaire or diary; systematic survey of patients? What was the source to assess harms (self-report vs. medical exam vs. PI opinion)? Who decided seriousness, severity, and causal relation with the treatments? On reporting: • Were any patients excluded from the adverse effects analysis? • Does the report provide numerical data by intervention group? • Which categories of adverse effects were reported by the investigators?
Case series	• Do the reports have good predictive value? • How was causality determined? • Is there a plausible biological mechanism linking the intervention to the adverse event? • Do the reports provide enough information to allow detailed appraisal of the evidence?
Case control	• Consider typical biases for this nonrandomized study design.

From Loke et al., 2011[23]

Chou and Helfand developed a tool for an AHRQ systematic review to assess the risk of bias of studies evaluating carotid endarterectomy; the primary outcome in these studies included adverse events.[24] Four of eight items within this tool were directed specifically to assessing bias associated with adverse events; however, these criteria are applicable to other interventions, although no formal validation has been undertaken.[24] The Chou and Helfand tool has been used in comparative studies (RCTs and observational studies). No formal reliability testing has been undertaken and the tool is interpreted as a summed score across eight items. One advantage of this tool is that it includes elements of study design (for example, randomization, withdrawal) and some items specific to harms. Table A-2 shows the items within this scale.

The McMaster University Harms scale (McHarm) was developed specifically for evaluating harms and is applicable to studies evaluating interventions (both randomized and nonrandomized studies). The criteria within McHarm are detailed in Table A-3. The McHarm tool is used in conjunction with other risk of bias assessment tools that evaluate basic design features (e.g., randomization). The McHarm assumes that some biases to study conduct are unique to harms collection and that these should be evaluated separately from outcomes of benefit; scoring is considered on a per item basis. Reliability was evaluated (in expert and nonexpert raters) in RCTs of drug and surgical interventions. Internal consistency and inter-rater reliability were evaluated and found to be acceptable (greater than 0.75) with the exception of drug studies for nonexperts; in this instance the inter-rater reliability was moderate. An intra-class correlation coefficient greater than 0.75 was set as the acceptable threshold level for reliability. With the exception of nonexpert raters for drug studies, all other groups of raters showed high levels of reliability (Table A-4).

Table A-2. Quality assessment tool for studies reported adverse events[24]

Criterion	Explanation	Score
Quality criterion 1: Nonbiased selection	1: study is a properly randomized controlled trial, or an observational study with a clear predefined inception cohort (that attempted to evaluate all patients in the inception cohort) 0: study does not meet above criteria (e.g., convenience samples)	
Quality criterion 2: Adequate description of population	1: study reports two or more demographic characteristics, presenting symptoms/syndrome and at least one important risk factor for complications 0: study does not meet above criteria	
Quality criterion 3: Low loss to follow-up	1: study reports number lost to follow-up, and the overall number lost to follow-up is low (threshold set at 5% for studies of carotid endarterectomy and 10% for studies of rofecoxib) 0: study does not meet above criteria	
Quality criterion 4: Adverse events prespecified and defined	1: study reports explicit definitions for major complications that allow for reproducible ascertainment (what adverse events were being investigated and what constituted an "event") 0: study does not meet above criteria	
Quality criterion 5: Ascertainment technique adequately described	1: study reports methods used to ascertain complications, including who ascertained, timing, and methods used 0: study does not meet above criteria	
Quality criterion 6: Nonbiased ascertainment of adverse events	1: independent or masked assessment or complications (for studies of carotid endarterectomy, someone other than the surgeon who performed the procedure; for studies of rofecoxib, presence of an external endpoint committee blinded to treatment allocation) 0: study does not meet above criteria	
Quality criterion 7: Adequate statistical analysis of potential confounders	1: study examines one or more relevant confounders/risk factors (in addition to the comparison group in controlled studies), using acceptable statistical techniques such as stratification or adjustment 0: study does not meet above criteria	
Quality criterion 8: Adequate duration of follow-up	1: study reports duration of follow-up and duration of follow-up adequate to identify expected adverse events (threshold set at 30 days for studies of carotid endarterectomy and 6 months for studies of rofecoxib) 0: study does not meet above criteria	
Total quality score = sum of scores (0-8)	>6: Good 4-6: Fair <4: Poor	

Reprinted from Chou R, Fu R, Carson S, et al. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. J Clin Epidemiol 2007 Jan;60(1):18–28, with permission from Elsevier.

Table A-3. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm)

	Question
1.	Were the harms PREDEFINED using standardized or precise definitions?
2.	Were SERIOUS events precisely defined?
3.	Were SEVERE events precisely defined?
4.	Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?
5.	Was the mode of harms collection specified as ACTIVE?
6.	Was the mode of harms collection specified as PASSIVE?
7.	Did the study specify WHO collected the harms?
8.	Did the study specify the TRAINING or BACKGROUND of who ascertained the harms?
9.	Did the study specify the TIMING and FREQUENCY of collection of the harms?
10.	Did the author(s) use STANDARD scale(s) or checklist(s) for harms collection?
11.	Did the authors specify if the harms reported encompass ALL the events collected or a selected SAMPLE?
12.	Was the NUMBER of participants that withdrew or were lost to follow-up specified for each study group?
13.	Was the TOTAL NUMBER of participants affected by harms specified for each study arm?
14.	Did the author(s) specify the NUMBER for each TYPE of harmful event for each study group?
15.	Did the author(s) specify the type of analyses undertaken for harms data?

From: hiru.mcmaster.ca/epc/mcharm.pdf
Note: The answers to each question are yes (implying less risk of bias), no (implying high risk of bias), and unsure.

Table A-4. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm): inter rater reliability (intra-class correlation coefficients and confidence intervals) within different groups of raters

	Drug studies	Surgery studies	All studies
Nonexpert Raters	0.69 (0.27, 0.91)	0.92 (0.80, 0.98)	0.88 (0.77, 0.94)
Experts Raters	0.89 (0.73, 0.97)	0.93(0.85,0.98)	0.92 (0.86, 0.97)
All Raters	0.89 (0.75, 0.97)	0.96 (0.92, 0.99)	0.95 (0.91, 0.98)

From: hiru.mcmaster.ca/epc/mcharm.pdf

References

1. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008 Feb;88(2):156–75. PMID: 18073267.
2. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36(3):677–8. PMID: 17470488.
3. Whiting P, Rutjes AW, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005 Jan;58(1):1–12. PMID: 15649665.
4. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii–x, 1–173. PMID: 14499048.
5. West SL, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality; 2002.
6. Hartling L, Bond K, Harvey K, et al. Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures. Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023. AHRQ Publication No. 11-EHC007-EF. Rockville, MD: Agency for Healthcare Research and Quality: June 2009.
7. Berger VW. The (lack of) quality in assessing the quality of transplantation trials. Transpl Int 2009 Oct;22(10):1029; author reply 3. PMID: 19497066.
8. Berger VW. Is the Jadad score the proper evaluation of trials? J Rheumatol 2006 Aug;33(8):1710–1; author reply 1–2. PMID: 16881132.

9. Jadad AR. The merits of measuring the quality of clinical trials: is it becoming a Byzantine discussion? Transpl Int 2009 Oct;22(10):1028. PMID: 19740247.
10. Hartling L, Ospina M, Liang Y, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ 2009;339:b4012. PMID: 19841007.
11. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 1998 Dec;51(12):1235–41. PMID: 10086815.
12. Yates SL, Morley S, Eccleston C, et al. A scale for rating the quality of psychological trials for pain. Pain 2005 Oct;117(3):314–25. PMID: 16154704.
13. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. In: Higgins JPT, Green S, eds.: The Cochrane Collaboration; 2011.
14. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Commun Health 1998;52:377–84.
15. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics 1989 Nov;84(5):815–27. PMID: 2797977.
16. Zaza S, Carande-Kulis VG, Sleet DA, et al. Methods for conducting systematic reviews of the evidence of effectiveness and economic efficiency of interventions to reduce injuries to motor vehicle occupants. Am J Prev Med 2001;21(4 Suppl):23–30.
17. Newcastle-Ottawa Quality Assessment Scale: Case control studies. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed January 2011.
18. An evaluation of the Newcastle Ottawa Scale: an assessment tool for evaluating the quality of non-randomized studies. XI Cochrane Colloquium: Evidence, Health Care and Culture; 2003 Oct 26–31; Barcelona, Spain.
19. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 3rd Symposium on Systematic Reviews: Beyond the Basics; 2000 Jul 3–5; Oxford, UK.
20. Shamliyan TA, Kane RL, Ansari MT, et al. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. J Clin Epidemiol 2011 Jun;64(6):637–57. Epub 2010 Nov 11. PMID: 21071174.
21. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol 2012 Feb;65(2):163–78. Epub 2011 Sep 29. PMID: 21959223.
22. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol 2010 Oct;63(10):1061–70. PMID: 20728045.
23. Loke YK, Price D, Herxheimer A. Adverse effects. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. updated March 2011: The Cochrane Collaboration; 2011.
24. Chou R, Fu R, Carson S, et al. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. J Clin Epidemiol 2007 Jan;60(1):18–28. PMID: 17161750.

Chapter 6. Assessing the Applicability of Studies When Comparing Medical Interventions

David Atkins, Stephanie Chang, Gerald Gartlehner, David I. Buckley, Evelyn P. Whitlock, Elise Berliner, David Matchar

Key Points

- The PICOS framework is a useful way of organizing the review and presentation of factors that affect applicability.
- Input from clinical experts and stakeholders can help identify specific study elements that should be routinely abstracted to examine applicability.
- Population-based surveys, pharmacoepidemiologic studies, and large case series or registries of devices or surgical procedures can be used to determine whether the populations, interventions, and comparisons in existing studies are representative of current practice.
- Reviewers should assess whether benefits or harms vary along with differences in patient or intervention characteristics (i.e. effect modification) or with differences in underlying risk.
- Reports should clearly highlight important issues relevant to applicability of individual studies in a "Comments" or "Limitations" section of evidence tables and in text.
- Meta-regression, sub-group analysis and/or separate applicability summary tables may help reviewers and those using the reports see how well the body of evidence applies to the question at hand.
- Judgments about applicability of the evidence should consider the entire body of studies.

Introduction

A defining characteristic of comparative effectiveness research is that it includes "the conduct and synthesis of research comparing the benefits and harms of different interventions… in 'real world' settings" with the purpose of determining "which interventions are most effective for which patients under specific circumstances."[1] A comparative effectiveness review must therefore make judgments about whether the available research evidence reflects "real world" practice and should make clear for which patients and which circumstances the review's conclusions can be used to make clinical or policy decisions. Existing guidance on conducting systematic reviews has focused on the risk of bias in individual studies and judging whether conclusions of the review are internally valid, rather than this equally important aspect of the review process.[2]

A variety of terms have been used to describe this aspect—*applicability, external validity, generalizability, directness,* and *relevance*. Shadish and Cook define *external validity* as "inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments and outcomes."[3] The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group has used the term *directness* to cover applicability as well as other distinct aspects of the relationship between the evidence and making recommendations[4]. We prefer *applicability*, which we define as the extent to which the

effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under "real-world" conditions. This better reflects the perspective of reviews conducted by the Agency for Healthcare Research and Quality (AHRQ) Effective Health Care (EHC) Program and by many other groups (for example, guideline developers) in which systematic review aim to answer specific clinical or policy questions involving particular populations and then must make judgments about whether the available evidence is *applicable* to the questions at hand.

Relatively few clinical trials are designed with applicability in mind and furthermore, clinical studies typically report only a few of the factors needed to fully assess applicability. In contrast to the accumulating body of empiric data on factors affecting the risk of bias, or internal validity, there has been much less empiric data to determine which factors affect applicability. For these reasons, to date there has not been any detailed guidance for assessing applicability of evidence in producing systematic reviews.

This paper outlines specific steps to ensure that systematic reviews describe and characterize the evidence so that users of a review can apply it appropriately in their decisions. The first step, identifying factors that may affect applicability, should be considered at the very earliest stages of a review, when defining key questions and the populations, interventions, comparators, and outcomes of interest. Defining inclusion and exclusion criteria inevitably takes into account factors that may affect the applicability of studies—for example, reviews meant to inform decision-makers in developed countries exclude studies in developing countries because they may not be applicable to the patients and health care settings in Western countries. This paper focuses on subsequent steps in a review to describe a systematic but practical approach for considering applicability in the process of reviewing, reporting, and synthesizing evidence from eligible studies.

To develop this guidance, we searched the literature using the terms *applicability* and *external validity* and reviewed our own experience with working with users of reviews produced by the Evidence-based Practice Center (EPC) program. We extracted specific study characteristics which were proposed as relevant to external validity or applicability in the literature; the paper of Rothwell[5] provided an extensive list to which we added from other literature, prioritized based on the experience of our program, and organized under the PICOS framework (Patient, Intervention, Comparator, Outcome, Setting). We presented draft guidance at in-person meetings of the EPC program and circulated multiple drafts for review by EPC investigators. Parts of an earlier draft were posted for public comment. The final guidance document has incorporated peer and public review comments.

General Guidance

Applicability Should Be Judged Separately for Different Outcomes

The most applicable evidence may differ when considering benefits or harms since these often depend on distinct physiologic processes. For example, evidence of the benefits of aspirin for prevention of cardiovascular events from patients with heart disease cannot be readily applied to healthy populations. However, studies of patients with and without heart disease may be useful for estimating the gastrointestinal risks of aspirin which act through different mechanisms and do not vary with underlying cardiac risk.[6]

Applicability Depends on Context and Cannot Be Assessed With a Universal Rating System

Several investigators have proposed series of questions or checklists for rating applicability.[5,7-9] Critical elements vary with the clinical area and intervention studied, thus it is not clear that developing a single universal checklist is feasible. For example, there is little overlap between the items identified by Piboleau[9] for assessing applicability of orthopedic studies and those identified for assessing community interventions by Green.[8] Since we also found no empiric data validating the use of checklists for rating applicability across a range of clinical topics, we do not recommend use of any single checklist to rate applicability, but existing ones may provide a useful guide for factors to consider.

Applicability Is Best Reported Separately From the Strength of a Body of Evidence

GRADE incorporates considerations of applicability or directness into their assessments of the quality (or strength) of evidence from a body of studies, defined as the "level of confidence that an estimate of effect is correct."[4] This approach, however, does not recognize that a body of evidence with limited applicability may nonetheless provide strong evidence for one set of decisions or users but poor evidence for another. For example, early trials of thrombolysis for acute stroke may provide strong evidence for clinical decisions in specialized stroke centers but poor evidence for decisions in small rural emergency departments. We thus recommend reporting and discussing factors that limit or strengthen applicability of a body of evidence separately, rather than including it with judgments about risk of bias and other factors to determine overall quality or strength of evidence.[10] It may be reasonable to incorporate applicability into strength of evidence where reviews are created with a single primary audience in mind [11] with common, well-defined perspectives—for example, reviews for the U.S. Preventive Services Task Force incorporate into their recommendations considerations about whether the evidence is applicable to a representative North American population cared for in primary care.[12]

Four Specific Steps

We outline below four steps in assessing and reporting applicability. We distinguish the reporting and assessment of applicability of individual studies (steps 1-3) from reporting and assessment of the applicability of a body of evidence (step 4).

Step 1. Determine the Most Important Factors that May Affect Applicability

Identify potential factors. The PICOS is a useful way of organizing factors that may affect applicability. Including "setting" separately may capture information not reliably reported in population or intervention characteristics. For example, studies that recruit or treat patients in specialty settings may not be applicable to primary care populations due to differences that may not be apparent from other reported details.

Table 1 lists a variety of factors organized by the PICOS framework that may limit the applicability of individual research studies. Many of these elements are routinely captured in most systematic reviews (for example, demographics, event rates, etc.) but many other specific factors are often overlooked.

Table 1. Characteristics of individual studies that may affect applicability

	Condition that may limit applicability	Example	Feature that should be abstracted into evidence tables
Population	Narrow eligibility criteria and exclusion of those with comorbidities	In the FIT trial,13 the trial randomized only 4000 of 54,000 originally screened. Participants were healthier, younger, thinner, and more adherent than typical women with osteoporosis.	Eligibility criteria and proportion of screened patients enrolled; presence of comorbidities
	Large differences between demographics of study population and community patients	Cardiovascular clinical trials used to inform Medicare coverage enrolled patients who were significantly younger (60.1 vs. 74.7 years) and more likely to be male (75% vs. 42%) than Medicare patients with cardiovascular disease.[14]	Demographic characteristics: age, sex, race and ethnicity
	Narrow or unrepresentative severity, stage of illness, or comorbidities	Two-thirds of patients treated for congestive heart failure (CHF) would have been ineligible for major trials. Community patients had less severe CHF, more comorbidities and were more likely to have had a recent cardiac event or procedure.[14]	Severity or stage of illness; comorbidities; referral or primary care population; volunteers vs. population-based recruitment strategies.
	Run in period with high-exclusion rate for nonadherence or side effects	Trial of etanercept for juvenile arthritis used an active run in phase and excluded children who had side-effects, resulting in study with low rate of side-effects.[13]	Run in period; include attrition before randomization and reasons (nonadherence, side-effects, nonresponse)[14,15]
	Event rates much higher or lower than observed in population-based studies	In the Women's Health Initiative trial of post-menopausal hormone therapy, the relatively healthy volunteer participants had a lower rate of heart disease (by up to 50%) than expected for a similar population in the community.[16]	Event rates in treatment and control groups
Intervention	Doses or schedules not reflected in current practice	Duloxetine is usually prescribed at 40-60mg/d. Most published trials, however, used up to 120 mg/d.[17]	Dose, schedule, and duration of medication
	Intensity and delivery of behavioral interventions that may not be feasible for routine use	Studies of behavioral interventions to promote healthy diet employed high number and longer duration of visits than is available to most community patients.[18]	Hours, frequency, delivery mechanisms (group vs. individual) and duration.
	Monitoring practices or visit frequency not used in typical practice	Efficacy studies with strict pill counts and monitoring for antiretroviral treatment does not always translate to effectiveness in real world practice.[19]	Interventions to promote adherence (e.g., monitoring, frequent contact). Incentives given to study participants.
	Older versions of an intervention no longer in common use	Only one of 23 trials comparing coronary artery bypass surgery with percutaneous coronary angioplasty used the type of drug eluting stent that is currently used in practice.[15]	Specific product and features for rapidly changing technology
	Cointerventions that are likely to modify effectiveness of therapy	Supplementing zinc with iron reduces the effectiveness of iron alone on hemoglobin outcomes.[20] Recommendations for iron are based on studies examining iron alone, but patients most often take vitamins in a multivitamin form.	Cointerventions
	Highly selected intervention team or level of training/proficiency not widely available	Trials of carotid endarterectomy selected surgeons based on operative experience and low complication rates and are not representative of community experience of vascular surgeons.[21]	Selection process, training and skill of intervention team.

Table 1. Characteristics of individual studies that may affect applicability (continued)

	Condition That May Limit Applicability	Example	Feature that should be abstracted
Comparator	Inadequate dose of comparison therapy	A fixed dose study[20] by the makers of duloxetine compared 80 and 120 mg/d of duloxetine (high dose) with 20 mg of paroxetine (low dose).[22]	Dose and schedule of comparator, if applicable
	Use of substandard alternative therapy	In early trials of magnesium in acute myocardial infarction, standard of treatment did not include many current practices including thrombolysis and beta-blockade.[23]	Relative comparability to the treatment option.
Outcomes	Composite outcomes that mix outcomes of different significance	Cardiovascular trials frequently use composite outcomes that mix outcomes of varying importance to patients.[24]	Effects of intervention on most important benefits and harms, and how they are defined
	Short-term or surrogate outcomes	Trials of biologics for rheumatoid arthritis used radiographic progression rather than symptoms.[25] Trials of Alzheimer's disease drugs primarily looked at changes in scales of cognitive function over 6 months which may not reflect their ability to produce clinically important changes such as institutionalization rates.[26]	How outcome defined and at what time
Setting	Standards of care differ markedly from setting of interest	Studies conducted in China and Russia examined the effectiveness of self breast exams on reducing breast cancer mortality, but these countries do not routinely have concurrent mammogram screening as is available in the United States.[27]	Geographic setting
	Specialty population or level of care differs from that seen in community	Early studies of open surgical repair for abdominal aortic aneurysms found an inverse relationship between hospital volume and short-term mortality.[28]	Clinical setting (e.g. referral center vs. community)

Select a limited number of the most important factors that may affect applicability. Table 1 presents a wide range of items to consider. It is not feasible or necessary to record and report all of these items regardless of topic. Reviewers must instead exercise judgment to select a subset of the most important study parameters for the clinical topic. Foremost are any factors that have been associated with differences in treatment outcomes.

The observation that effectiveness of an intervention varies in different populations or settings is known as *heterogeneity of treatment effect.*[29] One cause of heterogeneity is true *effect modification*, defined when characteristics of the patient, intervention, or setting modify the relative effect of the intervention on the main outcome. Rothwell[30] notes the example where the benefits of carotid endarterectomy after a transient ischemic attack vary dramatically with the severity of the carotid stenosis and the timing of the surgery. We recommend reviewers solicit input from clinical experts and stakeholders to identify specific biologic, clinical, or health system factors that are known or suspected effect modifiers. Emphasis should be given to factors where statistically significant interactions or sub-group differences have been demonstrated in multiple studies. These factors should be identified a priori and stated in the protocol which factors will be captured in data extraction. For example, if age is a known effect modifier, evidence from studies of middle-aged adults will not be applicable to older populations.

Additionally, emerging evidence has identified a number of genetic variations that modify the effectiveness of various drugs.

A more common source for heterogeneity in treatment effect is varying baseline rates of events. Even when an intervention has constant relative effects, *the absolute benefits and harms* will vary among populations with different baseline risks. For example, although statins reduce risks of fatal and nonfatal coronary events comparably in populations at high or lower risk of heart disease, the absolute benefits in high-risk populations such as those with a previous myocardial infarction are much larger (and thus not applicable) to lower risk populations.[31] Reviewers should routinely capture information on baseline or control group risk as a factor that may affect applicability.

Finally, intervention features may affect the *ability to generalize the effectiveness or safety of the intervention to use in everyday practice*. For example, outcome studies suggest that mortality after carotid surgery is affected by the experience of the center where surgery is performed, thus evidence from trials at selected tertiary centers may not be applicable to most community populations.[21] Clinical experts, population based surveys, outcome studies, and disease or procedure registries can provide information on current treatment context and whether typical populations, settings and interventions are represented in available studies.

Step 2. Systematically Abstract and Report Key Characteristics that May Affect Applicability in Evidence Tables; Highlight any Effectiveness Studies

Once the most important factors are selected, reviewers should abstract the relevant information into evidence tables under the relevant PICOS categories. Evidence tables should also highlight effectiveness trials. These studies (also referred to as "pragmatic" or "practical" trials) are designed to give more broadly applicable results than more common efficacy studies,[32] typically by enrolling more representative populations, letting interventions vary as they often do in practice, and focusing on the most important clinical benefits and harms.[32-34] Published criteria can be used to distinguish effectiveness trials from efficacy trials.[35,36] If data from both efficacy and effectiveness studies are available, comparing findings may indicate whether more narrowly designed studies are applicable to broader populations. At the same time, reviewers must also examine whether effectiveness studies conceal important subgroup differences.[33]

Step 3. Make and Report Judgments About Major Limitations to Applicability of Individual Studies

Describe impact of applicability on interpretation of individual studies. To make applicability information useful, a review should address how specific aspects of the design of the study affected the final population or the quality of the intervention, and how greatly (and in which direction) these may differ from more representative populations in practice. For example, surgical studies that recruited surgeons based on good operative outcomes had significantly lower perioperative mortality than those observed in national Medicare hospitals,[21] (1.4 percent vs. 1.7, 1.9, or 2.5 percent for those high, average, or low volume). Thus, the balance of benefits and harms in the study are likely to overestimate those that would be expected for older patients treated in the community. Although this step involves judgment, such judgments can be made more explicit by considering how different this study is from a true *effectiveness* study and how those differences might have affected baseline risks of the population or the effectiveness or harms of the intervention.

Step 4. Consider and Summarize the Applicability of a Body of Evidence.

Applicability of a body of studies is not the same as applicability of the individual studies. A collection of studies addressing one intervention or comparison generally provides more broadly applicable evidence than any individual study. Consistent results across studies that represent an array of different populations and settings increases our confidence that results are applicable across a broad set of conditions. For example, the individual trials of statin drugs to treat high cholesterol each selected specific and discrete populations, used different drugs, different dosages, and different cointerventions. While few would qualify as effectiveness trials individually, consistent findings across trials enrolling populations of differing risks, nationalities, and underlying conditions provides evidence that the benefits of statin drugs apply across a broad range of patients.

When the number of studies is large enough, the influence of specific factors (for example, age or gender) may be explored in additional analysis such as a subgroup analysis or meta-regression. If studies vary substantially in the underlying risk or event-rate, reviewers can test whether the effectiveness of treatment varies in high- and low-risk populations and judge which studies most closely approximate the typical risk in a more representative sample—this may require analysis of more representative registry or cohort data. We caution that meta-regression or other comparisons based on group level characteristics, such as the proportion of women in each trial, can be prone to bias (the "ecological fallacy").[37] Meta-analysis based on individual-patient data is more powerful.[37]

Describe the limitations of aggregate evidence using PICOS structure. Describe whether the collected body of evidence includes relevant populations, interventions, and appropriate comparisons, includes most important outcomes, and uses representative settings. Note whether studies share features that limit applicability—for example, did all the studies exclude older, sicker patients? Where studies vary in important features, inspect whether this variation is associated with differences in measures of effectiveness or safety. Reviewers should then describe how the available body of evidence differs from "ideal" evidence to answer the question and indicate which characteristics of the evidence limit the applicability of the available evidence.

Use a summary table for applicability to highlight significant limitations to applicability. When there is a large body of evidence or when there are significant issues relevant to applicability, a summary table displays important applicability issues across a diverse body of evidence (see Table 2). One table may suffice for multiple questions if the same collection of studies is used to answer multiple questions (for example, the benefits and harms of an intervention). Critical concerns about applicability, however, can and should be described in the text.

Table 2. Elements to be included in a summary table characterizing the applicability of a body of studies

Domain	Description of applicability of evidence
Population	Describe general characteristics of enrolled populations, how this might differ from target population, and effects on baseline risk for benefits or harms. Where possible, describe the proportion with characteristics potentially affecting applicability (e.g. % over age 65) rather than the range or average.
Intervention	Describe general characteristics and range of interventions and how they compare to those in routine use and how this might affect benefits or harms from the intervention
Comparators	Describe comparators used. Describe whether they reflect best alternative treatment and how this may influence treatment effect size
Outcomes	Describe what outcomes are most frequently reported and over what time period. Describe whether the measured outcomes and timing reflect the most important clinical benefits and harms.
Setting	Describe geographic and clinical setting of studies. Describe whether or not they reflect the settings in which the intervention will be typically used and how this may influence the assessment of intervention effect.

Include the applicability of evidence in summary statements and tables addressing key questions. Comparative effectiveness reviews typically describe overall conclusions on the key questions in summary text and tables, including the effect for important outcomes and a characterization of the strength of evidence. Since we recommend separating applicability from "quality of evidence," summary conclusions should also describe the key issues affecting applicability. For example, when concluding that there is high quality evidence that carotid endarterectomy can reduce the risk of stroke and death in patients with asymptomatic carotid stenosis, it is important to specify that the evidence is applicable to patients treated at centers where the perioperative risk is less than 3 percent and who were followed an average of 4 years.[38]

Limitations of This Approach

This paper provides guidance for conducting comparative effectiveness reviews or other systematic reviews which address relatively broad clinical or policy questions in representative patient populations—for example, what is the comparative effectiveness of carotid endarterectomy vs. carotid stenting for patients with carotid stenosis? When the clinical question of interest has a much narrower focus—for example, is carotid stenting as safe and effective as carotid endarterectomy for women with a recent transient ischemic attack—it is better to restrict the review to studies which report results directly applicable to the specific question.

A related but distinct set of considerations are involved in applying evidence clinical decisions for an individual patient. Individual studies and systematic reviews give the best estimates of the average effects but these averages may not apply to many individuals.[29] As Sackett has noted, clinical decisions need to incorporate best evidence, individual patient information (e.g. disease severity, life-expectancy, comorbidity), and individual preferences.[39]

Conclusions

Understanding the applicability of scientific evidence is an important but under-examined aspect of the systematic review process. Frequently, systematic reviews collect and present an abundance of details on elements of individual studies that are relevant to the applicability of the results, but few reviews organize this information to focus attention on specific concerns related to applicability. We describe an explicit approach to identifying, reporting and synthesizing information to allow consistent and transparent consideration of the applicability of the evidence in a systematic review. Although the exact process needs to be flexible and will likely evolve,

attention to the general concepts described here will improve the ability of clinicians and policy makers to understand better to whom the conclusions of a systematic review apply, and under what conditions. In some instances it may lead to more cautious conclusions due to limitations in applicability. In others, a careful consideration of applicability may give decision makers greater confidence that the evidence summarized is appropriate and applicable for clinical and policy decisions. In both cases, it should improve the usefulness of systematic reviews, in informing practice and policy.

Author Affiliations

Office of Research and Development, Department of Veterans Affairs, Washington, DC, (DA). Center for Outcomes and Evidence, Agency for Healthcare Research and Quality**,** Rockville, MD (SC). Department for Evidence-based Medicine and Clinical Epidemiology, Danube University, Krems, Austria (GG). Oregon Evidence-based Practice Center, Oregon Health & Science University, Portland, OR (DB). Center for Health Research, Kaiser Permanente Northwest, Portland, OR (EPW). Center for Outcomes and Evidence, Agency for Healthcare Research and Quality**,** Rockville, MD (EB). Duke Center for Clinical Health Policy Research, Durham, NC, (DM), Duke-NUS Medical School, Singapore (DM).

This paper has also been published in edited form: Atkins D, Chang SM, Gartlehner G, et al. Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2011;63:481–483.

References

Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and the Congress on Comparative Effectiveness Research. Available at: http://www.hhs.gov/recovery/programs/cer/ceranualrpt.pdf. Accessed June 30, 2009.

Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2, updated September 2009. The Cochrane Collaboration 2009. Available at: http://www.cochrane-handbook.org.

Shadish, W, Cook T, Campbell D. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2002.

Guyatt GH, Oxman AD, Kunz R, et al. What is “quality of evidence” and why is it important to clinicians? BMJ 2008 May 3;336(7651):995–8.

Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” Lancet 2005 Jan 1-7;365(9453):82–93.

Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the Effective Health-Care Program. J Clin Epidemiol 2010 May;63(5):502–12.

Bornhöft G, Maxion-Bergemann S, Wolf U, et al. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Med Res Methodol 2006 Dec 11;6:56

Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 2006 Mar; 29(1):126–53

Pibouleau L, Boutron I, Reeves BC, et al. Applicability and generalisability of published results of randomised controlled trials and non-randomised studies evaluating four orthopaedic procedures: methodological systematic review. BMJ 2009 Nov 17;339:b4538.

Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions. Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol 2010 May;63(5):513–23.

Falck-Ytter Y, Schünemann H, Guyatt G. AHRQ series commentary 1: rating the evidence in comparative effectiveness reviews. J Clin Epidemiol 2010 May;63(5):474–5.

Guirguis-Blake J, Calonge N, Miller T, et al. Current processes of the U.S. Preventive Services Task Force: refining evidence-based recommendation development. Ann Intern Med 2007 Jul 17;147(2):117–22.

Cummings SR, Black DM, Thompson DE, et al. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from the fracture intervention trial. JAMA 1998;280(24):2077–82

Dhruva SS, Redberg RF. Variations between clinical trial participants and Medicare beneficiaries in evidence used for Medicare National Coverage Decisions. Arch Intern Med 2008 Jan; 169(2):136–40

Bravata DM, McDonald KM, Gienger AL, et al. Comparative Effectiveness of Percutaneous Coronary Interventions and Coronary Artery Bypass Grafting for Coronary Artery Disease. Comparative Effectiveness Review No. 9. (Prepared by Stanford-UCSF Evidence-based Practice Center under Contract No. 290-02-0017.) Rockville, MD: Agency for Healthcare Research and Quality; October 2007.

Anderson GL, Limacher M, Assaf AR, et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. JAMA 2004 Apr 14;291(14):1701–12.

Gartlehner G, Hansen RA, Thieda P, et al. Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression. Comparative Effectiveness Review No. 7. (Prepared by RTI International-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality; January 2007.

Whitlock EP, O'Connor EA, Williams SB, et al. Effectiveness of Weight Management Programs in Children and Adolescents. Evidence Report/Technology Assessment No. 170 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). AHRQ Publication No. 08-E014. Rockville, MD: Agency for Healthcare Research and Quality; September 2008.

Fletcher CV. Translating efficacy into effectiveness in antiretroviral therapy: beyond the pill count. Drugs 2007;67(14):1969–79.

Walker, CF, Kordas K, Stoltzfus, RJ, et al. Interactive effects of iron and zinc on biochemical and functional outcomes in supplementation trials. Am J Clin Nutr 2005 82:5–12.

Wennberg D, Lucas F, Birkmeyer J, et al. Variation in carotid endarterectomy mortality in the Medicare population. JAMA 1998;279:1278–81.

Detke MJ, Wiltse CG, Mallinckrodt CH, et al. Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. Eur Neuropsychopharmacol 2004 Dec;14(6):457–70.

Li J, Zhang Q, ZhangM, et al. Intravenous magnesium for acute myocardial infarction. Cochrane Database of Systematic Reviews 2007, Issue 2. Art. No.: CD002755. DOI: 10.1002/14651858.CD002755.pub2.

Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007;334;786; originally published online 2 Apr 2007

Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol 1997 Oct;50(10):1089–98.

Hansen RA, Gartlehner G, Kaufer D, et al. Drug class review of Alzheimer's drugs. Final report. 2006. Available at: http://www.ohsu.edu/drugeffectiveness/reports/final.cfm.

Humphrey L, Chan BKS, Detlefsen S, et al. Screening for Breast Cancer. Prepared by Oregon Health Sciences University under Contract No. 290-97-0018. Rockville, MD. Agency for Healthcare Research and Quality; August 2002.

Wilt TJ, Lederle FA, MacDonald R, et al. Comparison of Endovascular and Open Surgical Repairs for Abdominal Aortic Aneurysm. Evidence Report/Technology Assessment No. 144. (Prepared by the University of Minnesota Evidence-based Practice Center under Contract No. 290-02-0009.) AHRQ Publication No. 06-E017. Rockville, MD: Agency for Healthcare Research and Quality; August 2006.

Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661–87.

Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials 2006 May;1(1):e9

National Institute for Health and Clinical Excellence. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. London: NICE; 2008. Available at: www.nice.org.uk/ CG67

Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003 Sep 24;290(12):1624–32.

Godwin M, Ruhland L, Casson I, et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003 Dec 22;3:28.

Atkins D. Creating and synthesizing evidence with decision makers in mind: integrating evidence from clinical trials and other study designs. Med Care 2007 Oct; 45(10 Supl 2):S16–S22.

Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006 Oct;59(10):1040–8. Epub 2006 Aug 4.

Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009 May;62(5):464–75.

Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010 Feb 5;340:c221. doi: 10.1136/bmj.c221.

Chambers BR, Donnan G. Carotid endarterectomy for asymptomatic carotid stenosis. Cochrane Database of Systematic Reviews 2005, Issue 4. Art. No.: CD001923. DOI: 10.1002/14651858.CD001923.pub2.

Sackett DL, Richardson WS, Rosenberg W, et al. Evidence-based medicine—how to practice and teach EBM. New York: Churchill Livingstone; 1997.

Chapter 6 Appendix A—Example Adapted from Comparative Effectiveness Review of Therapies for Clinically Localized Prostate Cancer[A1]

We have augmented consideration of applicability from a previous comparative effectiveness review[A1] illustrating the different steps for assessing and reporting the applicability of the evidence to the following question:

How do the benefits and harms of radical prostatectomy compare to watchful waiting for treatment of early organ-confined prostate cancer?

Step 1. Determine the Most Important Factors that May Affect Applicability

In order to determine the important factors, the reviewers must consider the underlying biology and epidemiology as well as the historical and current clinical practice context.

Epidemiologic studies indicate that prostate cancer prognosis is tied to *grade* and, to a lesser extent, *stage* of cancer. Cancer registries in the United States indicate that most localized cancers are detected by PSA testing (Stage T1c), with the majority diagnosed in men over age 65. Clinical experts think that *age and comorbidity* affect benefits and risks of aggressive therapy (by creating competing risks which reduce the benefits of aggressive interventions and by increasing risks of surgery). Specific *cointerventions* or *surgical techniques* (e.g. nerve-sparing approaches or adjuvant hormonal therapy) and *experience of the participating centers and surgeons* may influence both the effectiveness of treatment and adverse event rates.

Step 2. Systematically Abstract and Report Characteristics that May Affect Applicability in Evidence Tables; Highlight Any Effectiveness Studies

Table A-1 is an abbreviated version of an evidence table, into which the reviewer extracts relevant data from individual studies, used to judge both internal validity and applicability. However, this example table focuses only on data related to applicability of the study.

Table A-1. Example evidence table of individual studies with key applicability factors abstracted and judgment of applicability

Trial (including date, setting)	Population Demographic, Disease state	Intervention	Comparator	Outcomes and timing	Comments
Bill-Axelson et al.[A2] (SPCG-4) 1989-1999, Sweden	Mean age 65 78% T2 60% Gleason 6 or lower. Few detected by PSA	Radical prostatectomy at 18 centers; standard current protocol	Watchful waiting with deferred hormonal therapy	Prostate-specific antigen and all cause mortality; metastasis and disease progression; median follow-up of 8.3 years	Some indications of an effectiveness trial. Unclear how highly selected the enrolled patients were. Limited standardization of the intervention. Unclear whether the participating centers and surgeons are representative of the larger population.
Iversen et al.[A3] 1967-1975 Denmark	Mean age 64.2 46.5% Stage 2 86.5% Gleason 6 or lower. None detected by PSA.	Radical prostatectomy in one Veterans Administration center, protocol from 1967-1975	Watchful waiting with oral placebo	Overall mortality; Median follow-up 23 years	Results may not be applicable to current practices due to the evolving techniques in both stage and grade classification since PSA screening.

Step 3. Make and Report Judgments About Major Limitations to Applicability of Individual Studies

Once the appropriate data for assessing applicability of individual studies has been identified, the reviewer must then consider what impact it will have when interpreting the results of the study in relation to the question being asked.

The reviewer can then highlight and summarize the key concerns or strengths of an individual study for its applicability to the question, highlighting effectiveness studies. We illustrate how this might be done in the comments column of Table A-1 above.

Step 4. Consider and Summarize the Applicability of a Body of Studies

After identifying the major strengths and limitations in applicability for individual studies, the reviewer must then consider the applicability of the body of evidence and considering how the limitations may impact the interpretation of the evidence in answering the question. In order to do this, it may be helpful to use a summary table for applicability, as illustrated in Table A-2.

Table A-2. Example summary table characterizing the applicability of a body of studies

Domain	Description of applicability of evidence
Population	Available trials included few patients with PSA detected by screening (T1c), whose prognosis may be different. The age of enrolled patients was representative of prostate cancer patients in the community, but subgroup results from one study suggest that benefits of treatment may be smaller in patients over age 65 than those under age 65.
Intervention	The prostatectomy treatment in the Scandinavian study[A2] is applicable to current surgical methods although it is not clear if nerve-sparing surgery was common. The smaller trial[A3] was conducted over 20 years ago and may not be applicable.
Comparators	Watchful waiting is an appropriate comparator in both studies but only the more recent study used hormonal therapy for patients whose disease progresses.
Outcomes	Available trials use a reasonable array of health outcomes. Additional follow-up from one study suggests that outcomes at 10 years are representative of longer-term outcomes. For older patients, prostate cancer mortality may represent a small portion of overall mortality and thus be less relevant than overall mortality.
Setting	One study was conducted across a broad cross section of Scandinavian centers, whereas the other was conducted in a highly selected population from one Danish Veterans Administration center in the 1960's-1970's. It is not clear in what direction this may affect the results. They may be a healthier population from having regular access to medical care, but may be more likely to have other comorbidities such as heart disease than a highly selected population.

With use of a summary applicability table, it becomes easier for a reviewer to describe in the text how aspects of the study may impact the interpretation of the study results in answering the question. An example of a text summary of applicability and their implications is provided below.

Two trials have addressed the benefits of surgical therapy compared to deferred therapy or watchful waiting. Results are dominated by one trial, which demonstrated important but modest benefits of prostatectomy. There are important concerns about the applicability of this evidence to the population of interest. These results are most applicable to patients under 65 with T2 prostate cancer but cannot be assumed to apply to the largest group of prostate cancer patients in the United States, those with cancers detected by PSA screening (T1c). Such patients have a substantially better untreated prognosis and would be unlikely to benefit as much from surgery, at least over the 8 to 10 year time period of the available trials. Whether results apply to older patients is unclear. Patients over age 65 had smaller benefits in a subgroup analysis of the

Swedish trial but this difference was not statistically significant; nonetheless the high risk of competing causes of death reduces the number of patients that will live long enough to benefit.

Finally, at the level of synthesis, the reviewer should describe the applicability of the evidence in the highest level of summary conclusions. This is often presented in the form of the summary table (Table A-3).

Table A-3. Example summary table for body of evidence

Comparison	Strength of Evidence	Conclusions with description of applicability
Radical prostatectomy vs. watchful waiting	Medium	Compared with men who used watchful waiting, men with localized prostate cancer detected by methods other than PSA testing and treated with radical prostatectomy (RP) experienced fewer deaths from prostate cancer and fewer distant metastases. The benefits of RP on cancer-specific and overall mortality appears to be limited to men under 65 years of age but is not dependent on baseline PSA level or histologic grade.

References

A1. Wilt TJ, Shamliyan T, Taylor B, et al. Comparative Effectiveness of Therapies for Clinically Localized Prostate Cancer. Comparative Effectiveness Review No. 13. (Prepared by Minnesota Evidence-based Practice Center under Contract No. 290-02-0009.) Rockville, MD: Agency for Healthcare Research and Quality; February 2008.

A2. Bill-Axelson A, Holmberg L, Ruutu M, et al. Scandinavian Prostate Cancer Screening Group Study No. 4. Radical prostatectomy versus watchful waiting in early prostate cancer. N Engl J Med 2005 May 12;12(19):1977–84.

A3. Iversen P, Madsen PO, Corle DK. Radical prostatectomy versus expectant treatment for early carcinoma of the prostate. Twenty-three year followup of a prospective randomized study. Scan J Urol Nephrol Suppl 1995;172:65–72.

Chapter 7. Assessing Harms When Comparing Medical Interventions

Roger Chou, Naomi Aronson, David Atkins, Afisi S. Ismaila, Pasqualina Santaguida, David H. Smith, Evelyn Whitlock, Timothy J. Wilt, David Moher

Key Points

- Assess all important harms, whenever possible.
- Use multiple sources of information, including clinical experts and stakeholders, to identify important harms.
- Use consistent and precise terminology when reporting data on harms, and avoid terms implying causality unless causality is reasonably certain.
- 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.
- Be cautious about drawing conclusions on harms when events are rare and estimates of risk are imprecise.
- Include placebo-controlled trials, particularly for assessing uncommon or rare harms, but be cautious about relying on indirect comparisons to judge comparative risks, and evaluate whether studies being considered for indirect comparisons meet assumptions for consistency of treatment effects.
- Avoid inappropriate combining of data on harms, and thoroughly investigate inconsistent results.

Introduction

Comparative Effectiveness Reviews (CERs) are systematic reviews that evaluate evidence on alternative interventions in order to help clinicians, policymakers, and patients make informed treatment choices.[1] To generate balanced results and conclusions, it is important for CERs to address both benefits and harms.[2] However, assessing harms can be difficult. Benefits have been accorded greater prominence when reporting trials, with little effort to balance assessments of benefits and harms. In addition, systematically reviewing evidence for all possible harms is often impractical, as interventions may be associated with dozens of potential adverse events. Furthermore, there are often important tradeoffs between increasing comprehensiveness and decreasing quality of harms data.[3]

Adequately assessing harms requires CER authors to consider a broad range of data sources. For that reason, they need to deal with important challenges, such as choosing which types of evidence to include, identifying studies of harms, assessing their quality, and summarizing and synthesizing data from different types of evidence.

Identifying Harms To Be Evaluated

CERs should always assess harms that are important to decisionmakers and users of the intervention under consideration.[4] High-priority harms should include the most serious adverse events; they may also include common adverse events and other specific adverse events important to clinicians and patients. CER authors should examine previously published reviews, review publicly available safety reports from the U.S. Food and Drug Administration (FDA), and consult with technical experts and patients to set priorities for evaluating harms. Searches on postmarketing surveillance databases may also help identify important potential harms. The methods sections of the CER should specify the process used to identify harms of interest and list the specific harms for which evidence was sought.

Terminology

Terminology related to reporting of harms is poorly standardized.[5] This can cause confusion or result in misleading conclusions. CER authors should strive for consistent and precise usage of terminology when reporting data on harms. 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 harms, unless causality is reasonably certain.

Definitions for commonly used terms for harms reporting are summarized in Table 1, along with suggested usage.[4-6]

Table 1. Terminology for reporting on harms

Active surveillance of harms	Participants are asked in structured questionnaires or interviews about the occurrence of specific adverse events, or predefined laboratory or other diagnostic tests are performed at prespecified time intervals.
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 the purpose of the Comparative Effectiveness Review is to establish causality, "adverse event" should generally be the default term over "adverse effect" or "adverse reaction/adverse drug reaction."
Adverse reaction/adverse drug reaction	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	Participants are not specifically asked about or tested for the occurrence of adverse events. Rather, adverse events are identified based on patient reports made on their own initiative.
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*."

Table 1. Terminology for reporting on harms (continued)

Safety	Substantive evidence of an absence of harm. Do not use this term (or the term "safe") when evidence on harms is simply absent or insufficient.
Serious adverse event	Any adverse event with serious medical consequences, including death, hospital admission, prolonged hospitalization, and persistent or significant disability or incapacity.
Severe adverse event	An adverse event whose intensity is considered severe (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) given at 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	The term "toxicity" is used in pharmacology and microbiology to refer to the quality of being poisonous, especially the degree of virulence of a toxic microbe or of a poison. It is often measured in terms of the specific target affected (e.g., cytotoxicity or hepatotoxicity). In the context of systematic reviews, the term is often used to refer to laboratory-determined abnormalities, such as elevated liver function tests. However, the terms "abnormal laboratory measurements" and "laboratory abnormalities" are more specific and appropriate.

Sources of Evidence on Harms

Randomized Controlled Trials

Published trials. 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 RCTs to evaluate harms in CERs is problematic. First, most RCTs lack prespecified hypotheses for harms.[5] 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 is frequently inadequate.[7,8]

Second, few RCTs have large enough sample sizes or are long enough in duration to adequately assess uncommon or long-term harms.[9]

Third, most RCTs are explanatory, rather than pragmatic, in design—i.e., they assess benefits and harms in ideal, homogeneous populations and settings.[10] Patients who are more susceptible to adverse events are often underrepresented in such "efficacy" trials. Even when harms are appropriately assessed and reported, the applicability of efficacy trials to general practice is limited.

Fourth, relatively few RCTs directly compare alternative treatment strategies. Although CER authors can evaluate benefits or harms of two competing interventions based on trials in which each is compared with a common third treatment (usually placebo), the results of indirect comparisons do not always agree with direct comparisons.[11,12]

Fifth, publication and selective outcome(s) reporting bias can lead to distorted conclusions about harms when data are unpublished, partially reported, downplayed, or omitted.[13,14]

Finally, in some cases, RCTs may not be available. 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.[15]

Despite these limitations, RCTs are the gold standard for demonstrating efficacy, the basis for most regulatory approvals, and the source of most advertising and other claims made on behalf of drugs and other interventions. For this reason, CERs must address harms data from RCTs in detail when they are available.

"Head-to-head" RCTs provide the most direct evidence on comparative harms. However, placebo-controlled RCTs may also provide important information on absolute and relative risks and contribute to more precise estimates of harms. In addition, placebo-controlled trials can provide information about risks that may not be apparent from head-to-head trials. For example, a systematic review of nonsteroidal anti-inflammatory drugs (NSAIDs) found cyclo-oxygenase-2 selective NSAIDs associated with greater myocardial risk vs. placebo, but differences were not apparent vs. nonselective NSAIDs, which were also associated with increased risk.[16] In general, CERs should routinely include placebo-controlled trials for assessment of harms, particularly for rare or uncommon adverse events. In lieu of examining individual placebo controlled trials, CERs may incorporate findings of well-conducted systematic reviews, provided they evaluate the specific harms of interest.

Unpublished supplemental trials data. In addition to evaluating results of published RCTs, CER authors should consider including results of completed or terminated but unpublished RCTs, as well as unpublished results from published trials. Such information has several potentially valuable uses:

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

Unpublished clinical trials tend to report lower estimates of treatment benefits than published trials (i.e., weaker intervention effects).[17,18] The impact of unpublished trials on assessments of harms has not been extensively studied, but a systematic review of antidepressants in children found that addition of data from unpublished trials changed conclusions about the balance of risks and benefits from favorable to unfavorable for several drugs.[19]

Data from unpublished trials can be difficult to locate systematically. At a minimum, material from the FDA Web site should routinely be examined in order to assess what effect unpublished (completed or terminated) trials submitted for regulatory approval may have on conclusions regarding harms. In addition, starting in 2009, trial sponsors are required by the 2007 FDA reform bill to report results to a clinical trial results database (www.ClinicalTrials.gov).[20] Other resources for identifying unpublished trials include obtaining information from non-U.S. regulatory agencies and directly querying funding sources. Once unpublished trials are located, two caveats should also be considered. Frequently, there is insufficient information from unpublished trials to assess fully the risk of bias. Also, the results

and conclusions of trials may change between initial presentation of data and publication in a peer-reviewed journal.[21]

Even when a trial is published, important information may be omitted because of space limitations or other reasons.[22,23] For example, before the publication of the Vioxx Gastrointestinal Outcomes Research Study (VIGOR) in 2001,[24] information on myocardial infarctions was absent from most published reports of trials evaluating selective or nonselective NSAIDs because an association with cardiovascular events was not suspected. A systematic review that obtained unpublished myocardial infarction data from older trials found an increased risk with high doses of all evaluated NSAIDs (selective or nonselective) other than naproxen.[16] An analysis of myocardial infarction risk based on only published information would have been seriously compromised by incomplete data.

Drug approval information—for example, the clinical and statistical reviews prepared by staff of the 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 the *Journal of the American Medical Association* as a 6-month study and reported fewer gastrointestinal adverse events for celecoxib than for two nonselective NSAID comparators.[25] The JAMA article did not mention that some patients in the trial had been observed for longer than 6 months.[26] In contrast, the FDA review reported all the outcomes data, including data that showed no difference in gastrointestinal adverse events at the end of followup.[27]

Limited evidence suggests an inverse relationship between the proportion of included trials reporting a specific outcome and the estimates of treatment benefit for that outcome, possibly due to selective reporting of favorable outcomes.[28] How the proportion of included trials reporting outcomes affects estimates of harms has not been well studied. Nonetheless, when a significant proportion of published trials fail to report an important or critical adverse event, CER authors should report on this gap in the evidence and consider efforts to obtain unpublished data (e.g., by querying study authors, funding sources, or clinical trials results databases, or performing more detailed reviews of FDA documents).

Observational Studies

Observational studies are almost always necessary to assess harms adequately. The exception is when there are sufficient data from RCTs to reliably estimate harms. However, even though observational studies are more susceptible to bias than well-conducted RCTs, for some comparisons there may be few or no long-term, large, head-to-head, or effectiveness RCTs.[29] Observational studies may also provide the best (or only) evidence for evaluating harms in minority or vulnerable populations (such as pregnant women, children, elderly patients, or those with multiple comorbidities) who are underrepresented in clinical trials.

The term “observational studies” is commonly used to refer to cohort, case-control, and cross-sectional studies,[30] but can refer to a broad range of study designs, including case reports, uncontrolled series of patients receiving surgery or other interventions, and others.[31] All can yield useful information as long as their specific limitations are understood.

The types of observational studies included in a CER will vary depending on the type or frequency of adverse events being evaluated. The choice of study designs also depends on whether investigators are seeking to determine what harms might be associated with a treatment (hypothesis generating) or whether certain harms are more likely (hypothesis testing). Different

types of observational studies might be included or rendered irrelevant by availability of data from stronger study types.

Cohort and case-control studies. CER authors should routinely search for and include well-designed and reported case-control and population-based cohort studies.[30,32] 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. They also take stronger precautions against bias than do other observational designs, and their strengths and weaknesses are well understood. For unexpected adverse events, for example, confounding by indication may not be as important an issue in case-control and cohort studies as when evaluating beneficial effects because their occurrence is usually not associated with the reasons for choosing a particular treatment.[29,33] Although cross-sectional studies have features in common with cohort studies, it is difficult to establish causality because exposures, and outcomes are evaluated simultaneously. Indeed, associations in cross-sectional studies may sometimes be due to reverse causality.[34]

A recent report found that large observational studies usually report smaller absolute risks of harm than do large randomized trials.[35] There was no clear tendency for randomized trials or observational studies to report larger 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, including criteria used to identify harms; differential effects of biases; or some combination of these factors.

Observational studies based on patient registries. Patient registries collect information on clinical outcomes in populations defined by a particular disease, condition, or exposure.[36] Clinical data are prospectively collected for specific research purposes using active methods to identify outcomes, although registry information can be supplemented by information from administrative databases and other sources. Registries can be designed as an active surveillance system for identifying harms and may be particularly useful for assessing long-term or uncommon adverse events.

Observational studies based on analyses of large databases. Pharmacoepidemiologic studies using large databases to identify exposures and outcomes may be valuable for comparing the risk of uncommon adverse events.[37] However, additional empirical research is needed to identify methods for collecting and analyzing data in pharmacoepidemiologic studies that are associated with valid findings.[38] Unlike studies based on patient registries, large administrative databases usually contain information routinely collected during clinic, hospital, laboratory, or pharmacy encounters, rather than for a specific research purpose. Such studies are probably most useful for evaluating serious harms that are more reliably reported and recorded (for example, death or acute myocardial infarction) than less serious harms that may not generate a specific clinic visit or diagnostic code (for example, sedation or nausea). In some cases, administrative data may be supplemented or verified by more detailed clinical information. Regardless of how data are obtained, 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.[39] Case reports can be useful for identifying uncommon, unexpected, or long-term adverse events, particularly for new drugs or other interventions.[40] The adverse events identified by case reports often differ from those detected in clinical trials.[41] However, case reports are usually considered to be hypothesis generating because it is difficult to calculate information from them about the frequency or comparative risk of adverse events.

In the United States, the FDA receives about 280,000 reports of postmarketing adverse events annually, collects them into a database,[42] and issues information about adverse drug events on its MedWatch Web site (http://www.fda.gov/medwatch/). Although pharmaceutical companies and other investigators may also perform passive surveillance of harms on postmarketing data, such analyses are not always made public in a timely fashion.[43] Active, hypothesis-driven postmarketing surveillance systems have been developed recently for identifying and evaluating serious adverse drug events.[44]

Case reports and other hypothesis-generating studies may be useful for CERs evaluating new drugs suspected of being associated with serious but uncommon adverse events. For other topics, CER authors 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 RCTs or well-designed case-control or cohort studies, 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.[45] In addition, because such studies lack control groups, evaluating effects of confounding is difficult, as is comparing risks of adverse events across interventions.

Other quasi-experimental study designs may not offer any advantage over RCTs in terms of their applicability to routine practice. For example, open-label extensions of clinical trials may follow patients for an extended period of time, but they usually enroll a more highly selected population (patients who completed the randomized trial, tolerated the medication, and agreed to participate in the extension), are unblinded, and often lack a comparison arm. Such studies 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 impractical 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 match inclusion criteria to the reasons for including observational studies. For example, inclusion criteria might specify minimum duration of followup if a priority is to identify evidence on long-term harms. If large, higher quality studies are available, it could be reasonable to specify a minimum sample size threshold in order to utilize resources efficiently. Methods sections should clearly describe

selection criteria along with the rationale for choosing the criteria. Commonly used inclusion criteria for observational studies are shown in Table 2.

Table 2. Example criteria for selecting observational studies on harms for inclusion in a Comparative Effectiveness Review

Studies meet certain study design definitions (e.g., cohort and case-control studies)
Studies do not exceed a defined threshold for risk of bias (e.g., studies assessed as being at low risk of bias or meeting certain prespecified quality criteria)
Studies meet a defined threshold for duration of followup
Studies meet a sample size threshold
Studies evaluate a specific population of interest (e.g., studies evaluating populations underrepresented in randomized trials, such as elderly, women, or minority populations)

Assessing Risk of Bias (Quality) of Harms Reporting

Randomized Trials

A number of features of RCTs have been empirically tested and proposed as markers of higher quality (i.e., lower risk of bias). These include use of appropriate randomization generation and allocation concealment techniques; blinding of participants, health care providers, and outcomes assessors; and analysis according to intention-to-treat principles.[46] Whether these are equally important in protecting against bias in studies reporting harms is unclear. Moreover, because evaluating harms is often a secondary consideration in randomized trials, the quality of harms assessment and reporting can be inadequate even when assessment of the primary (beneficial) outcome is appropriate.

When evaluating the quality of harms assessment, CER authors should consider whether adequate methods were used to identify adverse events in the primary studies. Active methods, such as querying patients using a comprehensive checklist or standardized laboratory tests, are more likely to completely identify adverse events than passive methods, such as relying on patient self-report.[47] In addition, specific data on adverse events are likely to be more accurate and informative than 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 CER authors to assume that they were not ascertained or not recorded than to assume that the prevalence or incidence was zero.[4]

It is also important to assess how adverse events are assessed and categorized. Studies should 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.[48,49] 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 or may be poorly reproducible. Similarly, methods for classifying adverse events as “treatment related” are largely subjective, with unknown validity, and such data may be particularly unreliable.

It is not always necessary for trials to prespecify or define adverse events. For example, studies reporting unexpected outcomes can be very valuable for identifying previously unrecognized harms. However, when evaluating known harms, using validated or standardized criteria for adverse events may help reduce subjectivity or bias in their assessment and classification. In drug trials, use of an independent external endpoint committee may provide less

biased estimates of harms than outcomes assessment performed by investigators connected to the study.[50]

"Withdrawals due to adverse events" are commonly reported in trials, and they are often used in systematic reviews as a marker for intolerable or severe adverse events. However, the Cochrane Adverse Effects Methods Group suggests caution in interpreting withdrawals attributed to adverse events in this manner, for the following reasons:[4]

- Attribution of reasons for discontinuation is likely to be imprecise and to vary across trials.
- Pressure to keep dropouts low in trials may result in rates that do not reflect real-world practice.
- 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.

Observational Studies

Because observational studies lack randomization, they should adhere to high methodological standards to be considered valid.[30,32,51] RCTs 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 a clearly identified inception cohort (e.g., "new users")[52] is at least as important when assessing observational studies.

Instruments for assessing risk of bias in observational studies vary greatly in scope, number and types of items used, and developmental rigor.[53] Further study is needed to determine which methodological shortcomings in observational studies are consistently associated with bias in assessment and reporting of harms. However, some consensus exists on the major domains that should be considered when evaluating the overall validity of an observational study. For cohort studies, important factors include assembly of an inception cohort, complete followup, appropriate assessment of potential confounders, accurate determination of exposures and outcomes, and blinded assessment of outcomes.[30,52-54]

Several studies have empirically evaluated effects of specific methodological characteristics on estimates of harms from observational studies. They found that prospective or retrospective design,[55,56] case-control compared with cohort studies[57,58] and smaller compared with larger case series[55] did not have consistent effects on estimates of harms. Two studies found that industry-funded studies tended to report more favorable outcomes than did studies with other funding sources.[57,59] Because all of these studies evaluated fairly limited samples of studies, their wider applicability 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).[37]

For all observational study designs, estimates of harms are less likely to be confounded when evaluating previously unsuspected adverse events than when evaluating a known harm or intended effects. For example, the finding that cyclo-oxygenase-2-selective NSAIDS were associated with an increased risk of myocardial infarction vs. nonselective NSAIDs was an unexpected finding from an RCT examining a different outcome.[24] This risk could be confirmed in observational studies, in part because the choice of type of NSAID in typical practice was unrelated to the patients' risk for myocardial infarction. In contrast, gastrointestinal bleeding was a known risk of nonselective NSAIDS, and clinicians were more likely to prescribe selective NSAIDs in patients at higher risk for gastrointestinal bleeding. Such "confounding by indication" led to the appearance of an apparent association between selective NSAID use and bleeding in epidemiologic studies.[60] In some cases, such spurious associations may remain despite adjustment for known confounders ("residual confounding").

Uncontrolled Studies

Studies of surgery, medical devices, and other nonpharmacologic interventions are often uncontrolled series of patients who received the therapy and then were followed over a period of time. Such studies can provide some information about rates of adverse events in clinical practice, and they may be most informative when the incidence of such events in untreated patients is low. Unfortunately, such studies frequently do not meet standards for accurate and comprehensive reporting of harms.[61] Even when harms data are well described, an important limitation of uncontrolled studies is that it is difficult to evaluate confounding by indication. Authors are also more likely to submit for publication studies showing the best outcomes.

For some interventions, CER authors must consider including uncontrolled studies for assessing harms, as little or no other evidence may be available. Proposed criteria for evaluating case series are likely to promote improved reporting of results,[62] but may provide only limited information about risk of bias. Important factors to consider when evaluating uncontrolled studies include whether the study enrolled or attempted to enroll all patients meeting prespecified inclusion criteria and whether the study clearly describes loss to followup.[45] When uncontrolled studies do not meet these criteria, determining the reliability and applicability of even well-described results may be 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 issues remain unclear: 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 an instrument for assessing quality of harms assessment and reporting in randomized trials and observational studies of carotid endarterectomy for symptomatic carotid artery stenosis.[63] This approach involved four criteria: nonbiased selection of subjects, low loss to followup, adverse events prespecified and defined, and adequate duration of followup. Studies meeting at least three of the four criteria reported a rate of postsurgical complications of 5.7 percent (95 percent confidence interval [CI], 4.8 percent to 6.6 percent), compared with 3.7 percent (95 percent CI, 3.1 percent to 4.3 percent) for studies meeting fewer than three of the criteria. However, the generalizability of this instrument to other datasets or interventions is unclear. When the authors applied these criteria to studies of

rofecoxib, they were unable to show differences in estimates of risk of myocardial infarction. In addition, caution should be used when considering use of summary scores to assess risk of bias.[64] At a minimum, key methodological aspects should be assessed individually and their influence on estimates of harms explored.

Santaguida et al. have also developed a quality-rating instrument (McHarm) for evaluating studies reporting harms (Table 3).[65] The tool was developed from quality rating items generated by a review of the literature on harms and from 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 standardized quality-assessment tools for design-specific internal validity issues.

Table 3. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm)

1. Were the harms PRE-DEFINED using standardized or precise definitions?
2. Were SERIOUS events precisely defined?
3. Were SEVERE events precisely defined?
4. Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?
5. Was the mode of harms collection specified as ACTIVE?
6. Was the mode of harms collection specified as PASSIVE?
7. Did the study specify WHO collected the harms?
8. Did the study specify the TRAINING or BACKGROUND of who ascertained the harms?
9. Did the study specify the TIMING and FREQUENCY of collection of the harms?
10. Did the author(s) use STANDARD scale(s) or checklist(s) for harms collection?
11. Did the authors specify if the harms reported encompass ALL the events collected or a selected SAMPLE?
12. Was the NUMBER of participants that withdrew or were lost to follow-up specified for each study group?
13. Was the TOTAL NUMBER of participants affected by harms specified for each study arm?
14. Did the author(s) specify the NUMBER for each TYPE of harmful event for each study group?
15. Did the author(s) specify the type of analyses undertaken for harms data?

Source: Santaguida PL, Raina P. The development of the McHarm quality assessment scale for adverse events: Delphi consensus on important criteria for evaluating harms. http://hiru.mcmaster.ca/epc/mcharm.pdf. Accessed May 14, 2008.

Case reports may provide valuable information about the possibility of rare or previously unrecognized adverse events. A 1982 study examined 47 case reports published in 1963 in four major general medical journals and judged that 35 of them were subsequently proved to be "clearly" correct.[66] 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.[67] Nonetheless, statistical modeling study suggests 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.[68] Case reports, however, cannot be used to estimate the rate of an adverse event, which may be critical to any decisions.

Several disease-specific[69] and non-disease-specific[70] methods for assessing the probability of causality from case reports of adverse events have been developed. These methods represent expert opinion and have not been validated empirically. Factors believed to increase the likelihood of causality are shown in Table 4.[69,70]

Table 4. Criteria for evaluating the likelihood of a causal relationship in case reports

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)
Confirmation of adverse event by objective information

Guidelines for improving the reporting of suspected adverse drug events in case reports have also recently been proposed.[71] In 35 reports of 48 patients published in the *British Medical Journal*, the median number of recommended items that were reported was 9 of 19 (range 5-12), although effects of missing information on the validity of case reports have not been studied.

Synthesizing Evidence on Harms

CER authors should follow general principles for synthesizing evidence when evaluating data on harms. Such principles include: combining studies only when they are similar enough to warrant combining;[72] adequately considering risk of bias, including publication and other related biases;[73] and exploring potential sources of heterogeneity.[23] Several other issues are especially relevant for synthesizing evidence on harms.

Uncommon or Rare Adverse Events

Evaluating comparative risks of uncommon or rare adverse events in CERs can be particularly challenging. A frequent problem in RCTs and systematic reviews is interpreting a nonsignificant probability value 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.[74,75] 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 more than 0.20.[76] However, the 95 percent confidence intervals for estimates of relative risks for these three adverse events 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). In such cases, CERs should acknowledge the lack of statistical power to assess risk adequately and should interpret the confidence intervals, including the possibility or probability of excess harm.

Equivalence and Noninferiority

CER authors should draw conclusions about "equivalence" or "noninferiority" of interventions with regard to harms only when there are appropriate data to justify such statements.[77] Few CERs will have the statistical power to adequately assess noninferiority when the risk of an adverse event is on the order of 1 percent or lower. For example, about 100,000 patients would have been needed in the COBALT or GUSTOIII trials to rule out an excess relative death rate of 5 percent from alternative thrombolytic agents with 80 percent power.[78] Ruling out smaller event rates would require even higher sample sizes.

Indirect Analyses

Placebo-controlled trials can be helpful for evaluating absolute risks associated with an intervention. When head-to-head trials are sparse or unavailable, placebo-controlled trials may also be useful for indirectly evaluating comparative harms, particularly for rare or uncommon adverse events. However, for indirect analyses to be reliable, all studies should be comparable in terms of quality, factors related to applicability (population, dosing, co-interventions, and settings), measurement of outcomes, and incidence of adverse events in control groups.[12,79]

For example, a meta-analysis found that rofecoxib was associated with an increased risk of arrhythmia compared with control treatments; celecoxib was not.[80] However, the rate of arrhythmia in the control arms was tenfold higher in trials of celecoxib (0.27 percent, or 18 of 6,568 subjects) than in trials of rofecoxib (0.02 percent, or 2 of 10,174 subjects). In this situation, indirect comparisons about the relative safety of celecoxib compared with rofecoxib are likely to be problematic. 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 comparisons.

More studies are needed to determine when indirect comparisons are most likely to be valid. In the meantime, CER authors considering indirect analyses to assess harms should carefully consider whether assumptions underlying valid indirect comparisons are likely to be met, compare results of indirect comparisons with head-to-head data if available, and draw conclusions from indirect comparisons cautiously.

Combining Data from Different Types of Studies

Most CERs will include data on harms from different types of studies. Statistical combination of data from observational studies is often inappropriate and should be avoided unless there is a clear rationale to do so.[81] If such analyses are undertaken, the justification should be clearly explained.

Discrepancies Between Randomized Trials and Observational Studies

A separate challenging situation occurs when results on harms from randomized trials and observational studies are discordant. Some reasons for discrepancies between randomized trials and observational studies are shown in Table 5. A reasoned analysis of potential sources of discrepancy is generally more helpful than simply presenting the different results.

Table 5. Sources of discrepancy between randomized controlled trials and observational studies

Differences in risk of bias (study quality)
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
Differential effects related to funding source (observational studies less likely to be funded by industry)

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.[2] A critical role of CERs is to report clearly on the limitations of the evidence on harms and to analyze and interpret thoughtfully how these limitations may affect

estimates of the balance of benefit and harm. Suggested elements to focus on when reporting harms are shown in Table 6.

Table 6. Elements to report when describing results for harms in Comparative Effectiveness Reviews

Element	Factors
Risk of bias (quality)	Study design, number of studies, study quality, consistency of evidence, directness of evidence, other modifying factors
Applicability	Population characteristics, interventions, co-interventions, comparisons, outcomes, duration of followup for various harms
Results	Number of patients, absolute and relative estimates of risks
Publication bias or incomplete outcomes data	Graphic and/or statistical assessments for publication bias, known unpublished studies, number of studies not reporting key harms
Additional analyses	Sensitivity analyses, subgroup analyses, metaregression, etc.

Summary

A summary of the key points about assessment of harms discussed in this report is shown in Table 7.

Table 7. Summary of key points on assessment of harms in Comparative Effectiveness Reviews

Assess all important harms, whenever possible.
Use multiple sources of information, including clinical experts and stakeholders, to identify important harms.
Use consistent and precise terminology when reporting data on harms, and avoid terms implying causality unless causality is reasonably certain.
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.
Be cautious about drawing conclusions on harms when events are rare and estimates of risk are imprecise.
Include placebo-controlled trials, particularly for assessing uncommon or rare harms, but be cautious about relying on indirect comparisons to judge comparative risks, and evaluate whether studies being considered for indirect comparisons meet assumptions for consistency of treatment effects.
Avoid inappropriate combining of data on harms, and thoroughly investigate inconsistent results.

Acknowledgments

The authors would like to acknowledge Gail R. Janes for participating in the workgroup calls.

This paper has also been published in edited form: Chou R, Aronson N, Atkins D, et al. AHRQ Series Paper 4: Assessing harms when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2010;63:502–512.

Author Affiliations

Oregon Evidence-based Practice Center, Oregon Health & Science University, Portland, OR (RC, DS, EW). Blue Cross Blue Shield Evidence-based Practice Center, Blue Cross Blue Shield Association, Chicago, IL (NA). Department of Veterans Affairs, Washington, DC (DA). McMaster Evidence-based Practice Center, McMaster University, Hamilton, ON (ASI, PS). Oregon Evidence-based Practice Center, Kaiser Center for Health Research, Portland, OR (DHS, EW). Minnesota Evidence-based Practice Center, Minneapolis VA Center for Chronic Disease

Outcomes Research, MN (TJW). University of Ottawa Evidence-based Practice Center, University of Ottawa, Ottawa, ON (DM).

References

1. Lohr KN. Emerging methods in comparative effectiveness and safety: symposium overview and summary. Med Care 2007;45(10 Suppl 2):S5–S8.
2. GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.
3. McIntosh HM, Woolacott NF, Bagnall A.-M. Assessing harmful effects in systematic reviews. BMC Med Res Meth 2004;4:19.
4. Loke YK, Price D, Herxheimer A. Systematic reviews of adverse effects: framework for a structured approach. BMC Med Res Methodol 2007;7:32.
5. Ioannidis JPA, Evans SJW, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004;141(10):781–8.
6. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000;356(9237):1255–9.
7. Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001;285(4):437–43.
8. Loke Y, Derry S. Reporting of adverse drug reactions in randomised controled trials—a systematic survey. BMC Clin Pharmacol 2001;1:3.
9. Vandenbroucke JP. Benefits and harms of drug treatments. BMJ 2004;329(7456):2–3.
10. Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” Lancet 2005;365(9453):82–93.
11. Chou R, Fu R, Huffman LH, et al. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. Lancet 2006;368(9546):1503–15.
12. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003;326(7387):472.
13. Chan A, Hrobjartsson A, Haahr M, et al. Empirical evidence for selective reporting of outcomes in randomized trials. JAMA 2004;291(20):2457–65.
14. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337(8746):867–72.
15. McDonagh M, Helfand M, Carson S, et al. Hyperbaric oxygen therapy for traumatic brain injury: a systematic review of the evidence. Arch Phys Med Rehabil 2004;85(7):1198–1204.
16. Kearney PM, Baigent C, Godwin J, et al. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? Meta-analysis of randomized trials. BMJ 2006;332:1302–8.
17. Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7(1):1–76.
18. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008;358:252–60.
19. Whittington CJ, Kendall T, Fonagy P, et al. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. Lancet 2004;363(9418):1341–5.
20. Laine C, Goodman SN, Griswold ME, et al. Reproducible research: moving toward research the public can really trust. Ann Intern Med 2007;146(6):450–3.
21. Toma M, McAlister FA, Bialy L, et al. Transition from meeting abstract to full-length journal article for randomized controlled trials. JAMA 2006;295(11):1281–7.
22. Ridker PM, Torres J. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000-2005. JAMA 2006;295(19):2270–4.
23. Sterne JA, Egger M, Smith GD. Investigating and dealing with publication and other biases in meta-analysis. BMJ 2001;323(7304):101–5.

24. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group [see comment]. N Engl J Med 2000;343(21):1520–8.
25. Silverstein FE, Faich G, Goldstein JL, et al. 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 2000;284(10):1247–55.
26. Hrachovec JB, Mora M. Reporting of 6-month vs 12-month data in a clinical trial of celecoxib. JAMA 2001;286(19):2398.
27. Witter J. Medical review part 1. Center for Drug Evaluation and Research. Available at: http://www.fda.gov/cder/foi/nda/2002/20-998S009_Celebrex_medr_P1.pdf. Accessed April 3, 2008.
28. Furukawa TA, Watanabe N, Montori VM,et al. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses [letter]. JAMA 2007;297(5):468–70.
29. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004;363(9422):1728–31.
30. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573–7.
31. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. Belmont, CA: Wadsworth; 1982.
32. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007;147(8):W163–W194.
33. Psaty BM, Koepsell T, Lin D, et al. Assessment and control for confounding by indication in observational studies. J Am Geriatr Soc 1999;47(6):749–54.
34. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998.
35. Papanikolaou P, N, Christidi GD, Ioannidis J. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ 2006;174(5):635–41.
36. Gliklich R, Dreyer NA, eds. Registries for evaluating patient outcomes: a user's guide. AHRQ Publication NO. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality; 2007.
37. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323–37.
38. Sturmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration-a simulation study. Amer J Epidemiol 2007;165(10):1110–8.
39. Aronson JK, Derry S, Loke YK. Adverse drug reactions: keeping up to date. Fundam Clin Pharmacol 2002;16:49–56.
40. Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. BMJ 2004;329(7456):44–7.
41. Loke YK, Derry S, Aronson JK. A comparison of three different sources of data in assessing the frequencies if adverse reactions to amiodarone. Br J Clin Pharmacol 2004;57(5):616–21.
42. Strom BL. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: a counterpoint. JAMA 2004;292(21):2643–6.
43. Psaty BM, Furberg CD, Ray WA, Weiss NS. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis. JAMA 2004;292(21):2622–31.
44. Bennett CL, Nebeker JR, Lyons EA, et al. The Research on Adverse Drug Events and Reports (RADAR) project. JAMA 2005;293 (17):2131–40.
45. Oleson O. 2. Types of study design. The Cochrane Non-Randomised Studies Methods Group (NRSMG); 1999. Available at: http://www.cochrane.dk/nrsmg/docs/chap2.pdf. Accessed April 3, 2008.
46. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323(7303):42-46.
47. Bent S, Padula A, Avins AL. Brief communication: better ways to question patients about adverse medical events: a randomized, controlled trial. Ann Intern Med 2006;144(4):257-261.
48. NCI. Common Terminology Criteria for Adverse Events v3.0 (CTCAE); 2006. Available at: http://ctep.cancer.gov/reporting/ctc_v30.html. Accessed April 3, 2008.

49. NIAID. Division of AIDS table for grading the severity of adult and pediatric adverse events; 2004. Available at: http://www3.niaid.nih.gov/research/resources/DAIDSClinRsrch/Safety/. Accessed April 3, 2008.
50. Sydes MR, Spiegelhalter DJ, Altman DG, et al. Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. Clinical Trials 2004;1:60–79.
51. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66(5):688–701.
52. Rochon PA, Gurwitz JH, Sykora K, et al. Reader's guide to critical appraisal of cohort studies: 1. Role and design. BMJ 2005;330(7496):895–7.
53. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies Health Technol Assess 2003;7(27):iii–x, 1–173.
54. West S, King V, Carey TS. Systems to rate the strength of scientific evidence. Rockville, MD: Agency for Healthcare Research and Quality; 2002.
55. Dalziel K, Round A, Stein K, et al. Do the findings of case series studies vary significantly according to methodological characteristics? Health Technol Assessment 2005;9(2):1–146.
56. Rothwell PM, Slattery J, Warlow CP. A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis. Stroke 1996;27(2):260–5.
57. Juni P, Nartey L, Reichenbach S, et al. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet 2004;364(9450):2021–9.
58. Ofman JJ, MacLean CH, Straus WL, et al. A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs [see comment]. J Rheumatol 2002;29(4):804–12.
59. Shah RV, Albert TJ, Buegel-Sanchez V, et al. Industry support and correlation to study outcome for papers published in Spine. Spine 2005;30:1099–1104.
60. Laporte JR, Ibanez L, Vidal X, et al. Upper gastrointestinal bleeding associated with the use of NSAIDs: new versus older agents. Drug Saf 2004;27(6):411–20.
61. Martin RCG, Brennan MF, Jacques DP. Quality of complication reporting in the surgical literature. Ann Surg 2002;235:803–13.
62. Carey TS, Boden SD. A critical guide to case series reports. Spine 2003;28:1631-1634.
63. Chou R, Fu R, Carson S, et al. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. J Clin Epidemiol 2006;60(1):18–28.
64. Juni P, Witschi A, Bloch R,e t al. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282(11):1054–60.
65. Santaguida PL, Raina P. The Development of the McHarm Quality Assessment Scale for adverse events: Delphi Consensus on important criteria for evaluating harms. 2008. Available at: http://hiru.mcmaster.ca/epc/mcharm.pdf. Accessed May 14, 2008.
66. Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. BMJ 1982;284:249–52.
67. Loke YK, Price D, Derry S, et al. Case reports of suspected adverse drug reactions-systematic literature survey of follow-up. BMJ 2006;332(7537):335–9.
68. Begaud B, Moride Y, Tubert-Bitter P, et al. False-positives in spontaneous reporting: should we worry about them? Br J Clin Pharmacol 1994;38(5):401–4.
69. Danan G, Benichou C. 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 1993;46(11):1323–30.
70. Michel DJ, Knodel LC. Comparison of three algorithms used to evaluate adverse drug reactions. Am J Hosp Pharm 1986;43(7):1709–14.
71. Aronson JK. Anecdotes as evidence. BMJ 2003;326:1346.
72. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med 1997;127(9):820–6.
73. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352(9128):609–13.
74. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuses of power when interpreting results. Ann Intern Med 1994;121:200–6.
75. Jonville-Bera AP, Giraudeau B, Autret-Leca E. Reporting of drug tolerance in randomized clinical trials: when data conflict with authors' conclusions. Ann Intern Med 2006;144:306–7.

76. de Gans J, van de Beek D. Dexamethasone in adults with bacterial meningitis. N Engl J Med 2002;347:1549–56.
77. Piaggio G, Elbourne DR, Altman DG, et al. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006;295(10):1152–60.
78. Ware JH, Antman EM. Equivalence trials. N Engl J Med 1997;337(16):1159–61.
79. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50(6):683–91.
80. Zhang J, Ding EL, Song Y. Adverse effects of cyclooxygenase 2 inhibitors on renal and arrhythmia events: meta-analysis of randomized trials. JAMA 2006;296:1619–32.
81. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. BMJ 1998;316(7125):140–4.

Chapter 8. Expanded Guidance on Selected Quantitative Synthesis Topics

Joseph Lau, Norma Terrin, Rochelle Fu

Abstract

This report provides expanded guidance on several topics that originally appeared in Chapter 9 ("Conducting Quantitative Synthesis When Comparing Medical Interventions") of the 2007 draft "Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews." Selected topics from this chapter were posted on the Effective Health Care Program Web site after public comments and were also published as a journal manuscript. The topics in the current report were cut from the 2007 draft methods reference guide to make the currently posted quantitative synthesis document a manageable length. The current report complements the posted document and includes the following topics: combining a small number of studies, combining composite outcome, control rate meta-regression, and interpretation and translation of results of meta-analyses.

The first three topics of this report focus on whether meta-analyses should be conducted in the settings encountered and on the selection of appropriate methods should it be decided to carry out meta-analyses. The section on combining small number of studies provides the rationale for why meta-analyses of small number (two to four) of studies could be unreliable and gives guidance on performing meta-analyses that have only few studies. The section on combining composite outcome discusses the rationale for using composite outcomes as well as the potential for misinterpretation of clinical trials when such outcomes are used and provides guidance on carrying out the proper analyses and interpretation. The section on control rate meta-regression discusses settings in which heterogeneous treatment effects may be related to varying baseline risk. The proper method of performing control rate meta-regression is discussed. Finally, the section on interpretation and translation of results of meta-analyses provides practical guidance on interpreting meta-analysis results of binary and continuous outcomes, as well as time to event and count data. This report ends with a section that provides instructions for reporting of meta-analyses.

Background

This report provides expanded guidance on several topics that originally appeared in Chapter 9 ("Conducting Quantitative Synthesis When Comparing Medical Interventions") of the 2007 draft "Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews."[1] Selected topics from this chapter were posted on the Effective Health Care Program Web site[2] after public comments and were also published as a journal manuscript.[3] The topics in the current report were cut from the 2007 draft methods reference guide to make the currently posted quantitative synthesis document a manageable length. The current report complements the posted document and includes the following topics: combining a small number of studies, combining composite outcome, control rate meta-regression, and interpretation and translation of results of meta-analyses.

The first three topics of this report focus on whether meta-analyses should be conducted in the settings encountered and on the selection of appropriate methods should it be decided to

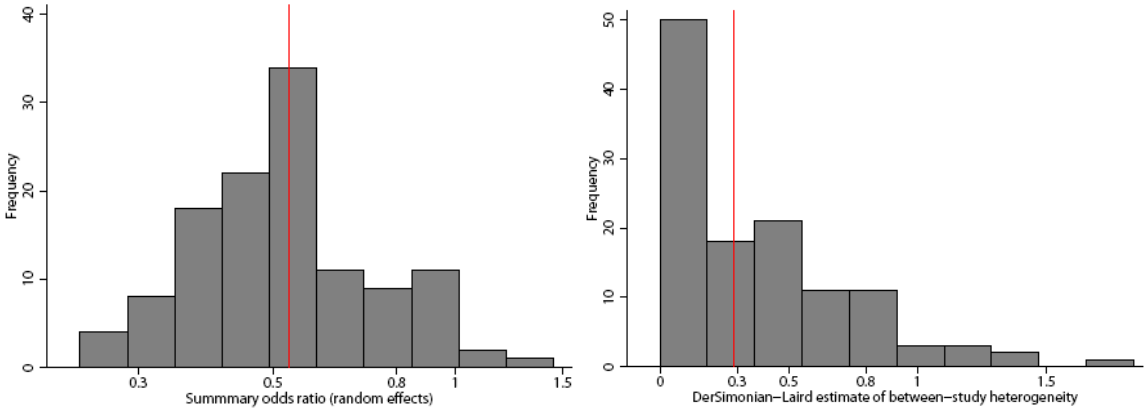
carry out meta-analyses. The section on combining small number of studies provides the rationale why meta-analyses of small number (two to four) of studies could be unreliable and gives guidance on performing meta-analyses that have only few studies. The section on combining composite outcome discusses the rationale for using composite outcomes as well as the potential for misinterpretation of clinical trials when such outcomes are used and provides guidance on carrying out the proper analyses and interpretation. The section on control rate meta-regression discusses settings in which heterogeneous treatment effects may be related to varying baseline risk. The proper method of performing control rate meta-regression is discussed. Finally, the section on interpretation and translation of results of meta-analyses provides practical guidance on interpreting meta-analysis results of binary and continuous outcomes, as well as time to event and count data. This report ends with a section that provides instructions for reporting of meta-analyses.

Combining a Small Number of Studies

There is no general rule for deciding the minimum number of studies in a meta-analysis, and it is possible to combine results even if there are only two studies. When interpreting the results, the precision of the studies is as relevant as the number of studies. Thus the meta-analysis of three "mega-trials" will be more reliable than the meta-analysis of three small trials, all other factors being equal. Therefore, determining whether to include studies in a meta-analysis will depend on the extent of their clinical and methodological diversity.

As an example of the hazards of relying on too little information, consider a meta-analysis of palliative chemotherapy versus supportive care and/or delayed chemotherapy for the treatment of advanced or metastatic colorectal cancer.[4] The summary random effects odds ratio for death within 12 months is 0.53 (95% CI: 0.34, 0.83) in favor of palliative chemotherapy (10 studies). The studies are statistically heterogeneous, with significant Q statistic and I^2=60 percent. Suppose that only 3 of the 10 studies had actually been completed, while the others never made it past the planning stages. The results would be quite different depending on which 3 studies had been completed. Figure 1 displays the wide variation in the summary odds ratio and estimated between-study variability among all 120 possible subsets of size 3.

Figure 1. Distributions of meta-analysis results (summary odds ratio and between-study variance) in subsets of 3 studies drawn from a set of 10.



Note: Vertical lines indicate the estimates for the entire set of 10 studies.

Statistical heterogeneity is difficult to infer when the total amount of information (sum of the precisions in the individual studies) is low,[5] although the I^2 statistic can be used as a descriptor[6] to help with the determination of whether to combine. Meta-regression should generally be avoided when there are few studies, because of low power.

Although it is not feasible to determine whether there is statistical heterogeneity among a small number of studies, a random effects model is preferred when heterogeneity is suspected. The classical random effects models (e.g., DerSimonian and Laird[7]) assume that the between-study variance is known, when actually it is estimated from the data. Hence the methods tend to underestimate the error associated with parameter estimates, particularly when the number of studies is small. The larger the true between-study variance, the less accurate the DerSimonian and Laird confidence limits.[8] A Bayesian meta-analysis with a vague prior distribution on the between study variance is recommended when there are too few studies to accurately estimate the between-study variance.[9]

One should consider whether the rarity of eligible studies is an indication of publication bias or selective outcome reporting, or whether the intervention being studied is novel and the specific scientific field is relatively new and immature. It is not unusual for estimates based on a handful of early studies to shift considerably over time as more studies are published on the same topic.[10] Thus, the interpretation of results should take into account the number of years since the first publication, or meta-analysis should be deferred until more studies are available.

In summary, when few studies (i.e., two to four) are available for meta-analysis:

- Clinical and methodological similarity should be taken into consideration when determining whether to combine them.
- Statistical heterogeneity is more difficult to address, but can be handled better with Bayesian random effects models than with classical methods.
- Meta-regression should be avoided.
- Interpretation should take into account the precision of the individual studies as well as the number of studies.
- Maturity of the field of investigation also needs to be considered.

Combining Composite Outcomes

A composite outcome can be binary (0/1) or time-to-event. If it is binary, it takes the value 1 if any of several possible events occurs. For example, “cardiovascular event (yes/no)” could be defined as a composite of MI, stroke, and death from cardiovascular disease. If a composite outcome is time-to-event, it takes the value of the time until the first event. Although the use of a composite outcome as the primary outcome in a clinical trial can reduce sample size requirements, that approach may lead to serious misinterpretation of the data. In meta-analysis, sample size is less of a concern than in clinical trials, and thus the motivation for using composite outcomes is diminished.

Statistical Efficiency in Clinical Trials

Composite outcomes can improve statistical efficiency, increasing power for a given sample size. If power is projected to be insufficient to analyze each of several outcomes separately while maintaining a low overall Type I error rate, investigators may be interested in using a composite as the primary outcome. Furthermore, composites have a larger number of events than the component outcomes, and thus they can increase the power for time-to-event

analyses and binary analyses of relative measures (odds ratio or relative risk). So composites can improve power both by handling the multiple testing problem and increasing the number of events. However, a composite can also reduce power by diluting component outcomes that are affected by the treatment with others that are not.[11] Composite outcomes may be considered in the context of clinical trials for which there are several relevant outcomes of similar clinical importance pertaining to the same disease process. In addition to homogeneity of clinical importance, there should be an expectation that the risk ratio of treatment benefit will be similar across the component outcomes.[11]

Misinterpretation in Clinical Trials

Composite outcomes pose a dilemma with regard to interpretation. For example, if an intervention results in a reduced risk for the composite of hospitalization and death, the intervention may have decreased hospitalizations while having no effect or a negative effect on survival. In reporting the result, it would be hard to avoid the suggestion of a reduction in mortality even if there was none. Requirements for meaningful composites include homogeneity of clinical importance as well as homogeneity of treatment benefit across the component outcomes.[12,13] Furthermore, the statement of the result should make clear the extent to which the component outcomes contributed to the finding. Empirical research found that in most clinical trials with composite endpoints, there was heterogeneity of clinical importance of the component outcomes; in about a third of the trials, results for the components were not reported; and in those trials that did report results for the components, more than half had heterogeneity of treatment effect.[14] Another review of composite outcomes found that only 60 percent of trials provided reliable estimates for both the composite and its components. The components were judged to be of similar importance in only 18 percent of trials. Indeed, death was the most important component in 83 percent of trials. Other problems included post hoc and inconsistent definitions of the composite.[15]

Example

In a trial that randomized 120 patients with in-stent stenosis of a saphenous vein graft to radiation or placebo, the composite outcome of death from cardiac causes, Q wave myocardial infarction, and revascularization of the target vessel, there were 43 events in the placebo arm compared with 22 for the intervention. Death and MI together accounted for 6 events in the placebo arm and 5 in the intervention arm. Thus, despite the composite outcome definition's inclusion of death and MI, the trial provided little information on these outcomes.[14,16]

Composite Outcomes in Meta-Analysis

Because of the large number of patients contributing data, systematic review diminishes and may eliminate the primary motivation for analyzing composites; that is, increasing statistical power. Furthermore, meta-analyses of the individual components of the composite yield more meaningful results. When a meta-analysis of a composite outcome is undertaken, trials without data for all component outcomes should be graded as having high risk of bias. Only composite outcomes that are generally agreed upon and in wide usage by the research community should be used in meta-analysis, and the meta-analyses of individual components should also be performed. Creating de novo composite outcomes without a precedent by the meta-analysts should be avoided. Statistical and clinical homogeneity of the components should be verified.

Summary

- Composite outcomes typically increase statistical efficiency.
- The additional power may not be necessary for many meta-analyses.
- Interpretation of composite outcome results is fraught.
- Only widely accepted composite outcomes should be used in meta-analysis.
- The components of the composite should be homogeneous with respect to clinical importance and magnitude of treatment benefit.
- For most composites used in clinical trials, there is heterogeneity of clinical importance across the components.
- Meta-analyses of the individual components should also be performed.

Control Rate Meta-Regressions

Patients with higher underlying risk for mortality and other outcomes may experience different benefits or harms from treatment than patients with lower underlying risk.[17] For studies with binary outcomes, the "control rate" refers to the proportion of subjects in the control group who experience the event. The control rate may be affected by disease severity, concomitant treatments, followup duration, as well as other factors that differ across studies,[18,19] and may thus be viewed as a study-level proxy for these factors. It is used to test for interaction between underlying population risk and treatment benefit, via control-rate meta-regression. However, advanced methods must be employed to obtain the correct level of statistical significance.

Even in the absence of a true linear relationship between treatment effect and control rate, the expected slope for the regression of treatment effect on control rate is non-zero. This bias is caused by measurement error in the control rate estimate and correlation between the control rate and treatment effect estimates.[20,21] Simple weighted regressions tend to identify a significant relation between control rate and treatment effect twice as often as more suitable approaches including hierarchical meta-regression models[19] and Bayesian meta-regressions.[21]

Thompson, Smith, and Sharp[21] illustrated the hazards of using a naïve meta-regression model to assess the relation between the control rate and mortality in a meta-analysis of the effectiveness of endoscopic sclerotherapy in patients with cirrhosis and esophagogastric varices.[22] The naïve approach estimated a statistically significant negative slope for the regression of odds ratio on control rate, implying that the higher the underlying risk, the more effective the treatment. In contrast, a Bayesian analysis that accounted for all sources of variability and correlation found a much weaker relation.[21]

The presence of a control rate effect varies according to the metric. The risk difference is more highly correlated with the control rate than is the relative risk or odds ratio and is constrained by the control rate particularly when the control rate is small. Schmid et al. demonstrated this empirically and showed that the relationship with the control rate is inflated using the risk difference metric.[19] In an empirical evaluation control rate effects were seen in 14 percent, 13 percent or 31 percent of 115 meta-analyses of binary outcomes when the measure of choice was the odds ratio, the risk ratio, or the risk difference, respectively.[19] The differences in the percentages between the relative (odds ratio, risk ratio) and the absolute (risk difference) metrics is related to the greater heterogeneity of the risk difference. For example, a risk ratio of 1.5 corresponds to very different risk differences at various levels of baseline risk (0.5 percent at 1 percent control rate, and 5 percent at 10 percent control rate).

A scatter plot of treatment effect against control rate is a useful ad hoc approach to visually assess whether there may be a relation between the two. A quick way to rule out the presence of a control rate effect is by a weighted regression of the effect size on the control rate. A negative finding would be most likely replicated by the more complicated methods; a positive finding would need to be verified by a more comprehensive method.

In summary, if the control group event rate is a plausible proxy for average within-study severity of illness of the study population, then:

- Consider a control rate meta-regression to explain between-study treatment effect heterogeneity.
- The use of a relative metric (risk ratio, odds ratio) is preferred in control rate meta-regression.
- Use a scatter plot to search for a systematic change in the effect size at different control rates.
- Use a simple weighted regression of the effect size on the control rate to rule out presence of a control rate effect; if the slope is significantly different than 0, advanced methods must be used to obtain the correct level of statistical significance.

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.

Binary Outcomes

Three effect measures could be used for binary outcomes in meta-analyses including risk difference (RD), relative risks (RR) and odds ratios (OR). It should be noted that there is no single perfect metric that is adequate in all settings. Each has its limitations and the proper interpretation requires additional data in order to fully inform the decision maker.

RD is generally considered as being most easily understood by clinicians and patients, and is the absolute difference in probabilities of an event between two intervention groups. Interpretation of RD is straightforward. For example, a RD of 5 percent between the intervention and placebo groups indicates that the risk of an event in the intervention group is 5 percent higher than the risk in the placebo group. Investigators should note that the clinical relevance of RD (as well as for RR and OR) 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 risk of an event from 78 percent to 76 percent. Therefore, when reporting a RD, the underlying event risks from each study should be reported as well, and investigators should comment on the clinical significance of the RD. Furthermore, the proportion of event for each intervention group usually increases with the increase of study duration and the estimated RD may increase accordingly. While it is not recommended to combine studies using RD when baseline risks are different among studies, when it is appropriate to combine RD, investigators should be clear about the length of followup periods of included studies. For example, for a group of studies with about 3 months' followup, the risk of an event in the intervention group in an average of 3 months is 5 percent higher than the risk in the placebo group.

RR (and OR) provide estimates that are less likely to vary over different populations and study durations, compared with RD. RR is interpreted as the ratio of probabilities of an event

between two intervention groups. Therefore, a RR of 2 means a twofold increased risk of an event in patients receiving 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, patients on tiotropium were twice as compliant as patients using ipratropium (RR: 2.0; 95% CI, 1.8–2.3).[23] Likewise, a meta-analysis of crystalline silica, subjects exposed to crystalline silica were shown to have a twofold incidence of lung cancer compared with those not exposed to crystalline silica (RR: 2.0, 95% CI, 1.8–2.3).[24]

Alternatively, investigators could present results as a relative risk reduction or relative risk increase, especially when the RR is below 2. For example, a CER on second-generation antidepressants compared discontinuation due to adverse events between venlafaxine and the class of selective serotonin reuptake inhibitors (SSRIs), and the combined RR was 1.36 (95% CI, 1.09–1.69).[25] This finding could be expressed as a relative risk increase, that is, venlafaxine had a 36 percent higher risk of causing discontinuation due to adverse events than SSRIs as a class. Similarly, if a combined RR is 0.74 to compare an intervention to the placebo, the finding could be interpreted as that the risk of the intervention is 36 percent less. However, investigators must be aware that the meaning of RR is not symmetric around 1. For example, the RR of 0.5 of dying is not the same as RR of 2 of not dying (living); whereas the OR calculation is valid.

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, and odds is not intuitive to communicate the magnitude of risk. Mathematically one could choose either the RR or OR metric in the analyses of data and their results would be similar when the event rates are low. Investigators should avoid the common misinterpretation of treating odds and odds ratios as risks and relative risks, especially when event risks are high (> 10%). This misinterpretation could lead to an overstatement of the actual effect size. For example, a survey designed to examine 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),[26] causing concerns in the media about gender disparities. Schwartz et al. reanalyzed the same data using RRs and found that the gender disparities is actually small (RR 1.07 95% CI 1.01–1.16).[27]

To facilitate interpretation when RR or OR is used, we recommend calculating a RD or number need to treat (NNT) or number needed to harm (NNH) and the corresponding 95% confidence interval using the combined estimates at typical proportions of events in the control group, to provide enough information for readers to assess the clinical relevance. For the above comparison between venlafaxine and SSRIs, given that a typical proportion of discontinuation is 8 percent for the SSRI group, the corresponding NNH to prevent one additional discontinuation is 35 (95% CI, 18–139).

NNTs and NNHs are frequently used because they portray the absolute effect of an intervention that is believed to be intuitive.[28] NNTs and NNHs themselves do not reflect variations attributable to underlying event rates; and they do not have a standardized unit of time. Therefore, when NNTs or NNHs are presented, investigators should report these measures with an appropriate time frame and make clear that they are based on an average estimate. For example, one correct interpretation of a NNT of 10 over 3 years could be that “On average, 10 patients would have to be treated for 3 years with treatment A to observe one fewer event after 3 years”.[29] A different and less used way to interpret a NNT (or NNH) would be as a treatment frequency. For example, a NNT of 100 could be presented as 10 in 1,000 treated people will benefit from treatment. If substantial variations in NNTs (NNHs) exist based on different event

rates, dosages, or subgroups, then investigators should report them separately for each group. However, the use of NNT and NNH is not universally recommended.[30] Empirical studies have questioned whether this metric is really intuitive to patients.[31]

Finally, the terms “risk difference” or “relative risk” themselves can be confusing however if they refer to a beneficial outcome. Investigators should avoid the use of “risk” when reporting beneficial outcomes. Instead, investigators could interpret the results in terms of the probability of the beneficial outcome directly. For example, if a meta-analysis produced a RD of 10 percent when combining studies comparing the effectiveness of a drug vs. placebo to achieve a 50 percent pain reduction, it could reported as that comparing with the placebo group, the probability of achieving a 50 percent pain reduction was 10 percent higher in the treatment group. When RR is used, 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.[32] For the outcome of being a responder, the result was reported as “suggested a modest additional treatment effect (relative benefit, 1.10 [95% CI, 1.01–1.22]) for sertraline compared with fluoxetine.”[32]

Continuous Outcomes

The weighted mean difference (WMD) and the standardized mean difference (SMD) or 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, and easily interpreted as the mean difference between two comparison groups. The summary effect has the same unit as the scale employed in the included studies. For example, in a meta-analysis of differences in points on the Montgomery-Asberg Depression Scale (MADRS) between escitalopram and citalopram,[25] the WMD was estimated to be 1.51 (95% CI, 0.58–2.45). This finding can be interpreted as escitalopram having an additional treatment effect of 1.51 points on the MADRS, or escitalopram having a 1.51 higher points on the MADRS. Although this finding was statistically significant, the clinical significance of a difference of 1.51 points must be determined independently.

Standardized mean difference or 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.[33] combined functional outcomes measured on different scales in placebo-controlled studies of Alzheimer’s drugs using standardized mean different and the combined estimate was 0.25 (95% CI 0.13, 0.37) for trials less than 24 weeks, and 0.29 (95% CI 0.22, 0.36) for trials more than 24 weeks. Although these results were interpreted as small based on the most widely used classification, where standardized effect sizes of 0.2, 0.5 and 0.8 are suggested corresponding to small, medium, and large referents,[34] the clinical significance of the additional treatment effect of Alzheimer’s drugs compared with placebo is difficult to determine. This is an inherent problem of using standardized mean difference where currently there is no better interpretation available. To facilitate interpretation, the investigators could consider calculating an approximation of mean difference on the included measurement scales by multiplying the standardized effect sizes by the combined standard deviation for each included scale.

Time to Event Data and Count Data

Hazard ratio (HR) is the measure typically used for time to event data. Interpretation of HR is similar to RR, so a HR of 2 could also be interpreted as a twofold increased risk of an event in patients receiving a treatment compared with those not receiving the treatment. However, there is a subtle difference between HR and RR where RR is a ratio of two probabilities and HR is a ratio of two hazard rates (instantaneous risk). Such distinction is not important for the clinical implication of the results and informing patients, clinicians and health policy makers. Rate ratio (RR) is the measure typically used for count data, and as the term indicates, the ratio of two rates. For a rate ratio of 2, it means the rate of an event in patients receiving a treatment is 2 times the rate of an event in patients not receiving the treatment. Similar to binary outcome, we recommend reporting both the event rate for each treatment arm and the rate ratio. The estimate of rate takes into account both the number of new cases, and followup time of population. Its interpretation depends on the selection of the time unit. For example, a rate of 0.097/person-years could be expressed as 0.008/person-months, or 97/1000 person-years. It is essential in presenting incidence rates with appropriate time units. For clarity, the numerator is often expressed as a power of 10.

Similar to relative risk, investigators should calculate NNT/NNH based on combined hazard ratio or rate ratio while incorporating the time frame associated with such calculations. Smeeth et al.[30] provided a good example, and calculated NNT with statins to prevent one cardiovascular event and mortality over 5 years. Although they combined studies to achieve a summary NNT, they also presented NNTs for individual studies with varying baseline risks (Table 1). The combined NNT to prevent one death was 20 over 5 years. NNTs of individual studies, however, ranged from 8 to 28 corresponding to different baseline risks. A 95% CI was provided for each NNT from the combined estimates, and we recommend providing a 95% CI for all estimates. A similar table could also be used for reporting relative risk and NNT from binary data with minor modifications. For example, for the column of baseline risk, it could be replaced with control rate (proportion of event in the control group) if the event rate is not available. If it is appropriate to use risk difference to combine data, the columns of rate ratio could be replaced by risk difference to report the results.

Table 1. Number needed to treat with statins to prevent one cardiovascular event in 5 years

Trials	Number 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/TexCAPS	6,605	0.1	1.04	1.36	0.69	167*	1,000*	28
WOSCOPS	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 ischemic disease	9,014	1.4	0.78	0.77	0.8	41	64	17
Combined 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)

Adapted by permission from BMJ Publishing Group Limited. BMJ. Smeeth L, Haines A, Ebrahim S, vol. 318, pp. 1548-51, 1999.
CHD = coronary heart disease; CV = cardiovascular
* AFCAPS/TexCAPS study reported a nonsignificant increased total and CHD mortality in the intervention group. Numbers needed to treat are derived from the lower limit of the 95% CIs of the risk differences in event rates to illustrate the lower limit within which the numbers might lie.
**No upper number needed to treat can be calculated as the upper 95% CI of pooled absolute risk difference is greater than zero. In these circumstances, the number needed to treat is a number needed to harm.

Key Points

- Investigators should present summary effects in a way that makes it easy for readers to interpret and apply these findings appropriately.
- Investigators should interpret results accordingly based on the type of measure and data.
- For binary outcomes, report underlying event rates along with the effect measure used in the meta-analysis.
- For binary outcomes, consider calculating number need to treat (NNT) or number needed to harm (NNH) and the corresponding 95% confidence interval to provide information for readers to assess the clinical relevance.
- For binary outcome, NNTs and NNHs should be interpreted as "on average" within a specific time frame. If NNTs (NNHs) differ substantially based on control event rates, dosages, or subgroups, they should be presented separately. A confidence interval should be presented for each NNT or NNH.
- If ORs are used in a meta-analysis, results should be interpreted in terms of odds. Only when the event rate is low (< 10%), the OR can be interpreted approximately in the same way as RRs.
- If meta-analysis using standardized effect sizes is performed, standard deviations could be used to convert standardized effect sizes back to a unit on a specific scale to facilitate the interpretation.
- For time to event data and count data, investigators should also calculate NNT/NNH while incorporating the timeframe associated with such calculations.

Instructions for Reporting the Quantitative Synthesis of Studies

The purpose of the following summary of headings (Tables 2 and 3) for reporting quantitative syntheses of studies is to ensure some degree of uniformity in how EPCs present CER methods and results. The summary is not entirely prescriptive because CERs do not have to include all headings at all times. Rather, 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, the section of the quantitative chapter that discusses the pertinent issues is identified.

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.

If the review does not touch upon a specific area, then no mention of the associated heading or subheading is necessary. If no meta-analysis is conducted, then the authors would not

have to include a heading about methods of combining studies. The exact titles of headings and subheadings are left to the discretion of authors.

Table 2. Summary of headings for reporting the quantitative synthesis of studies: methods section

Headings	Subheadings	Required Reporting
Rationale to combine	Clinical heterogeneity	Specify important clinical characteristics which may differ among studies (e.g., intervention, dosage, baseline disease severity, length of followup) 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. For example, for length of followup, define the range of lengths of included studies that could be combined in one meta-analysis.
	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	Outcome definitions	Specify whether outcome definitions or the way outcomes were measured differed among studies. Specify whether surrogate outcomes or composite endpoints were used. If observational studies are included, describe the definition and measurement of confounding factors/effect modifiers considered in the analyses of individual studies.
	Primary vs. secondary outcomes	Specify whether outcomes were primary or secondary outcomes in the original studies. Specify benefit and harm outcomes clearly.
	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	Study design	Specify what type of study designs are being combined (e.g., RCT [crossover, cluster randomized, factorial], observational [cohort, case-control, cross-sectional]).
	Rationale for inclusion of observational studies	If observational studies are included, then provide a rationale (e.g., to broaden generalizability, to examine longer followup periods, inadequate data from RCTs, etc.).
Explanation of choice of effect measure		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 difference). 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.

Table 2. Summary of headings for reporting the quantitative synthesis of studies: methods section (continued)

Headings	Subheadings	Required Reporting
Methods for combining study estimates	Statistical procedure and justification of model chosen	First specify whether direct or indirect comparisons are being made.
	Direct comparison	For direct comparison, describe and justify the statistical model used to combine effect measures (e.g., random effects model, fixed effects model, Bayesian model).
	Indirect comparison	If indirect comparisons are being made, clearly state the rationale. Describe the methods used for indirect comparison and specify the analyses done to ensure the validity and robustness of results from indirect comparison.
	Special considerations	For rare binary outcome, describe and justify the statistical methods used.
Statistical heterogeneity	Statistical tests	Specify how statistical heterogeneity is assessed and the criteria used to identify "important" heterogeneity.
	Quantifying heterogeneity	Specify methods used to quantify statistical heterogeneity (e.g., I^2).
	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 prespecified and post hoc analysis. The exploratory nature of these analyses should be clear.
Sensitivity analyses		Specify what sensitivity analyses are being done and how they relate to key decisions and assumptions made in the systematic review.

Table 3. Summary of headings for reporting the quantitative synthesis of studies: results section

Headings	Recommendations
Descriptive study information	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 also 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 combined 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 (or represented by the size of plot symbol), 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.
Reporting of individual and combined study estimates	Provide interpretation for the individual and combined study estimates based on the type of data and choice of effect measure. Provide interpretation for results from test and exploration of heterogeneity. If additional analyses were conducted (e.g., sensitivity analysis), report the results of all additional analyses undertaken.

Author Affiliations

Tufts Evidence-based Practice Center (JL, NT). Oregon Evidence-based Practice Center (RF).

References

1. Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews, Version 1.0 [Draft posted Oct. 2007]. Rockville, MD; Agency for Healthcare Research and Quality. http://effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf. Accessed September 23, 2012.
2. Fu R, Gartlehner G, Grant M, et al. Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program. In: Methods Guide for Comparative Effectiveness Reviews [posted October 2010]. AHRQ Publication No. 10(12)-EHC063-EF. Rockville, MD; Agency for Healthcare Research and Quality. Chapters available at: http://effectivehealthcare.ahrq.gov/.
3. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011; 64:1187-97.
4. Palliative chemotherapy for advanced or metastatic colorectal cancer. Colorectal Meta-analysis Collaboration. Cochrane Database Sys Rev. 2000;(2):CD001545.
5. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17:841-56.
6. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539-58.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-88.
8. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20:825-40.
9. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Meth Med Res. 2001;10:277-303.
10. Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. J Clin Epidemiol. 2004;57:1124-30.
11. Freemantle N, Calvert M, Wood J, et al. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003; 289:2554-9.
12. Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, et al. Validity of composite end points in clinical trials. BMJ. 2005; 330:594-6.
13. Pogue J, Thabane L, Devereaux PJ, et al. Testing for heterogeneity among the components of a binary composite outcome in a clinical trial. BMC Medical Research Methodology. 2010;10:49.
14. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ. 2007; 334:786.
15. Cordoba G, Schwartz L, Woloshin S, et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ. 2010; 341:c3920.
16. Waksman R, Ajani AE, White RL, et al. Intravascular gamma radiation for in-stent restenosis in saphenous-vein bypass grafts. N Eng J Med. 2002; 346:1194-9.
17. Glasziou PP, Irwig LM. An evidence based approach to individualizing treatment. BMJ. 1995;311:1356-9.
18. Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. Lancet. 1998;351:123-7.
19. Schmid CH, Lau J, McIntosh M, et al. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med. 1998;17:1923-42.
20. McIntosh M. The population risk as an exploratory variable in research synthesis of clinical trials. Stat Med 1997;15:1713-28.
21. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med. 1997;16:2741-58.
22. Pagliaro L, D'Amico G, Sorenson TIA, et al. Prevention of first bleeding in cirrhosis a meta-analysis of randomized trials of nonsurgical treatment. Ann Intern Med. 1992;117, 59:70.

23. Breekveldt-Postma NS, Koerselman J, Erkens JA, et al. Enhanced persistence with tiotropium compared with other respiratory drugs in COPD. Resp Med. 2007;101:1398-405.
24. Smith AH. Lopipero PA. Barroga VR. Meta-analysis of studies of lung cancer among silicotics. Epidemiol. 1995;6:617-24.
25. Gartlehner G, Gaynes BN, Hansen RA, et al. Comparative benefits and harms of second-generation antidepressants: background paper for the American College of Physicians. Ann Intern Med. 2008;149:734-50.
26. Schulman KA, Berlin JA, Harless W, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med. 1999; 340:618-26.
27. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effect of race and sex on physicians' referrals for cardiac catheterization. N Engl J Med. 1999; 341:279-83.
28. Laupacis A, Sackett DL, Roberts RS, An assessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988; 318:1728-33.
29. Hutton JL. Number needed to treat: properties and problems. J Royal Statist Soc A. 2000;163:403-19.
30. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. BMJ. 1999;318:1548-51.
31. Sheridan SL, Pignone MP, Lewis CL. A randomized comparison of patients understanding of number needed to treat and other common risk reduction formats. J Gen Intern Med. 2003;18:884-92.
32. Hansen RA, Gartlehner G, Lohr KN, et al. Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. Ann Intern Med. 2005;143:415-26.
33. Hansen RA, Gartlehner G, Lohr KN, et al. Functional outcomes of drug treatment in Alzheimer's disease: a systematic review and meta-analysis. Drugs Aging. 2007;24:155-67.
34. Cohen J. Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.

Chapter 9. Conducting Quantitative Synthesis When Comparing Medical Interventions

Rongwei Fu, Gerald Gartlehner, Mark Grant, Tatyana Shamliyan, Art Sedrakyan, Timothy J. Wilt, Lauren Griffith, Mark Oremus, Parminder Raina, Afisi Ismaila, Pasqualina Santaguida, Joseph Lau, Thomas A. Trikalinos

Introduction

Comparative effectiveness reviews (CERs) are systematic reviews that summarize comparative effectiveness and harms of alternative clinical options, and aim to help clinicians, policy makers, and patients make informed treatment choices. Quantitative synthesis, or meta-analysis, is often essential for CERs to provide scientifically rigorous summary information. Quantitative synthesis should be conducted in a transparent and consistent way, and methodologies reported explicitly. Reasons for this were made clear during the controversy around the safety of rosiglitazone, where a systematic review that found increased risk for myocardial infarction[1] spurred heated debate on issues around choosing appropriate methods for quantitative syntheses;[2-4] and the subsequent Congressional hearing[5] brought these issues further into spotlight. This story highlighted the fact that basic issues in quantitative syntheses, such as choice of an effect measure or a model or how to handle heterogeneity, remain crucial considerations and are often the subject of controversy and debate.

A CER typically evaluates the evidence on multiple alternative interventions whereas most published meta-analyses compared one intervention with a placebo. Inclusion of multiple interventions increases the complexity of quantitative synthesis and entails methods of comparing multiple interventions simultaneously. Evaluation of multiple interventions also makes the assessment of similarity among studies and the decision to combine studies even more challenging. Presenting results of a meta-analysis from a CER in a way that is useful to decisionmakers is also a challenge.

The Evidence-based Practice Center (EPC) program of the Agency for Healthcare Research and Quality (AHRQ)[6] is the leading U.S. program providing unbiased and independent CERs. The goal of this article is to summarize our recommendations in conducting quantitative synthesis of CERs for therapeutic benefits and harms for the EPC program with the goal to improve consistency and transparency. The recommendations cover recurrent issues in the EPC program and we focus on methods for combining study-level effect measures. First, we discuss considerations for deciding whether to combine studies, followed by discussions on indirect comparison and incorporation of indirect evidence. Then we describe our recommendations for choosing effect measures and statistical models, giving special attention to combining studies with rare events; and on testing and exploring heterogeneity. Finally, we briefly present recommendations on combining studies of mixed design and on sensitivity analysis. This article is not a comprehensive review of methods.

The recommendations were developed using group discussion and consensus based on current knowledge in the literature.[7] EPC investigators are encouraged to follow these recommendations but may choose to use alternative methods if deemed appropriate. If alternative methods are used, the investigators are required to provide rationales for their choice,

and if appropriate, to state the strengths and limitations of the chosen method in order to promote consistency and transparency. In addition, several steps in conducting a meta-analysis require subjective decisions, for example, the decision to combine studies or the decision to incorporate indirect evidence. For each subjective decision, investigators should fully explain how the decision was reached.

Decision To Combine Studies

The decision to combine studies to produce an overall estimate should depend on whether a meaningful answer to a well formulated research question can be obtained. The purpose of a meta-analysis should be explicitly stated in the methods section of the CER. The overall purpose of the review is not in itself a justification for conducting a meta-analysis, nor is the existence of a group of studies that address the same treatments. Investigators should avoid statements such as "We conducted a meta-analysis to obtain a combined estimate of…" Rather, explain the reason a combined estimate might be useful to decision makers who might use the report or products derived from the report.

Study Similarity Is a Requirement for Quantitative Synthesis

Combining studies should only be considered if they are clinically and methodologically similar. There is no commonly accepted standard defining which studies are "similar enough." Instead, the similarity of selected studies is always interpreted in the context of the research question, and to some extent, is subjective. In addition, judging similarity among studies depends on the scope of the research question. A general question may allow inclusion of a broader selection of studies than a focused question. For example, it may be appropriate to combine studies from a class of drugs instead of limiting only to a particular drug, if the effect of the drug class is of interest, and the included studies are methodologically comparable.

Statistical Heterogeneity Does Not Dictate Whether or Not To Combine

Variation among studies can be described as[8]:

Clinical diversity. Variability in study population characteristics, interventions and outcome ascertainments.

Methodological diversity. Variability in study design, conduct and quality, such as blinding and concealment of allocation.

Statistical heterogeneity. Variability in observed treatment effects across studies. Clinical and/or methodological diversity, biases or even chance, can cause statistical heterogeneity.

Investigators should base decisions about combining studies on thorough investigations of clinical and methodological diversity as well as variation in effect size. Both the direction and magnitude of effect estimates should be considered. These decisions require clinical insights as well as statistical expertise.

Clinical and methodological diversity among studies always exists even if a group of studies meet all inclusion criteria and seem to evaluate the same interventions in similar settings. Incomplete description of protocols, populations, and outcomes can make it impossible to assess clinical and methodological diversity among trials; nor does it always result in detectable statistical heterogeneity.[9] Further, evolving disease biology, evolving diagnostic criteria or

interventions, change in standard care, time-dependent care, difference in baseline risk, dose-dependent effects and other factors may cause seemingly similar studies to be different. For example, the evolution of HIV resistances makes the HIV population less comparable over time, while the effectiveness of the initial highly-active antiretroviral therapy improves rapidly over time. These increased the complexity in the evaluation of clinical and methodological diversity.

Statistical tests of heterogeneity are useful to identify variation among effects estimates, but their performance is influenced by number and size of studies[10] or choice of effect measures.[11] As a general rule, however, investigators should *not* decide whether to combine studies based on the p-value of a test of heterogeneity. When there is a large amount of clinical and methodological diversity along with high statistical heterogeneity such that any combined estimate is potentially misleading, the investigators should not combine the studies to produce an overall estimate. Instead, investigators should attempt to explore heterogeneity using subgroup analysis and meta-regression if there is sufficient number of studies (see section on Test and Explore Statistical Heterogeneity) or describe the heterogeneity qualitatively. However, combining clinically or methodologically diverse studies can make sense if effect sizes are similar, particularly when the power to detect variation is large. In this situation, investigators should describe the differences among the studies and population characteristics, as well as the rationale for combining them in light of these differences. Ultimately the decision will be judged on whether combining the studies makes sense clinically, a criterion that is qualitative and perhaps subjective. Examples to illustrate how to make appropriate decisions based on evaluation of different types of heterogeneity are helpful to guide the consistent implementation of these principles and need to be developed by the EPC program.

Indirect Comparisons and Consideration of Indirect Evidence

Multiple alternative interventions for a given condition usually constitute a network of treatments. In its simplest form, a network consists of three interventions, for example, interventions A, B, and C. Randomized controlled trials (RCT) of A vs. B provide direct evidence on the comparative effectiveness of A vs. B; trials of A vs. C and B vs. C would provide indirect estimates of A vs. B through the “common reference,” C. The inclusion of more interventions would form more complex networks and involve more complex indirect comparisons.[12,13]

Consideration of Indirect Evidence

Empirical explorations suggest that direct and indirect comparisons often agree,[13-18] but with notable exceptions.[19] In principle, the validity of indirect comparison relies on the invariance of treatment effects across study populations. However, in practice, trials can vary in numerous ways including population characteristics, interventions and cointerventions, length of followup, loss to followup, study quality, etc. Given the limited information in many publications and the inclusion of multiple treatments, the validity of indirect comparisons is often unverifiable. Moreover, indirect comparisons, like all other meta-analyses, essentially constitute an observational study, and residual confounding can always be present. Systematic differences in characteristics among trials in a network can bias indirect comparison results. In addition, all other considerations for meta-analyses, such as choice of effect measures or heterogeneity, also apply to indirect comparisons.

Therefore, in general, investigators should compare competing interventions based on direct evidence from head-to-head RCTs whenever possible. When head-to-head RCT data are

sparse or unavailable but indirect evidence is sufficient, investigators could consider indirect comparisons as an additional analytical tool.[20] If the investigators choose to ignore indirect evidence, they should explain why.

Approaches of Indirect Comparison

The naïve indirect comparison—where the summary event rate for each intervention is calculated for all studies and compared—is unacceptable. This method ignores the randomized nature of the data and is subject to a variety of confounding factors. Confounders will bias the estimate for the indirect comparison in an unpredictable direction with uncertain magnitude.[21]

An alternative approach of indirect comparison is to use qualitative assessments by comparing the point estimates and the overlap of confidence intervals from direct comparisons. Two treatments are suggested to have comparable effectiveness if their direct effects versus a common intervention have the same direction and magnitude, and there is considerable overlap in their confidence intervals. Under this situation, the qualitative indirect comparison is useful by saving the resources of going through formal testing and more informative than simply stating that there is no available direct evidence. However, the degree of overlap is not a reliable substitute for formal testing. It is possible that the difference between two treatment effects is significant when there is small overlap of confidence intervals. When overlap in confidence intervals is less than modest and a significant difference is suspected, we recommend formal testing.

Indirect comparison methods range from Bucher's simple adjusted indirect comparisons[15] to more complex multi-treatment meta-analysis (MTM) models.[12,13,22,23] When there are only two sets of trials, say, A vs. C and B vs. C, Bucher's method should be enough to get the indirect estimate of A vs. B. More complex network needs more complex MTM models. Currently the investigators may choose any of the MTM models and further research is required to evaluate their comparative performance and the validity of the model assumptions in practice. However, whichever method the investigators choose, they should assess the invariance of treatment effects across studies and appropriateness of the chosen method on a case-by-case basis, paying special attention to comparability across different sets of trials. Investigators should explicitly state assumptions underlying indirect comparisons and conduct sensitivity analysis to check those assumptions. If the results are not robust, findings from indirect comparisons should be considered inconclusive. Interpretation of findings should explicitly address these limitations. Investigators should also note that simple adjusted indirect comparisons are generally underpowered, needing 4 times as many equally sized studies to achieve the same power as direct comparisons, and frequently lead to indeterminate results with wide confidence intervals.[15,17]

MTM models provide the ability to check and quantify consistency or coherence of evidence for complex networks.[12,13,22,23] Consistency or coherence describes the situation that direct and indirect evidence agrees with each other, and when the evidence of a network of interventions is consistent, investigators could combine direct and indirect evidence using MTM models. Conversely, they should refrain from combining multiple sources of evidence from an incoherent network where there are substantial differences between direct and indirect evidence. Investigators should make efforts to explain the differences between direct and indirect evidence based upon study characteristics, though little guidance and consensus exists on how to interpret the results.

Choice of Effect Measures

Effect measures quantify differences in outcomes, either effectiveness or harms, between treatments in trials (or exposure groups in observational studies). The choice of effect measures is first determined by the type of outcomes. For example, relative risk and odds ratio are used for a binary outcome and mean difference is for a continuous outcome. They could also be broadly classified into absolute measures—such as risk differences or mean differences—and relative measures—such as odds ratio or relative risk. The number needed to treat (NNT) or harm (NNH) may also be considered effect measures, though they are usually not considered for meta-analyses as the standard error is rarely calculated or reported and normal approximation does not apply to NNT and NNH.

Binary Outcomes

Three measures are routinely used in a meta-analysis: the relative risk (RR), odds ratio (OR), and risk difference (RD). Criteria used to compare these measures include consistency over a set of studies, statistical property, and interpretability.[24] No single measure excels in all criteria.

The RD is most easily understood by clinicians and patients, and most useful to aid decision making, though it tends to be less consistent than relative measures (RR and OR) across studies. It is a preferred measure whenever estimates of RD are similar across studies and appropriate to be combined. Usually in such cases, the proportions of events among control groups are relatively common and similar among studies. When events are rare, we don't recommend RD because combined estimates based on RD are often biased and have conservative confidence interval coverage and low statistical power.[25] When RD is not appropriate, RR is preferred over OR because it is easier to interpret clinically. RR and OR are effectively equivalent for rare events. However, RR is not a reversible measure in terms that if the definition of an outcome event and nonevent is switched, for example, from death to survival, the estimate of RR will be affected substantially and RR for death is not the reciprocal of RR for survival. The precision of the estimated RR would be affected, too. For RD and OR, such switch has no major consequence as OR for death is the reciprocal of OR for survival and the switch only changes the sign of RD. Therefore, while the definition of the outcome event needs to be consistent among the included studies when using any measure, the investigators should be particularly attentive to the definition of an outcome event when using a RR.

The reported measures or study design could prescribe the choice of effect measures. Case-control studies only allow the estimation of an OR. For observational studies, usually only relative measures are reported from a model adjusted for confounding variables. In another situation, when a subset of included studies only report, say, RR, without reporting raw data to calculate other measures, the choice could be determined by the reported measure in order to include all studies in the analysis.

To facilitate interpretation when a relative measure (RR or OR) is used, we recommend calculating a RD or NNT/NNH using the combined estimates at typical proportions of events in the control group. We also encourage the calculation of NNT/NNH when using RD. Investigators should calculate a confidence interval for NNT/NNH as well.[26,27]

Note that both absolute and relative effect measures convey important aspects of evidence. We consider it good practice to report the proportion of events from each intervention group in addition to the effect measure.

Continuous Outcomes

The two measures for continuous outcomes are mean difference and standardized effect sizes. The choice of effect measure is determined primarily by the scale of the available data. Investigators can combine mean differences if multiple trials report results using the same or similar scales. Standardized mean difference (SMD) is typically used when the outcome is measured using different scales. SMD is defined as the mean difference divided by a measure of within-group standard deviation and several estimators of SMD have been developed including Glass's Δ, Cohen's *d* and Hedge's *g*. Hedge also proposed an unbiased estimator of the population SMD.[28] Hedge's unbiased estimator should be used whenever possible; otherwise, Hedge's *g* is generally preferred over Cohen's *d* or Glass's Δ. Standardized mean differences of 0.3, 0.5 and 0.8 are suggested corresponding to small, medium, and large referents[29] and widely used, though they were not anchored in meaningful clinical context.

For some continuous outcomes, a meaningful clinically important change is often defined and patients achieving such change are considered as "responders."[30] Understanding the relationship between continuous effect measures and proportion of "response" is nascent and not straightforward. Further research is necessary and we currently recommend against inferring response rate from a combined mean difference.

Count Data and Time to Events

Rate ratio is used for count data and often estimated from a Poisson regression model. For time to event data, the measure is hazard ratio (HR), and most commonly estimated from the Cox proportional hazards model. Investigators can also calculate HR and its variance if observed and expected events can be extracted,[31, 32] although this is often quite difficult.[33]

Choice of Statistical Model for Combining Studies

Meta-analysis can be performed using either a fixed or a random effects model. A fixed effects model assumes that there is one single treatment effect across studies. Generally, a fixed effects model is not advised in the presence of significant heterogeneity. In practice, clinical and methodological diversity are always present across a set of included studies. Variation among studies is inevitable whether or not the test of heterogeneity detects it. Therefore, we recommend random effects models, with exceptions for rare binary outcomes (discussed in more details under Combining Rare Binary Outcomes). We recommend against choosing a statistical model based on the significance level of heterogeneity test, for example, picking a fixed effect model when the p-value for heterogeneity is more than 0.10 and a random effects model when $P < 0.10$.

A random effects model usually assumes that the treatment effects across studies follow a normal distribution, though the validity of this assumption may be difficult to verify, especially when the number of studies is small. When the results of small studies are systematically different from those of the large ones, the normality assumption is not justified either. In this case, neither the random effects model nor the fixed effects model would provide an appropriate estimate[8] and we recommend not combining all studies. Investigators can choose to combine the large studies if they are well conducted with good quality and expected to provide unbiased effect estimates.

General Considerations for Model Choice

The most commonly used random effects model, originally proposed by DerSimonian and Laird,[34] does not adequately reflect the error associated with parameter estimation. A more general approach has been proposed.[35] Other estimates are derived by using simple or profile likelihood methods, which provide an estimate with better coverage probability.[36] Likelihood based random effects models also account better for the uncertainty in the estimate of between-study variance. All these models could be used to combine measures for continuous, count and time to event data, as well as binary data when the events are common. For OR, RR, HR and rate ratio, they should be analyzed on the logarithmic scale. For OR, a logistic random effects model is another option.[37] When the estimate of between-study heterogeneity is zero, a fixed effects model (e.g., the Mantel-Haenszel method, inverse variance method, Peto method (for OR), or fixed effects logistic regression) could also be used for common binary outcomes and provide similar estimate to the DerSimonian and Laird approach. Peto method requires that no substantial imbalance exists between treatment and control group sizes within trials and treatment effects are not exceptionally large.

A special case: combining rare binary outcomes. When comparing rare binary outcomes, few or zero events often occur in one or both arms in some of the included studies. The normal approximation of the binomial distribution does not hold well and choice of model becomes complicated. A fixed effects model is often more appropriate for rare events based on simulation study, even under the conditions of heterogeneity,[38] because it provides less biased results and better coverage property of the 95% confidence interval. However, investigators should note that no method gives completely unbiased estimates when events are rare.

When event rates are less than 1 percent, the Peto OR method is the recommended choice if the included studies have moderate effect sizes and the treatment and control group are of relatively similar sizes. This method provides the least biased, most powerful combined estimates with the best confidence interval coverage.[25] Otherwise when treatment and control group sizes are very different or effect sizes are large, or when events become more frequent (5 percent to10 percent), the Mantel-Haenszel method (without correction factor) or a fixed effects logistic regression provide better combined estimates and are recommended.

Exact methods have been proposed for small studies and sparse data.[39,40] However, simulation analyses did not identify a clear advantage of exact methods over a logistic regression or the Mantel-Haenszel method even in situations where the exact methods would theoretically be advantageous.[25] Therefore the investigators may choose to use exact methods but we don't specifically recommend exact methods over fixed effect models discussed above.

Considerations of correction factor for studies with zero events in one arm. In a study with zero events in one arm, estimation of effect measures (RR and OR) or their standard errors needs the addition of a correction factor, most commonly, 0.5 added to all cells. However, a combined estimate can be obtained using the Peto method, the Mantel-Haenszel method, or a logistic regression approach, without adding a correction factor. It has been shown that the Mantel-Haenszel method with the 0.5 correction does not perform as well as the uncorrected Mantel-Haenszel method or logistic regression,[25] nor as well as the Mantel-Haenszel method with alternative correction factors.[38] Therefore, we advise against the use of the Mantel-Haenszel method with the 0.5 correction. The investigators could choose adding no correction factors or exploring alternative correction factors using sensitivity analyses.[38]

Studies with zero events in both arms. When both arms have zero events, the relative measures (OR and RR) are not defined. These studies are usually excluded from the analysis as they do not provide information on the direction and magnitude of the effect size.[25,38] Others consider including studies without events in the analyses to be important and choose to include them using correction factors.[41,42] Inferential changes were observed when including studies without events[41] but the DerSimonian and Laird approach and RD[41] were used, which have been shown to have poor performance for rare events. [25]

We recommend that studies with zero events in both arms should be excluded from meta-analyses of OR and RR. The Peto method, fixed effects logistic regression (Bayesian or not), and the Mantel-Haenszel method effectively exclude these studies from the analysis by assigning them zero weight. Instead, the excluded studies could be qualitatively summarized, as in the hypothetical example below (Table 1), by providing information on the confidence intervals for the proportion of events in each arm. On the other hand, when the investigators estimate a combined control event rate, the zero events studies should be included and we recommend the random effects logistic model that directly models the binomial distribution.[43]

Table 1. Example of a qualitative summary of studies with no events in both groups

	Intervention A		**Intervention B**	
Studies with zero events in both arms	Counts	One sided 97.5% exact confidence interval for the proportion of events	Counts	One sided 97.5% 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)

Bayesian Methods

Both fixed and random effects models have been developed within a Bayesian framework for various types of outcomes. The Bayesian fixed effects model provides good estimates when events are rare for binary data.[38] When the prior distributions are vague, Bayesian estimates are usually similar to estimates using the above methods, though choice of vague priors could lead to a marked variation in the Bayesian estimate of between-study variance when the number of studies is small.[44] Bayesian random models properly account for the uncertainty in the estimate of between-study variance.

We support the use of Bayesian methods with vague priors in CERs, if the investigators choose Bayesian methods. The statistical packages such as WinBUGS provide the flexibility of fitting a wide range of Bayesian models.[45] The basic principle to guide the choice between a random effects and a fixed effect model is the same as that for the above non-Bayesian methods, though the Bayesian method needs more work in programming, simulation and simulation diagnostic.

Test and Explore Statistical Heterogeneity

Investigators should assess heterogeneity for each meta-analysis. Visual inspection of forest plots and cumulative meta-analysis plots[46] are useful in the initial assessment of statistical heterogeneity. A test for the presence of statistical heterogeneity, for example, Cochran's Q test, as well as a measure for magnitude of heterogeneity, e.g., the I^2 statistic,[11,47] is useful and should be reported. Further, interpretation of Q statistic should consider the limitations of the test that it

has low power when the number of studies is small and could detect unimportant heterogeneity when the number of studies is large. A p-value of 0.10 instead of 0.05 could be used to determine statistical significance. In addition, the 95% CI for I^2 statistic should also be provided, whenever possible, to reflect the uncertainty in the estimate.[48]

Investigators should explore statistical heterogeneity when present. Presentation and discussion of heterogeneity should distinguish between clinical, methodological and statistical heterogeneity when appropriate. Subgroup analysis or meta-regression with sensitivity analyses should be used to explore heterogeneity. When statistical heterogeneity is attributable to one or two "outlier" studies, sensitivity analyses could be conducted by excluding these studies. However, a clear and defensible rationale should be provided for identifying "outlier" studies. As discussed earlier, tests of statistical heterogeneity should not be the only consideration for the decision to combine studies or of the choice between a random or fixed effects model.

Subgroup analysis and meta-regression. Meta-regression models describe associations between the summary effects and study-level data, that is, it describes only *between-study*, not *between-patient*, variation. Subgroup analysis may be considered as a special case of meta-regression and involve comparison of subgroups of studies, for example, by study design, quality rating and other topic-specific factors such as disease severity. Investigators should note the difference between two types of study-level factors: (1) factors that apply equally to all patients in a study, e.g., study design, quality and definition of outcomes, and (2) study-level summary statistics of individual patient-level data, e.g., mean age, percentage of diabetic patients.[49-51] Meta-regression is most useful with the first type of study-level factors. A meta-regression on summarized patient-level factors may be subject to ecological fallacy,[51] a phenomenon in which associations present at the study level are not necessarily true at the patient level. Therefore, interpretation of meta-regression on summary data should be restricted to the study level.

We encourage the use of subgroup analysis and meta-regression to explore heterogeneity, to investigate the contribution of specific factors to heterogeneity and obtain combined estimates after adjusting for study level characteristics, when appropriate. A random effects meta-regression should always be used, to allow residual heterogeneity not explained by study level factors. Whenever possible, study level factors, including subgroup factors, considered in meta-regressions should be prespecified during the planning of the CER and laid out in the key questions, though the actual data may be known to some extent when the analyses are being planned for a meta-analysis. Variables that are expected to account for clinical or methodological diversity are typically included, e.g., differences in populations, or interventions, 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 variables. Use of permutation test for meta-regression can help assess the level of statistical significance of an observed meta-regression finding.[52]

When interpreting results, investigators should note that subgroup analyses and meta-regressions are observational in nature and suffer the limitations of any observational investigation, including possible bias through confounding by other study-level characteristics. As a general rule, association between effect size and the study-level variables (either pre- or post-specified) should be clinically plausible and supported by other external or indirect evidence, if they are to be convincing.

Number of studies required for a meta-regression. There is no universally accepted optimal minimum number of studies that are required for a meta-regression. The Cochrane handbook[8] suggests a minimum of 10 studies for each study-level variable without providing justifications, although fewer as six studies have been used in applied meta-regression empirical research.[50] The size of the studies and the distribution of subgroup variables are also important considerations. With the understanding that any recommended number has an arbitrary element, we advise a slightly different rule of thumb than the Cochrane handbook that when the sizes of the included studies are moderate or large, there should be at least 6 to 10 studies for a continuous study level variable; and for a (categorical) subgroup variable, each subgroup should have a minimum of 4 studies. These numbers serve as the lower bound for number of studies that investigators could start to consider a meta-regression. They are not the numbers that are sufficient for significant findings. The greater the number of studies, the more likely that clinically meaningful result is to be found. When the sizes of the included studies are small, it would take a substantial number of studies to produce useful results. When the number of studies is small, investigators should only consider one variable each time.

Combining studies of mixed designs. In principle, studies from different randomized trial designs, e.g. parallel, cross-over, factorial, or cluster-randomized design, may be combined in a single meta-analysis. Investigators should perform a comprehensive evaluation of clinical and methodological diversity and statistical heterogeneity to determine whether the trials should actually be combined, and consider any important differences between different types of trials. For cross-over trials, investigators should first evaluate whether the trial is appropriate for the intervention and medical condition in question. The risk of carryover and the adequacy of the washout period should be fully evaluated. Estimates accounted for within-individual correlation are best for meta-analysis. Similarly for cluster randomized trials, estimates accounted for intra-cluster correlation are best for meta-analysis. More discussion on combining studies of mixed randomized trial designs is provided in the online appendix.

In addition to randomized trials, CER also examines observational studies, especially for harms, adherence, and persistence.[53] Trial and observational evidence often agree in their results.[54-56] However, discrepancies are not infrequent.[57] Though there are several examples in the literature,[58,59] synthesis across observational and randomized designs is fraught with theoretical and practical concerns and much research is necessary to assess the consistency between clinical trials and observational studies and investigate the appropriateness of and develop statistical methods for such cross-design synthesis. Currently, we recommend against combining clinical trials and observational studies in the same meta-analysis.

Sensitivity Analyses

Completing a CER is a structured process. Investigators make decisions and assumptions in the process of conducting the review and meta-analysis; each of these decisions and assumptions may affect the main findings. Sensitivity analysis should always be conducted in a meta-analysis to investigate the robustness of the results in relation to these decisions and assumptions.[60] Results are robust if decisions and assumptions only lead to small changes in the estimates and do not affect the conclusions. Robust estimates provide more confidence in the findings in the review. When the results are not robust, investigators should employ alternative considerations. For example, if the combined estimate is not robust to quality rating, investigators should report both estimates including and excluding studies of lesser quality and

focus interpretation on estimates excluding studies of lesser quality. Investigators may also exclude studies of lesser quality.

Investigators should plan sensitivity analysis at the early stage of a CER, including tracking decisions and assumptions made along the way. Decisions and assumptions that might be considered in the sensitivity analysis include population or study characteristics, study quality and methodological diversity, choice of effect measures, assumptions of missing data, and so on. When necessary, multiple decisions and assumptions can be considered simultaneously.

Concluding Remarks

In this article, we provided our recommendations on important issues in meta-analyses to improve transparency and consistency in conducting CERs. The key points and recommendations for each covered issue are summarized in Table 2. Compared with the *Cochrane Handbook*, which explains meta-analysis methods in more detail, we focused on selected issues that present particular challenges in comparative effectiveness reviews. Overall there is no fundamental inconsistency between our recommendations and *Cochrane Handbook* on covered issues. We adopted the categorization of heterogeneity from the *Cochrane Handbook*, but provided more discussion of considerations for the decision to combine studies. For the choice of effect measures and statistical models, we favored RD and RR for binary outcome, and explicitly recommended random effects model except for rare binary outcome. Our recommendations and those of the *Cochrane Handbook* follow similar principles to test and explore heterogeneity though we proposed a slightly different rule on the number of studies adequate for meta-regression and distinguished between continuous vs. subgroup study level covariates.

Table 2. Summary of key points and recommendations for quantitative synthesis in Comparative Effectiveness Reviews

Decision to combine studies
The decision to combine studies should depend on whether a meaningful answer to a well formulated research question can be obtained.
Investigators should make decisions of combining studies based on thorough investigations of clinical and methodological diversity as well as variation in effect size.
Statistical tests of heterogeneity are helpful, but investigators should *not* make a decision on combining studies based *only* on tests of heterogeneity.
When there is a large amount of clinical and methodological diversity along with high statistical heterogeneity such that any combined estimate is potentially misleading, the investigators should not combine the studies.
Combining clinically or methodologically diverse studies may make sense if there is no real difference among effect sizes, particularly when the power to detect variation is large.
Reasons to combine or to not combine studies and steps taken to reach the decision should be fully explained.
The purpose of a meta-analysis should be explicitly stated in the methods section of the CER.
Indirect comparison
In the absence of sufficient direct head-to-head evidence and presence of sufficient indirect evidence, indirect comparisons can be considered as an additional analytic tool.
The unadjusted (naïve) indirect comparison method is not recommended in any case.
A qualitative indirect comparison may be useful to judge comparable effectiveness when there is a large degree of overlap in confidence intervals, but we recommend formal testing when significant difference is suspected.
Validity of the adjusted indirect comparison methods depends on the consistency of treatment effects across studies, and the appropriateness of an indirect comparison needs to be assessed on a case-by-case basis.
Adjusted indirect comparison methods, such as Bucher's method or mixed treatment comparison, should be used for indirect comparison.
Investigators should conduct sensitivity analysis to check the assumptions of the indirect comparison. If the results are not robust to the assumptions, findings from indirect comparisons should be considered as inconclusive.
Investigators should make efforts to explain the differences between direct and indirect evidence based upon study characteristics.

Table 2. Summary of key points and recommendations for quantitative synthesis in Comparative Effectiveness Reviews (continued)

Choice of effect measure
For dichotomous outcomes, RD is a preferred measure whenever appropriate. Otherwise, RR is preferred over OR.
A relative measure (RR or OR) instead of RD should be used when the events are rare.
When using a relative measure, risk differences and NNT/NNH should be calculated using the combined estimates at typical proportions of event in the control group. Calculation of NNT/NNH when using RD is also encouraged.
Calculation of NNT/NNH should include both point estimate and confidence interval.
Proportion of events from each intervention group should be reported in addition to the effect measure.
For continuous outcomes, mean difference should be used if results are reported using the same or similar scales and standardized mean difference should be used when results are reported in different scales.
For standardized mean difference, Hedge's unbiased estimator should be used whenever possible. Otherwise, Hedge's *g* is generally preferred over Cohen's *d* or Glass's Δ.
Rate ratio should be used for count data and hazard ratios for time-to-event data.
Choice of model
A random effects model is recommended since clinical and methodological diversity are inevitable among included studies.
A fixed effects model is recommended for rare binary events, and the choice of a fixed effects model depends on the event rate, effect size, and the balance of intervention groups.
For rare binary events: Studies with zero events in one arm should be included in the analyses. When event rates < 1%, the Peto OR method is recommended when no substantial imbalance exists between treatment and control group sizes within trials and treatment effects are not exceptionally large. In other situations, the Mantel-Haenszel method or a fixed effects logistic regression provides better combined estimates and are recommended. For the Mantel-Haenszel method, a correction factor of 0.5 is not recommended but using no correction factor or alternative correction factors could be considered, and investigated in sensitivity analyses when necessary. Studies with zero events in both arms should be excluded from the analyses but should be summarized qualitatively.
Use of Bayesian methods with vague priors in CERs is supported, if the investigators choose Bayesian methods.
Test and Explore Heterogeneity
Visual inspection of forest plots and cumulative meta-analysis plots are useful in the initial assessment of heterogeneity.
Heterogeneity should be assessed for each meta-analysis and both measures of the statistical significance and magnitude of heterogeneity should be reported.
Interpretation of statistical significance (for Q statistics) should consider the limitations of the test and the 95% CI for the estimate of magnitude of heterogeneity should be provided, whenever possible.
Presentation and discussion of heterogeneity should distinguish between clinical diversity, methodological diversity, and statistical heterogeneity when appropriate.
Heterogeneity should be explored using subgroup analysis or meta-regression or sensitivity analyses.
When heterogeneity is caused by one or two "outlier" studies, sensitivity analyses are recommended by excluding such studies.
Meta-regression (including subgroup analyses) is encouraged to explore heterogeneity.
Pre-specified meta-regression based on the key questions should be used to explore heterogeneity as much as possible.
A random effects meta-regression should be used.
Meta-regression is observational in nature, and if the results of meta-regression are to be considered valid, they should be clinically plausible and supported by other external or indirect evidence.
Combining Studies of Mixed Designs
If cross-over trials are appropriate for the intervention and medical condition in question, and there are no systematic differences between the two types of design, cross-over designs can be combined with parallel trials.
Meta-analysis of cross-over trials should use estimates from within-individual comparisons whenever available.
If cluster-randomization trials are appropriate for the intervention and medical condition in question, and there are no systematic differences between the different types of design, 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.
Clinical trials and observational studies should not be combined.
Sensitivity Analyses
A CER with a meta-analysis should always include sensitivity analyses to examine the robustness of the combined estimates in relation to decisions and assumptions made in the process of review.
Planning of sensitivity analysis should start at the early stage of a CER, and investigators should keep track of key decisions and assumptions.
When necessary, multiple decisions and assumptions may be considered at the same time.

This article does not address every major issue relevant to meta-analyses. Other interesting topics, such as meta-analysis of individual patient data, meta-analysis of diagnostic tests, assessing bias including publication bias, as well as more specific issues such as how to handle different comparators, composite outcomes or selective reporting will be considered in future versions of the EPC methods guide for CER. Meta-analysis methods for observational studies including combining observational studies, assessing bias for observational studies, incorporation of both clinical trials and observational studies, and even indirect comparison of observational studies will also be topics for both future version of guidelines and future research. As in most research areas, quantitative synthesis is a dynamic area with a lot of active research going on. Correspondingly, development of guidelines is an evolving process and we will update and improve recommendations with the accumulation of new research and improved methods to advance the goal for transparency and consistency.

Acknowledgements

This article was written with support from the Effective Health Care Program at the Agency for Healthcare Research and Quality (AHRQ).

The authors would like to acknowledge Susan Norris for participating in the workgroup calls and commenting on an earlier version of this manuscript, Ben Vandermeer for participating workgroup calls, Christopher Schmid for reviewing and commenting on the manuscript, Mark Helfand and Edwin Reid for editing the manuscript, and Brian Garvey for working on references and formatting the manuscript.

Author Affiliations

Oregon Evidence-based Practice Center, Department of Public Health and Preventive Medicine, Oregon Health & Science University, Portland, OR (RF). Danube University, Krems, Austria (GG). Technology Evaluation Center, Blue Cross Blue Shield Association (MG). Minnesota Evidence-based Practice Center, Division of Health Policy and Management, University of Minnesota, Minneapolis, MN (TS). Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD (AS). Minnesota Evidence-based Practice Center, Minneapolis VA Center for Chronic Disease Outcomes Research and the University of Minnesota Department of Medicine, Minneapolis, MN (TJW). McMaster Evidence-based Practice Center, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada (LG, MO, PR, AI, PS). Tufts Evidence-based Practice Center and Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA (JL, TAT).

This paper has also been published in edited form: Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2011 Nov;64(11):1187–1197.
PMID: 21477993.

References

1. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 2007;356:2457–71.
2. Dahabreh IJ, Economopoulos K. Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. Clin Trials 2008;5:116–20.
3. Diamond GA, Bax L, Kaul S. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. Ann Intern Med 2007;147:578–81.
4. Shuster JJ, Jones LS, Salmon DA. Fixed vs random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death. Stat Med 2007;26:4375–85.
5. Committee on Oversight and Government Reform. Hearing on FDA's Role in Evaluating Safety of Avandia. Available at: http://oversight.house.gov/index.php?option=com_content&view=article&id=3710&catid=44%3Alegislation&Itemid=1. Accessed May 31, 2010.
6. Agency for Healthcare Research and Quality. Evidence-based Practice Centers. Available at: www.ahrq.gov/clinic/epc/. Accessed May 31, 2010.
7. Helfand M, Balshem H. Principles for developing guidance: AHRQ and the effective health care program. J Clin Epidemiol 2010;63:484–90.
8. Higgins J. Cochrane handbook for systematic reviews of interventions. Available at: http://www.cochrane.org/resources/handbook/. Accessed May 31, 2010.
9. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
10. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med 1998;17:841–56.
11. Engels EA, Schmid CH, Terrin N, et al. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000;19:1707–28.
12. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105–24.
13. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313–24.
14. Baker SG, Kramer BS. The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C? BMC Med Res Methodol 2002;2:13.
15. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683–91.
16. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005;331:897–900.
17. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;9:1–148.
18. Song F, Glenny AM, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. Control Clin Trials 2000;21:488–97.
19. Chou R, Fu R, Huffman LH, et al. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. Lancet 2006;368:1503–15.
20. Ioannidis JP. Indirect comparisons: the mesh and mess of clinical trials. Lancet 2006;368:1470–2.
21. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003;326:472.
22. Dominici F, Parmigiani G, Wolpert R, et al. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. J Am Stat Assoc 1999;94:16–28.
23. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447–59.
24. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575–1600.
25. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007;26:53–77.
26. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452–4.

27. Schulzer M, Mancini GB. 'Unqualified success' and 'unmitigated failure': number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events. Int J Epidemiol 1996;25:704–12.
28. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. J Educ Stat 1981;6:107–28.
29. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: L. Erlbaum Associates; 1988.
30. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005;64:29–33.
31 Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998;17:2815–34.
32. Tierney J, Stewart L, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 2007;8:16.
33. Duchateau L, Collette L, Sylvester R, et al. Estimating number of events from the Kaplan-Meier curve for incorporation in a literature-based meta-analysis: what you don't see you can't get! Biometrics 2000;56:886–92.
34. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
35. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials 2007;28:105–14.
36. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001;20:825–40.
37. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 1995;14:2685–899.
38. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004;23:1351–75.
39. Mehta CR. The exact analysis of contingency tables in medical research. Cancer Treat Res 1995;75:177–202.
40. Mehta CR, Patel NR. Exact logistic regression: theory and examples. Stat Med 1995;14:2143–60.
41. Friedrich JO, Adhikari NK, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. BMC Med Res Methodol 2007;7:5.
42. Sankey S, Weissfeld L, Fine M, et al. An assessment of the use of the continuity correction for sparse data in meta-analysis. Communications in statistics—Simulation and computation 1996;25:1031–56.
43. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol 2008;61:41–51.
44. Lambert PC, Sutton AJ, Burton PR, et al. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med 2005;24:2401–28.
45. The BUGS Project. WinBUGS. Available at: http://www.mrc-bsu.cam.ac.uk/bugs/. Accessed May 31, 2010.
46. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48:45–57; discussion 9–60.
47. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
48. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007;335:914–6.
49. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet 1998;351:123–7.
50. Schmid CH, Lau J, McIntosh MW, et al. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med 1998;17:1923–42.
51. Schmid CH, Stark PC, Berlin JA, et al. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 2004;57:683–97.
52. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004;23:1663–82.
53. Slutsky J, Atkins D, Chang S, et al. Comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2008; in press.

54. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92.
55. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–86.
56. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286:821–30.
57. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–28.
58. Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. Int J Technol Assess Health Care 1993;9:440–9.
59. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. Stat Med 2000;19:3359–76.
60. Olkin I. Re: “A critical look at some popular meta-analytic methods”. Am J Epidemiol 1994;140:297-299; discussion 300–1.

Chapter 10. Grading the Strength of a Body of Evidence When Comparing Medical Interventions

Douglas K Owens, Kathleen N. Lohr, David Atkins, Jonathan R. Treadwell, James T. Reston, Eric B. Bass, Stephanie Chang, Mark Helfand

Key Points

- The EPC (Evidence-based Practice Center) approach is conceptually similar to the GRADE (Grading of Recommendations Assessment, Development and Evaluation) system of evidence rating.
- It requires assessment of four domains: risk of bias, consistency, directness, and precision.
- Additional domains to be used when appropriate include dose-response association, presence of confounders that would diminish an observed effect, strength of association, and publication bias.
- Strength of evidence receives a single grade: high, moderate, low, or insufficient.
- EPCs should grade strength of evidence separately for each major outcome and, for Comparative Effectiveness Reviews, all major comparisons.
- EPCs will collaborate with the GRADE group to address ongoing challenges in assessing the strength of evidence.

Introduction

Comparative Effectiveness Reviews (CERs), like systematic reviews in general, are essential tools for summarizing information to help make well-informed decisions about health care options.[1] CERs explicitly compare two or more screening or diagnostic strategies or therapeutic interventions. The Evidence-based Practice Center (EPC) program, supported by the U.S. Agency for Healthcare Research and Quality (AHRQ), produces substantial numbers of evidence reports and CERs. These reports are designed to accurately and transparently summarize a body of literature with the goal of helping clinicians, policymakers, and patients make well-informed decisions about health care. Reviews should provide clear judgments about the strength of the evidence that underlies conclusions to enable decisionmakers to use them effectively.[2]

In 2007, AHRQ supported a cross-EPC set of workgroups to develop guidance on major elements of designing, conducting, and reporting CERs.[3] This paper reports the outcomes of the EPC workgroup on grading strength of evidence. We briefly explore the rationale for grading strength of evidence, define the domains of concern for evidence strength, and describe our recommended grading system for such reviews. Our main objective was to give guidance to EPCs for grading strength of evidence in CERs, but this guidance may also apply to other systematic reviews.

The EPCs prepare reports that are used by a variety of decisionmakers, but they do not themselves develop recommendations. Therefore, the goal of our evidence rating system is to facilitate use of the reports by decisionmakers who may have differing perspectives. This separation of the raters of the strength of evidence from the decisionmakers led to some

differences in the system we propose relative to other rating systems that are designed to be used directly by decisionmakers.

The EPC approach is based in large measure on the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group approach.[4-6] We briefly discuss the differences in emphasis between the two systems. EPC and GRADE experts will explore ways to harmonize the two methods and to offer reviewers and decisionmakers a coordinated model for grading strength of evidence. This paper presents the approach that EPCs are expected to implement for CERs in the meantime.

Strength of Evidence: Rationale

Among organizations that make practice guidelines or coverage decisions and among experts who develop systematic reviews, assessment of the strength of a body of evidence is widely accepted. In drawing conclusions about strength of evidence, a growing number of organizations adopt systematic approaches to making judgments about the strength of evidence. A wide variety of grading systems is available for this purpose,[7] 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 of practical help to some organizations.

We note the important distinction between strength-of-evidence systems and evidence hierarchies. Evidence hierarchies traditionally focus only on study design, with systematic reviews of randomized controlled trials (RCTs) and individual RCTs at the highest levels. By contrast, strength-of-evidence systems incorporate not only study design but also many other facets of the evidence, including study conduct, presence or absence of bias, quantity of evidence, directness (or indirectness) of evidence, consistency of evidence, and precision of estimates. By including these additional components in our approach, we have attempted to give decisionmakers a more comprehensive evaluation of the evidence.

The aims of this work are to ensure appropriate methodologic consistency in how different EPCs grade the strength of evidence and to facilitate users' interpretations of those grades and how they apply them in guideline development or other decisionmaking tasks. Attaining these goals 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 reviews more useful to a broader range of potential audiences.

Strength of Evidence: Domains

The EPC approach to grading evidence begins with assessments of a set of agreed-upon domains pertaining to entire bodies of evidence about major outcomes (benefits and harms) and comparisons—i.e., outcomes and comparisons that are most important to decisionmakers in clinical practice and health policy. A determination of which outcomes and comparisons the EPCs consider important enough to warrant formal grading of the strength of the evidence will depend on the key questions, the clinical or policy context, and the purpose of the report. Major outcomes may include mortality, health-related quality of life, costs, potential harms, and for some reviews, intermediate end points or surrogate markers (for example, blood pressure control or cholesterol levels).

The four major domains are risk of bias, consistency, directness, and precision of the evidence. In selecting these domains, we reviewed work by the U.S. Preventive Services Task Force,[8] the GRADE Working Group[4] (http://www.gradeworkinggroup.org/), and other research by EPCs.[7,9] EPC reviewers aggregate judgments about the strength of evidence with respect to the domains into an overall evidence grade (explained below) for each major outcome. Tables 1 and 2 present two sets of domains: "required" and "additional," respectively. Because the strength of evidence may vary between key questions in a systematic review and among comparisons within a key question, the EPC should evaluate the strength of evidence separately for each important comparison for each key question.

Table 1. Required domains and their definitions

Domain	Definition and elements	Score and application
Risk of bias	Risk of bias is the degree to which the included studies for a 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 studies) Aggregate quality of the studies under consideration. Information for this determination comes from the rating of quality (good/fair/poor) done for individual studies	Use one of three levels of aggregate risk of bias: Low risk of bias Medium risk of bias High risk of bias
Consistency	The principal definition of consistency is the degree to which reported effect sizes from included studies appear to have the same direction of effect. 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 three levels of consistency: Consistent (i.e., no inconsistency) Inconsistent Unknown or not applicable (e.g., single study) As noted in the text, single-study evidence bases (even mega-trials) cannot be judged with respect to consistency. In that instance, use "consistency unknown (single study)."
Directness	The rating of directness relates to whether the evidence links the interventions directly to health outcomes. For a comparison of two treatments, directness implies that head-to-head trials measure the most important health or ultimate outcomes. Two types of directness, which can coexist, may be of concern: 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 the intermediate to most important (health or ultimate) outcomes. It uses two or more bodies of evidence to compare interventions A and B—e.g., studies of A vs. placebo and B vs. placebo, or studies of A vs. C and B vs. C but not A vs. B. 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 outcomes involved when assessing this domain.	Score dichotomously as one of two levels of directness: Direct Indirect If indirect, specify which of the two types of indirectness accounts for the rating (or both, if that is the case)—namely, use of intermediate/surrogate outcomes rather than health outcomes and use of indirect comparisons. 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 one of 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, results may be statistically compatible with both clinically important superiority and inferiority (i.e., the direction of effect is unknown), a circumstance that will preclude a valid conclusion.

Abbreviations: EPC= Evidence-based Practice Center; RCT=randomized controlled trial.

Table 2. Additional domains and their definitions

Domain	Definition and elements	Score and application
Dose-response association	This association, either across or within studies, refers to a pattern of a larger effect with greater exposure (dose, duration, adherence).	This additional domain should be rated if studies in the evidence base have noted levels of exposure. Use one of 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)
Plausible confounding that would decrease observed effect	Occasionally, in an observational study, plausible confounding factors would work in the direction *opposite* that of the observed effect. Had these confounders not been present, the observed effect would have been even larger than the one observed. In such a case, an EPC may wish to upgrade the level of evidence.	This additional domain should be considered if plausible confounding exists that would decrease the observed effect. Use one of two levels: Present: Confounding factors that would decrease the observed effect may be present Absent: Confounding factors that would decrease the observed effect are not likely to be present
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.	This additional domain should be considered if the effect size is particularly large. Use one of two levels: Strong: Large effect size that is unlikely to have occurred in the absence of a true effect of the intervention Weak: Small enough effect size that it could have occurred solely as a result of bias from confounding factors
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. The finding that only a small proportion of relevant trials (or other studies) 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.	Publication bias need not be formally scored. However, it can influence ratings of consistency, precision, magnitude of effect, and, to a lesser degree, risk of bias and directness. If EPCs identify unpublished trials and if the results differ from those of published studies, they can take these factors into account in their rating for consistency and in calculating a summary confidence interval for an effect. We encourage authors to comment on publication bias when circumstances suggest that relevant empirical findings, particularly negative or no-difference findings, have not been published or are not otherwise available.

Abbreviation: EPC=Evidence-based Practice Center

Required Domains

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

Before assessing the required domains, EPCs should first identify the studies that address the outcomes and comparisons of interest. When no study is available for an outcome or comparison of interest, the evidence should be graded simply as insufficient.

For the remaining major outcomes and comparisons of interest, the strength-of-evidence grade will depend on the required domains. EPCs have decided that focusing on consistency,

directness, and precision is more informative than emphasizing just the number of studies. Nevertheless, for CERs, EPCs should record the numbers of studies both in total and for specific comparisons. They should also indicate the numbers of studies that form the basis of given findings or conclusions. In this way, readers can better understand the available evidence for any given outcome or comparison.

Risk of Bias

As noted in Table 1, the risk of bias for an evidence base will be derived from assessment of the risk of bias in individual studies. Risk of bias incorporates both study design and study conduct. For strength-of-evidence grading, this domain requires reviewers to assess the aggregate quality of studies within each major study design and integrate those assessments into an overall risk-of-bias score.

Scores are denoted high, medium, or low. High risk of bias lowers the strength-of-evidence grade; low risk of bias raises it. If studies included in a systematic review differ substantially in risk of bias, EPCs may give greater weight or emphasis to the studies with a lower risk of bias. In formal meta-analyses, EPCs may choose to evaluate the influence of studies with differing risk of bias to aid in their assessment of the overall strength of evidence.

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 if the range of effect sizes is narrow, an evidence base can be judged to be consistent. This assessment enhances the overall strength-of-evidence grade. Nonoverlapping confidence intervals, significant unexplained clinical or statistical heterogeneity, or similar problems may reflect inconsistency. The presence of inconsistency is the chief concern for grading strength of evidence in this domain, and it would lead EPCs to reduce the overall strength-of-evidence grade.

If meta-analysis is appropriate, EPCs can evaluate consistency using statistical tests and measures of heterogeneity (such as Cochran's Q test or I^2 statistics, as discussed in the Quantitative Synthesis chapter of the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60).

Some bodies of evidence may show statistical heterogeneity in effect sizes but consistency in the direction of effect. Even if EPCs cannot explain the heterogeneity satisfactorily, they can still judge the evidence base to be consistent with respect to the direction of effect. With substantial unexplained heterogeneity, however, EPCs should be appropriately cautious about estimating treatment effects.

EPCs should designate an evidence base as inconsistent when different studies show statistically significant effect sizes in opposite directions. In the absence of statistical testing or measurement of heterogeneity, EPCs can assess consistency on the basis of similarity of populations, interventions, and outcome measures.

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,[10] and the methodology of mega-trials has been further questioned.[11]

Thus, EPCs cannot 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, with respect to consistency, we recommend that EPCs judge single-study evidence bases "consistency unknown (single study)," which would generally decrease the strength-of-evidence grade.

Directness

Directness concerns whether the evidence being assessed reflects a single, direct link between the interventions of interest and the ultimate health outcome under consideration (whether a benefit or harm). If direct evidence linking an intervention to the most ultimate outcomes is lacking, then two or more bodies of evidence are needed to link the intervention to health outcomes. When several bodies of evidence are involved, the ultimate decision about using an intervention 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. Thus, the final assessment of directness requires EPCs to consider the strength of evidence for each link as well as the importance of each link in the chain. Of particular salience is the extent to which evidence pertains to intermediate or surrogate outcomes rather than to ultimate patient-centered outcomes such as mortality, morbidity, and quality of life. More direct links enhance strength-of-evidence assessments (and vice versa).

In an example involving enteral feeding[12] used in this *Methods Guide* (see Principles for Developing Guidance for Comparing Medical Interventions),[3] 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. However, evidence of an association between a richer nutritional milieu and the ultimate outcome of complete healing may be weak. If this is a critical link in the causal chain, then the EPC can decide to grade the overall body of evidence as indirect, which would weaken the strength of evidence. As illustrated in the chapter on Principles for Developing Guidance for Comparing Medical Interventions of this *Methods Guide*,[3] use of an analytic framework is an important heuristic for determining how to evaluate evidence in a causal chain (e.g., in an overarching link or only in subsidiary linkages).

For CERs in particular, directness also applies to comparing interventions. For example, if there are three alternative interventions—A, B, and C—having evidence that compares them directly—A vs. B, A vs. C, and B vs. C—is desirable. In many circumstances, such head-to-head evidence is not available. Under these circumstances, reviewers must look to indirect evidence, such as evidence for A vs. C and B vs. C but not A vs. B. Grades for such indirect evidence will not be as strong as those obtained from truly direct evidence.

A single body of evidence is preferable to two bodies of evidence, particularly if the strengths 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 health outcome does or does not meet the ideal set of studies addressing the overarching question.

Precision

Precision is the degree of certainty surrounding an estimate of effect with respect to a specific outcome. EPCs should assess the boundaries of the pooled confidence interval for that effect estimate in relation to a threshold that would allow CER users to make judgments about the treatments being compared. Relevant thresholds for precision include the boundary of statistical significance—that is, whether the estimate of an effect reaches accepted levels for statistical significance. A precise estimate should enable decisionmakers to draw conclusions about whether one treatment is, clinically speaking, inferior, equivalent (neither inferior nor superior), or superior to another.[13,14]

Judgments about precision may depend on the importance of the outcome being measured, other clinically important outcomes, and the context of decisionmaking. They may also be contingent on whether the central issue is harms or benefits and the relative impact or size of those harms or benefits. This domain should be rated as precise or imprecise separately for each important outcome.

Substantial variability does not necessarily render an estimate imprecise. A truly imprecise estimate is one with a confidence interval so wide that it does not rule out the superiority or inferiority of either treatment being compared—that is, an estimate whose confidence interval includes two incompatible possibilities: one treatment is clinically significantly better than the other, and the difference is in the opposite direction. In this case, no conclusion can be reached about the relative effectiveness of the two treatments.

Additional Domains

The second set of domains, which supplement the four required domains, consists of secondary constructs that EPCs should use and report if they are relevant to a particular CER. These domains are dose-response association, existence of confounders that would diminish an observed effect, strength of association (i.e., magnitude of effect), and publication bias. These domains also derive from our review of other rating systems, including GRADE. Table 2 provides their definitions and ways to rate and apply them. Generally, we expect three of these domains—dose-response association, existence of confounding factors that would diminish an observed effect, and strength of association—to be applied more often to evidence from observational studies (of all types) than to evidence from RCTs.

The EPCs will invoke publication bias concerns when they have reason to believe that relevant empirical findings have not been published or are not otherwise available. Three situations are particularly relevant: (1) when negative, no-difference, or other studies with results that are substantially different from published studies are unavailable; (2) when the results of completed studies (e.g., those noted in ClinicalTrials.gov as having been ended 3 or more years in the past) have clearly not been published (save, perhaps, in abstract form); and (3) when trial protocols specify certain secondary end points for which results have not been reported (even if other results have been published). EPCs should consider and report on publication bias insofar as it appears to influence scores for either required or other domains (e.g., consistency or precision).

Applicability

A wide array of groups use EPC reports; not surprisingly, the context and populations these users consider relevant may differ. Thus, evidence that one group may consider applicable to the population of interest may not be applicable to the population of interest of another group.

For this reason, we have chosen to make our judgments about applicability explicit and separate from assessments of other domains of strength of evidence. In doing so, we aim to make it clear when our statements about the evidence are based on applicability rather than on other aspects of the evidence. Our goal in assessing applicability separately is to enable decisionmakers to take into account how well the evidence maps to the patient populations, settings, diseases or conditions, interventions, comparators, and outcomes that are most relevant to their decisions. Decisionmakers may determine that evidence is not readily applicable to their population of interest, and they should make recommendations accordingly.

Thus, we recommend that EPCs summarize characteristics that decisionmakers may need to consider in assessing the applicability of the evidence. In particular, EPCs should record information about applicability for the outcomes and comparisons for which they specify an overall strength-of-evidence rating. Summarizing such information in a separate table, which decisionmakers can review along with the strength-of-evidence table, may be helpful. Guidance for this process will be available in the Assessing Applicability paper of the *Methods Guide*, which was under review at the time of publication of this paper.

Procedures for Assessing Domains

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

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 directly into account and, as needed, incorporates judgments about the additional domains as well. For each comparison of interest, EPCs should rate strength of evidence for each major benefit (e.g., positive 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 both benefits and harms, EPCs should focus on the outcomes most relevant to patients, clinicians, and policymakers.

Systematic reviews and 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, reviewers should set clear priorities, assigning grades to those combinations (patients-interventions-outcomes) that are likely to be of greatest interest to users of the report. EPCs should also state clearly which interventions, outcomes, and comparators they included for each strength-of-evidence grade. For example, an evidence grade might apply to a link in an analytic framework, or it might apply to a specific intervention for a specific set of outcomes in a particular population. EPCs should also make clear which of the comparators or interventions is favored for each strength-of-evidence grade.

Table 3 summarizes the four levels of grades that EPCs should use. Each level has two components. The first, principal definition concerns the level of confidence the authors place in

the estimate of effect for the benefit or harm (i.e., their judgment that the evidence reflects the true effect). The second, 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.

Table 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 a conclusion.

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

High, moderate, or low strength of evidence. Assigning a grade of high, moderate, or low implies that an evidence base is available from which to estimate an effect. EPCs 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. EPCs should apply discrete grades and avoid designations such as “low to moderate” strength of evidence.

Insufficient. In some cases, the reviewers cannot draw conclusions for a particular outcome, specific comparison, or other question of interest. In these situations, the EPC should assign a grade of insufficient. Such situations arise in two main ways.

First, evidence for an outcome receives a grade of insufficient when no evidence is available from the included studies. This case includes the absence of any relevant studies whatsoever. In CERs, for example, certain drug comparisons may never have been studied (or published) in head-to-head trials and placebo-controlled trials of the multiple drugs of interest may not provide adequate indirect evidence for any comparisons.

Second, a grade of insufficient is also appropriate when evidence on the outcome is too weak, sparse, or inconsistent to permit any conclusion to be drawn. This situation can reflect several complicated conditions, 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 it is worse. 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 or if heterogeneity cannot be explained. This same evidence base 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 that overall assessment. In some systems, such as that of the GRADE working group,[4-6] the overall grade for strength of evidence (which it calls quality of evidence) is calculated from the ratings for each domain using a method that provides guidance on how to upgrade or downgrade the rating of the evidence. Such a system has the advantage of transparency because it clearly delineates a direct path from the evidence to its grade.

Although a system that uses such a method may offer advantages in terms of transparency, as yet there is not empirical evidence to support the superiority of a particular point system compared with a more qualitative approach. Furthermore, some evidence suggests no difference in accuracy between quantitative and qualitative systems.[7] Research is needed to compare the performance of a point system approach with other grading systems before we can recommend that EPCs use any specific system. Thus, EPCs may use different approaches to incorporate multiple domains into an overall strength-of-evidence grade.

The EPCs should explain the rationale for their approach to rating of strength of evidence and note which domains were important in upgrading or downgrading the strength of evidence. GRADE uses an algorithm to help reviewers to be clear about how they consider domains to produce the grade. EPCs may use the GRADE system or their own weighting system, or they may elect to use a qualitative approach, so long as the rationale for ratings of strength of evidence is clear. Several general principles that all should follow are important.

First, the risk of bias based on the design and conduct of the available studies is an essential component to rating the overall body of evidence. In considering the risk-of-bias domain, EPCs should consider which study design is most appropriate to reduce bias for each question. For many of the traditional therapeutic interventions, evidence that is based on well-conducted randomized trials will have less risk of bias than does evidence based on observational studies. For these outcomes, if randomized trial data are available, EPCs may choose to start with a rating of low for the risk-of-bias domain and change the assessment of this domain if the RCTs have important flaws. For these traditional therapeutic intervention questions, observational data would generally start with a high risk of bias but may be altered depending on the conduct of the study. As with all questions, the overall strength of evidence must incorporate assessments of other domains in addition to risk of bias.

Second, EPCs should assess each of the major domains for rating the overall strength of evidence. Assessment of consistency, directness, and precision may reveal strengths or weaknesses with the entire body of evidence and lead to a strength of evidence that is either higher or lower than would be obtained by considering only risk of bias. EPCs should also consider the additional domains when appropriate; they need not report on those domains when they regard them as irrelevant to the review in question. The strength of the evidence would be weakened by concerns about publication bias. In contrast, several factors may increase strength of evidence and are especially relevant for observational studies, where one may typically begin with a lower overall strength of evidence based on the risk of bias. Presence of a clear dose-response association or a very strong association would justify increasing strength of evidence. If the confounding that may exist in a study would decrease the observed effect, but an effect is observed despite this possible confounding, the EPC may wish to upgrade the strength of evidence.

Third, EPCs should decide a priori how to incorporate each domain into an overall strength of evidence and what measures they will use to ensure accuracy and consistency of

evidence ratings. The degree to which the overall strength of evidence is altered by additional domains that are used is a judgment that EPCs should explain in the report.

Key Procedures

EPCs should also take specific steps to ensure reliability and transparency within their own work (both in individual reviews and across them) when incorporating domains into an overall grade. As a first step, they should be explicit about whether the evidence grade will be determined by a point system for combining ratings of the domains or by a qualitative consideration of the domains. They should carefully document procedures used to grade strength of evidence and provide enough detail within the report to assure that the users can grasp the methods that were employed. EPCs should, furthermore, keep records of their procedures and results for each review so that they may contribute to the overall EPC expertise and science of grading evidence.

Second, EPCs should identify the domains that are most important for the targeted body of evidence and decide how to weight the domains when assigning the evidence grade. For the sake of consistency across reviews, the domains should be defined using the terminology presented in this chapter. In the absence of evidence to support specific systems for weighting of the domains, both qualitative and quantitative approaches are acceptable. EPCs may also choose to follow GRADE guidance for downgrading and upgrading evidence based on assessments of each domain. 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 and conduct have 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 with appropriate clinical and methodological expertise will rate each domain. 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 their CERs, they should keep records of this information. By documenting this information, 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 the first through third steps. If this action involves a qualitative approach with subjective weighting of the domains, EPCs should consider using at least two reviewers and assessing 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 readers to replicate it if desired.

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 decisionmakers. Thus, strength of evidence should relate to those important outcomes. 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 these reasons, EPCs should also present information that will help decisionmakers 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 4 illustrates one approach to providing actionable information to decisionmakers that reflects strength of evidence. 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).

Table 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:	Consistency	Directness	Precision	Absolute risk difference per 100 patients
Mortality					**Insufficient SOE**
1;80	RCT/Medium	Unknown	Direct	Imprecise	-1 (95% CI -4 to +3)
14;384	Retrospective cohort/ Medium	Inconsistent	Direct	Imprecise	-7 to +5 (range)
Myocardial infarction					**Low SOE**
7: 625	Retrospective cohort/ High	Consistent	Direct	Precise	-3 (95% CI -5 to -1)
Severe diarrhea					**Moderate SOE**
4; 256	RCTs/ Medium	Consistent	Direct	Imprecise	-4 (95% CI -8 to +1)
14; 28,400	Cohort/ Medium	Consistent	Direct	Precise	-5 (95% CI -8 to -2)
Improved quality of life					**High SOE**
6; 265	RCTs/ Low	Consistent	Direct	Precise	-5 (95% CI -1 to -7)
Ulcer healing					**High SOE**
6; 265	RCTs/ Low	Consistent	Direct	Precise	+12 (95% CI +4 to +27)
5; 684	Retrospective cohort/ Low	Consistent	Direct	Precise	+17 (95% CI +12 to +22)

Abbreviations: CI=confidence interval; RCT=randomized controlled trial; SOE=strength of evidence.

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 percent 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 and consistent (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. However, the evidence is insufficient to allow a conclusion for mortality.

Although Table 4 illustrates how EPCs might organize information about the strength of evidence and magnitude of effect in ways useful to decisionmakers, it is incomplete. First, the table does not convey any information about the applicability of the evidence, which would be presented through other means (text or table). Second, a narrative summary of the results is also essential for interpreting the results of a literature synthesis.

Discussion

The EPC approach to rating the strength of evidence draws heavily on the international GRADE system; both conceptually and substantively, it is similar to GRADE. Our recommendations address specific circumstances of the EPC program, which differ from those of some groups that use GRADE. The EPC program produces systematic reviews, but it is not involved directly in development of recommendations or guidelines. Rather, EPC reports are used by a spectrum of government agencies, professional societies, and other stakeholders. Our approach for grading strength of evidence and discussing applicability of the evidence is meant to facilitate use of the EPC reports by this broad group of users.

We recommend that EPCs rate strength of evidence based on a core group of domains that include risk of bias, consistency, directness, and precision. Randomized trials will generally be assessed to have a low risk of bias, which correlates with a high strength of evidence, but may be changed after evaluation of other domains. Evidence based on observational studies will generally have a high risk of bias, which correlates with a low strength of evidence, but may be rated higher after evaluating other domains. When appropriate, the EPCs can also use additional domains of dose-response association, the impact of plausible confounding, strength of association, and publication bias to upgrade or downgrade the strength of evidence.

This overall approach is similar to the methods used in the GRADE system. In GRADE, evidence based on observational studies starts with a strength of low and can be upgraded based on several factors. In the approach we describe here, the EPC may believe that, for certain outcomes, such as harms, observational studies have less risk of bias than do randomized trials or that the available randomized trials have a substantial risk of bias. In such instances, the EPC may either 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 or low.

We recognize that some types of evidence, such as evidence about public health interventions, quality improvement studies, and studies of diagnostic tests, may be challenging to rate. With these nontherapeutic intervention questions, the challenge to the EPCs is to determine the study design that is most appropriate to minimize the risk of bias. For example, the EPCs may find that particular types of studies, such as interrupted time series, reduce the risk of bias more than do other types of observational studies. Although the EPCs can take into account criteria other than those specified expressly by GRADE in assessing the risk of bias of observational (nonrandomized) studies as moderate, we caution that changing the assessment of observational studies for risk of bias should be done judiciously.

AHRQ CERs have often focused on pharmaceutical therapies, for which both efficacy and effectiveness trials[15] are a major source of information. The domains discussed above are directly relevant to studies of most drugs. In the future, CERs may increasingly assess diagnostic tests or strategies. For these technologies, RCTs may not be the origin of much relevant information, and the studies that are available may have special methodologic features. Further conceptual or empirical work may be warranted to explore whether the EPC approach to grading strength of evidence described here remains appropriate for such interventions. EPCs are

encouraged to keep careful records of the application of these methods to nonpharmacologic interventions.

In arriving at an overall strength-of-evidence grade, the crucial requirement is transparency. The EPC method implies that EPCs can, if they choose, make a global assessment of the overall quality of evidence rather than explicitly use scores for each domain and then combine them. Nevertheless, EPCs are encouraged to make judgments for individual domains as a first step and to be especially sensitive to the effects of any "borderline" scores for those domains and their impact on the overall score. Being explicit and transparent about what criteria are used to raise or lower grades is the essential element in this step.

As noted earlier, the EPC approach emphasizes assessment of applicability separately from strength of evidence. GRADE also addresses applicability, which is incorporated within the general concept of directness. The rationale for the EPC approach is that many stakeholders use EPC reviews for developing guidelines or making clinical or health policy decisions, and they may have quite different views on how much, or little, the evidence applies to populations of interest to them. Future EPC reports will have a discussion and information about applicability, and the intention is for the various users and audiences to read this section of the report and make their own judgments.

A consistent approach for grading the strength of evidence—one that decisionmakers can readily recognize and interpret—is highly desirable. To that end, the EPCs and the GRADE working group will continue to collaborate to facilitate consistency across grading systems. Refinements and modifications of the approach outlined here can be found at www.effectivehealthcare.ahrq.gov as they become available. Meanwhile, this paper codifies the interim guidance that EPCs can follow to strengthen the consistency within the AHRQ program's current and coming reports and products.

Acknowledgments

The authors thank Valerie King, M.D., M.P.H., of the John M. Eisenberg Center at Oregon Health & Science University, for her insightful comments on an earlier draft and Loraine Monroe, of RTI International, for superior assistance with manuscript preparation. We thank Gordon Guyatt, Holger Schünemann, and the members of the GRADE working group for their work on rating quality of evidence, for very helpful discussions about our approach, and for comments on the manuscript.

Author Affiliations

VA Palo Alto Healthcare System; Stanford-University of California San Francisco Evidence-based Practice Center; Center for Primary Care and Outcomes Research, Stanford University, Palo Alto, CA, (DKO). RTI International, Research Triangle Park, NC, (KNL). Health Services Research & Development Service, Department of Veterans Affairs, Washington, DC, (DA). ECRI Institute Evidence-based Practice Center, Plymouth Meeting, PA, (JRT). ECRI Institute, Plymouth Meeting, PA, (JTR). Johns Hopkins University Evidence-based Practice Center, Baltimore, MD, (EBB). Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD, (SC). Oregon Health & Science University Evidence-based Practice Center, Portland VA Medical Center, Portland, OR, (MH).

This report has also been published in edited form: Owens D, Lohr K, Atkins D, et al. AHRQ Series Paper 5: Grading the strength of a body of evidence when comparing medical

interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol* 2010; 63, 513–523.

References

1. Helfand M. Using evidence reports: progress and challenges in evidence-based decision making. Health Aff (Millwood) 2005;24(1):123–7.
2. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005 Jun 21;142(12 Pt 2):1035–41.
3. Helfand M, Balshem H. Principles for developing guidance: AHRQ and the Effective Health-Care Program. 2010 May;63(5):484–90. Epub 2009 Aug 27.
4. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches, The GRADE Working Group. BMC Health Serv Res 2004 Dec 22;4(1):38.
5. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924–26.
6. Guyatt GH, Oxman AD, Kunz R, et al.. What is “quality of evidence” and why is it important to clinicians? BMJ 2008 May 3;336(7651):995–8.
7. West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). Rockville, MD: Agency for Healthcare Research and Quality, 2002. AHRQ Publication No. 02-E016.
8. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001 Apr;20(3 Suppl):21–35.

9 Treadwell JR, Tregear SJ, Reston JT, et al. A system for rating the stability and strength of medical evidence. BMC Med Res Methodol 2006;6:52.

10. Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. J Clin Epidemiol 2000 Dec;53(12):1193–9.
11. Charlton BG. Fundamental deficiencies in the megatrial methodology. Curr Control Trials Cardiovasc Med 2001;2(1):2–7.
12. Langer G, Schloemer G, Knerr A, et al. Nutritional interventions for preventing and treating pressure ulcers. Cochrane Database Syst Rev 2003(4):CD003216.
13. Sackett DL. Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis. ACP J Club 2004 Mar-Apr;140(2):A11.
14. Sackett D. The principles behind the tactics of performing therapeutic trials. In: Haynes RBS, Guyatt DL, Gordon H, Tugwell P, eds. Clinical epidemiology: how to do clinical practice research. New York: Lippincott Williams & Wilkins; 2005.
15. Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006 Oct;59(10):1040–8.

Chapter 11. Using Existing Systematic Reviews To Replace De Novo Processes in Conducting Comparative Effectiveness Reviews

C. Michael White, Stanley Ip, Melissa McPheeters, Tim S. Carey, Roger Chou, Kathleen N. Lohr, Karen Robinson, Kathryn McDonald, Evelyn Whitlock

Key Points

- Using existing systematic reviews (SRs) has potential benefits and risks. Evidence-based Practice Centers (EPCs) and the relevant Task Order Officer should discuss these points.
- This chapter does not focus on the use of existing systematic reviews for obtaining background information, providing background or discussion context, or cross-checking references. Rather, it concerns the use of existing systematic reviews to replace a de novo process. It also does not consider the processes used to create separate products, called "umbrella" reviews, meta-reviews, or reviews of reviews.
- We propose a five-step process to standardize the approach that EPCs can use to decide whether existing systematic reviews might provide value (Figure 1).
- Transparency is a priority; users of a Comparative Effectiveness Review (CER) should be able to determine what was done (Figure 2).
- Two independent reviewers using a modified AMSTAR (Assessment of Multiple Systematic Reviews) instrument should assess the quality of relevant reviews (Table 1).
- EPCs should incorporate existing systematic reviews (i.e., use them to replace all or part of a de novo process) only if they are fully relevant and of high quality. Partly relevant or suboptimal quality reviews should not be incorporated, although they may be useful for cross-checking references and for providing background. It is important to discuss how the findings of the CER agree or disagree with particularly well known SRs (highly cited or published in a high-impact journal) not included in the CER's discussion section.
- Once EPCs identify relevant, high-quality systematic reviews, they may opt to use them in the following ways: adapting or adopting the search strategy, using the summarized evidence, or a combination of these.
- EPCs can choose to replace a de novo process to answer a key question by selecting the best review or may choose to summarize all of the relevant and high-quality reviews.
- EPCs should routinely review reference lists of such systematic reviews to identify relevant studies
- If EPCs do a de novo synthesis, they should routinely compare results with those of relevant, high-quality systematic reviews and formally address consistency or potential reasons for discrepancies in the discussion of the report.

Introduction and Rationale

Over a 4-year period (2005 to mid-September 2009), 11,390 citations for systematic reviews and 11,281 citations for meta-analyses were retrieved in an OvidSP search. In contrast,

over the previous 9 years (1996 to 2005) only 7,390 citations for systematic reviews and 9,251 citations for meta-analyses were retrieved. Approximately 2,500 new systematic reviews (SRs) and meta-analyses were published in 2006 alone.[1] A systematic review uses an explicit methodology for systematically searching and synthesizing the literature and for grading evidence. Given the extensive body of existing SR and meta-analysis literature, questions have been raised about whether Evidence-based Practice Centers (EPCs) should use existing SRs in a Comparative Effectiveness Review (CER) commissioned by the Agency for Healthcare Research and Quality (AHRQ) and, if so, in what capacity they should be used. Of course, examining existing SRs to provide background information or other useful references for a CER is a common practice in EPC work, and we do not discuss this procedure further in this chapter.

An informal survey of eight non-EPC centers that conduct systematic reviews in the United Kingdom, Australia, and New Zealand confirmed that they are facing these same questions about the use of existing SRs without any commonly accepted approach.[2] In summer 2008, the Existing SR Working Group queried EPC directors about their experiences (including experience with both EPC and non-EPC projects) in this area. Overall, EPCs considered the use of an existing SR 50 percent of the time and used existing SRs slightly more than 30 percent of the time. The most commonly stated reason for using an existing SR was for completeness, but existing SRs were also often used when EPCs faced a topic of extensive breadth, because of the sizable body of literature, or limitations in timeframe or budget. Some EPCs used the existing SR while updating the SR.

When queried about how they were using existing SRs, EPCs indicated that they used existing SRs predominantly (74 percent of the time) for background information or to ensure completeness of the literature search. EPCs sometimes used results of existing SRs to answer key questions in the new SR, but in more than two-thirds of these cases, at least a sample of the original trials or studies included in the existing SR were verified to ensure the quality of original data extraction.

When EPCs considered using existing SRs in a new SR, the most common reason given *not* to use one was that the identified reviews were not relevant to the specific questions being asked in the new SR. Other frequent reasons not to use existing SRs included: no time savings associated with using the existing SR vs. using de novo methods to answer the key question, poor quality of existing SRs after detailed assessment, outdated existing SRs, and uncertainty about how to include them in a new SR.

As a result of our queries and subsequent discussion within the Working Group, we identified six possible benefits associated with using existing SRs in CERs:

- Allows a cross-check to assure that relevant trials and studies are captured in a new CER.
- Allows EPCs to directly compare and contrast the present CER and previous SRs in terms of findings that may be relevant to health care decisionmakers.
- May save EPCs time, effort, and resources to answer key questions.
- May allow EPCs to anticipate and plan for context-specific methodological issues.
- May help avoid unnecessary redundancy among SRs.
- May provide analyses that are not readily available from other sources (e.g., subgroup analyses from a meta-analysis of individual patient data not available in constituent studies or published reports).

In addition, some existing SRs may contain additional information from primary studies not reported in the manuscripts resulting from author queries or by having a primary study author as an author on the SR.

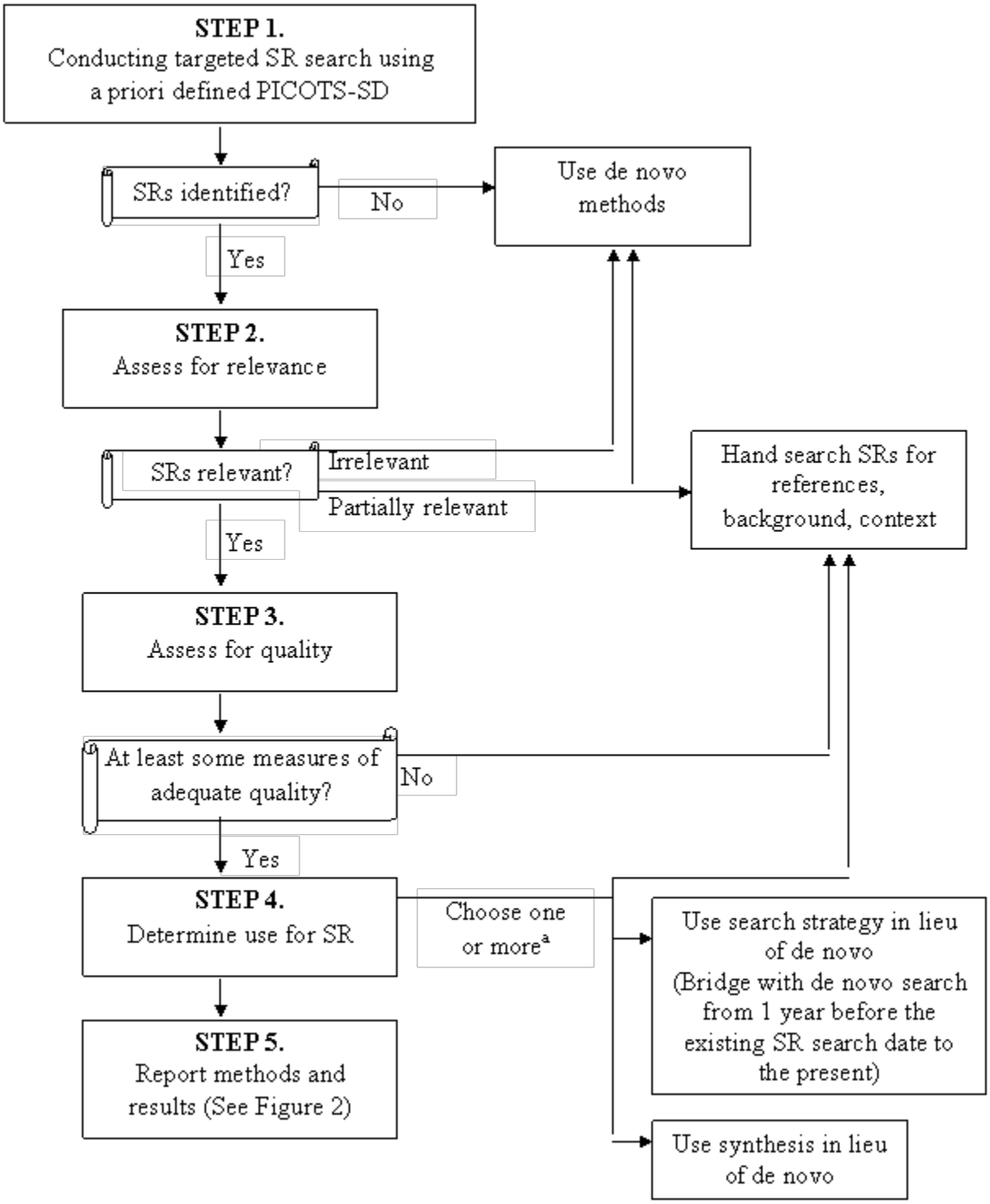
Conversely, five main risks are associated with using existing SRs in CERs that do not arise in a purely de novo process:

- If EPCs find numerous existing SRs, the time and resources required to evaluate them may be wasted because earlier reports may not be recent enough, not relevant enough to answer the key questions posed, or not of acceptable quality.
- Incorporating the results of existing SRs into a CER could propagate errors arising from errors in data abstraction, selection of studies, and qualitative or quantitative synthesis. Propagating errors can reduce credibility for the CER and the EPC program among stakeholders and users.
- Using an existing SR to answer key questions might create a perception that EPCs are not performing due diligence in conducting a CER. This perception might reduce credibility for the CER and the EPC program among stakeholders and users.
- If the existing SR does not provide evidence from primary studies and analyses in sufficient detail, the methodological process of the CER may be perceived to lack transparency.
- Ambiguity about how to compare multiple existing SRs on the same subject remains an important challenge. Lack of clear methodological guidance on selecting the most appropriate SRs could introduce reviewer bias, which is especially true if existing SRs have discordant results.

The use of existing SRs to substitute for purely de novo CER methods may provide benefits and risks. Ultimately, EPCs need to work with those who commission the work (i.e., their Task Order Officers at AHRQ and decisionmakers who nominated the topic) to determine whether the potential benefits associated with the incorporation of existing SRs are worth the risks to a CER's comprehensiveness and transparency or the risk of introducing bias. If a decision has been made to incorporate the use of existing SRs in answering one or more key questions in lieu of using a purely de novo process, we recommend that EPCs apply the following approaches.

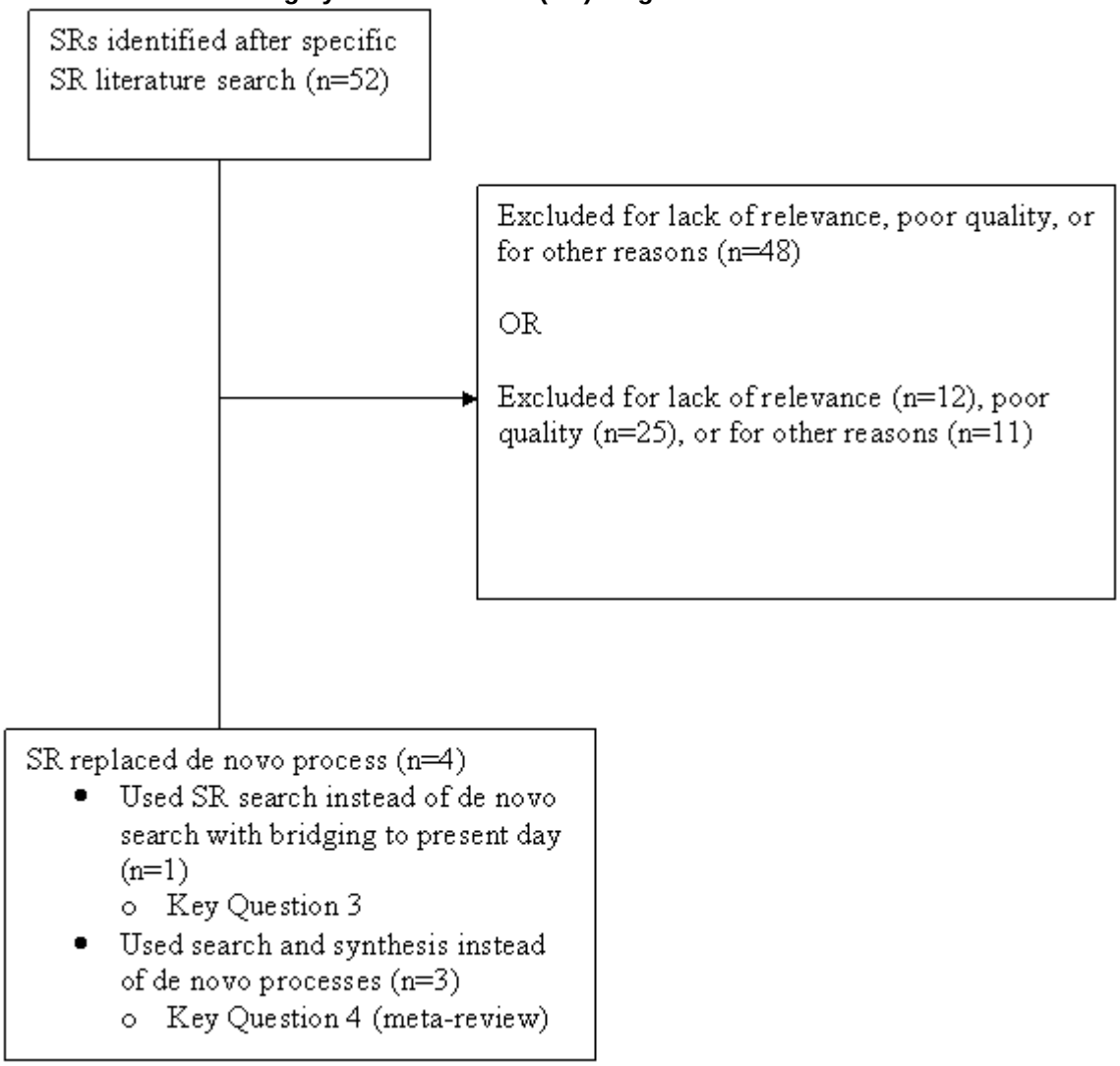
Figure 1 is a flow diagram adapted from a methods article by Whitlock and colleagues.[2] It will help guide EPCs as they move through the process of identification, assessment, and use of existing SRs. To ensure transparency, EPCs can include a graphic similar to the example shown in Figure 2 in a CER report so users can identify the number of original citations identified in an SR search, the number of articles that are excluded, and how the existing SRs are being used.

Figure 1. Systematic process for identifying, assessing, and using existing systematic reviews



Adapted from Whitlock EP, Lin JS, Shekelle P, et al. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008;148:776-782.

Figure 2. Illustrative existing systematic review (SR) diagram



Locating Existing Systematic Reviews

Using search terms that reflect a priori PICOTS-SD (population, intervention, comparator, outcome, setting, and study design) refines the search and decreases noise. Although EPCs can apply many possible approaches to identify existing SRs for a CER, we recommend two procedures. One strategy is to use a targeted search of higher yield databases.[2] Because SRs are a secondary literature source, identifying relevant, high-quality SRs is probably more important than identifying all SRs because redundancy of primary studies across SRs is likely. Higher yield databases include the output of the Evidence-based Practice Center program, MEDLINE's Top 120 Index Medicus Journals, Health Technology Assessments, Cochrane Database of Systematic Reviews, and Database of Abstracts of Reviews of Effects. EPCs can add other databases depending on the topic. Alternatively, EPCs can identify SRs during their title and abstract searches while conducting a broad de novo literature search for trials and studies, as long as the searches are structured not to exclude reviews. The EPC medical librarian is a valuable resource when making these decisions and developing the search strategy.

Assessing the Relevance of Existing Systematic Reviews

EPCs considering the inclusion of prior SRs in a CER should begin with a fundamental presumption—that the intent is to answer one or more key questions or a specific portion of a key question with an existing SR in lieu of a completely de novo process. Relevance requires consideration of the PICOTS-SD. Those SRs not completely relevant to the current review (partially relevant) may still be useful for background material or for cross-checking references. Some existing SRs will not be relevant at all and should be eliminated from any further consideration at this stage.

Initial Screening for Relevance

As depicted in Figure 1, after EPCs conduct a literature search for existing SRs (Step 1), they need to screen identified citations for relevance (Step 2). Citations that are not SRs (primary research, narrative reviews) or duplicate citations can be readily excluded.

Many factors that determine whether an existing SR is relevant or not are addressed in the SR's methods section. Timeliness of the existing SR is critical. Timeliness refers not to the publication date of the review, but to how recently the literature search was conducted. When considering issues of timeliness, reviewers should be aware that SRs can become outdated quickly.[3] Whether an SR is outdated depends primarily on the topic because some areas may not be as intensely researched and newer studies added only rarely. We generally recommend bridging any search date for an SR that ended a year or earlier than the present date. Given their clinical expertise, expert team members may be helpful in deciding acceptable date parameters; ideally they should make this decision a priori.

If EPCs regard an earlier SR to be outdated, they can still consider using the search results (obtaining data from the evidence tables) and then updating from 1 year before the date of the original literature search to the present time with a de novo process. By going back 1 year before the existing SR's search date, the lagtime between the publication of an article and its inclusion into standardized literature retrieval databases ought not to be a major factor. Using the search results from these existing SRs would require only that the earliest date for which studies could be included (e.g., 1960) is in line with the date the EPCs have set for their CER.

Focusing on Population, Intervention, Comparator, Outcomes, and the Timing of Their Measurement, Setting, and Study Design To Assess Relevance

For existing SRs that make it to this stage, EPCs should compare the PICOTS-SD in the earlier SRs with these elements in the new CER protocol.[4] Determining similarity will depend on how well the existing SR describes these elements. Poor reporting will make it impossible for an EPC to consider inclusion of an existing SR. Poor reporting, however, is an element of quality appraisal as well, so a poorly reported SR would not be eligible for incorporation for both relevance and quality reasons. Appreciating the subtle differences that may exist between an existing SR and the current CER is vital; this generally requires EPCs to give careful consideration of these elements.

Population. The need for the population in an existing SR to "match" completely the intended population in a new CER will depend to some degree on the clinical condition of interest and the questions being addressed. On the one hand, for example, a CER that is attempting to review

interventions for hemorrhagic stroke may not be well served by including an existing SR with studies of patients with any kind of stroke unless results clearly separate the subgroup of studies relevant to hemorrhagic stroke patients. On the other hand, a CER that is examining any kind of stroke might be able to incorporate a relevant, high-quality prior SR addressing hemorrhagic stroke only. Similarly, an existing SR restricted to adults will be of limited utility if the new key questions include young children. Other CERs, however, may require less rigidity, and modest differences in age range or geographic range (e.g., United States vs. North America) may be less important.

Intervention. To ensure that existing SRs evaluated the same intervention as intended for the new CER, the team should look carefully at criteria for inclusion used in the older review. It is particularly important to make sure that issues such as dosing and mode of delivery match as closely as possible. When the existing SR was either more or less inclusive than the CER is intended to be, the experts on the team need to determine that this factor will not fundamentally change the conclusions. This may become an issue when dosing regimens change over time, as has been the case with use of higher dose statins in recent years, or for example, in the evolution of cardiac devices such as pacemakers to newer, dual-chamber versions.

Comparator. EPCs should consider whether they are interested in the effect of the intervention of interest as it compares with usual practice or another intervention and ensure that the existing SR matches this criterion. EPCs should note, when comparing treatments with usual care, whether usual practice has changed significantly since the timeframe of the earlier SR; this would make older studies—and perhaps a review of those studies—not applicable to the current concern. Such evolution of usual practice has been a significant issue, for instance, in "medical treatment" after acute coronary syndrome; older versions of medical treatment are no longer comparable with current practice. In surgical reviews, it may be important to know what supportive treatments were used in the past compared to those associated with interventions being reviewed. For example, if patients previously spent longer in postoperative care in bed rather than in active rehabilitation, those older studies may not reflect current practice. For issues of this type, the input of clinical experts can be particularly useful to determine changes in usual care over time.

Outcomes. The outcomes assessed in existing SRs should be the same as or similar to the outcomes envisioned for the CER. The usual caveats regarding use of intermediate or nonpatient-oriented outcomes apply for existing SRs just as they apply to inclusion criteria for constituent studies.

Timing of outcome measurement. Some SRs are restricted to studies with relatively short periods of followup. The period of appropriate followup, of course, depends on the condition, intervention under consideration, and outcome being assessed. The rationale for such restriction may be the lack of availability of longer term followup; when such studies become available, the relevance of the older SR is reduced. Often, short periods of followup involve surrogate outcome measures; both factors (length of followup, surrogate or proxy outcomes) decrease an SR's relevance. Timing of outcome measurement is not the same as timeliness (how recent the existing SR is), which EPCs should examine early in the relevancy assessment.

Setting. Older SRs can address interventions in a broad or narrow range of settings, such as interventions to reduce falls in inpatient settings, in nursing homes, and in the home and other community settings. Although some of these distinctions will be clear by examining the populations addressed, a previous SR that covers a wider range of settings may not be relevant to a more narrowly scoped CER unless results of the former are stratified by setting.

Study design. SRs can differ appreciably in the types of study designs that they consider acceptable. EPCs may find that surveying inclusion criteria related to study design is a useful early step in an evaluation of relevance. If EPCs plan to include randomized and controlled clinical trials and high-quality comparative cohort studies as evidence in their CERs, but an existing SR covers only randomized controlled trials, then the latter is only partially relevant to the current effort.

The original author of the existing SR could be contacted for additional information if it is not clear whether or not sufficient relevance is present. Once EPCs have established relevance for an existing SR, they should assess and rate quality using the approach described below. Quality assessments (Figure 1, Step 3) are time intensive and should be conducted only on existing SRs found to be relevant.

Assessing the Quality of Relevant Systematic Reviews

Whatever aspect of an existing SR an EPC includes in the CER should adhere to a high methodological standard. EPCs should avoid routinely including all existing SRs in an attempt to be comprehensive. Note that this admonition is in contrast to another effort, a review of reviews, in which reviewers are asked to summarize the available evidence at the level of the systematic review.

Several instruments designed to rate quality of SRs are available.[5] Regardless of the specific instrument that is chosen for this purpose, the instrument should address all aspects of the review that the EPC plans to incorporate into the CER, including methods used to identify, select, appraise, and synthesize studies; the possibility of publication bias; and potential conflicts of interest.[6]

Commonly Used SR Quality Instruments

In assessing the quality (i.e., assessing the risk of bias) of existing SRs, EPCs should address both the methods used by the earlier systematic reviewers to minimize bias and the transparency and completeness with which they reported their methods, individual study details, and results. Checklists for improving reporting of SRs (e.g., QUOROM [recently renamed PRISMA], MOOSE) have been used as surrogate tools for quality assessment, although they were designed to improve transparency and consistency of reporting SR methods, not directly to assess methodological quality.[7-9] For example, the QUOROM checklist requires detailed descriptions of the literature search strategy terms and sources searched, but it does not provide criteria for distinguishing adequate from inadequate searches.[7] In addition, inadequate reporting of SR methods does not necessarily mean that the SR was conducted poorly. Nonetheless, rating the quality of an SR without understanding how it was conducted is difficult. Several items related to quality of reporting have been incorporated into instruments such as the ones from Oxman and Guyatt and AMSTAR.[6,10]

The Oxman and Guyatt instrument was one of the early widely used standardized quality rating indexes for evaluating the scientific quality of a review article; unlike other quality rating

instruments specifically developed for SRs, some empiric evidence supports its use.[10] Reviews with lower quality ratings on the Oxman and Guyatt instrument are more likely to show treatment benefit.[11,12] However, methods for evaluating SRs have evolved since the Oxman and Guyatt instrument was developed, and it does not address several methodological domains now thought to be important.[13]

The newer Assessment of Multiple Systematic Reviews (AMSTAR) tool includes additional criteria, such as whether study selection and data extraction were conducted in duplicate, whether publication bias was assessed, and whether conflicts of interest were reported.[6] Although more data are needed to determine its reliability and validity, AMSTAR has been proposed as the preferred instrument for assessing the quality of SRs by the World Health Organization and by the Canadian Optimal Medication Prescribing and Utilization Service (COMPUS), among others.[14,15] One domain that is not included in AMSTAR pertains to nonbiased application of inclusion and exclusion criteria, although EPCs can adapt the AMSTAR instrument to include such an item. (See recommendation.)

Limitations in Quality Rating Scales

As much as possible, CER investigators should apply objective and reproducible criteria when using quality assessment instruments such as Oxman and Guyatt or AMSTAR.[6,10] For example, a "comprehensive" literature search could be defined as requiring searches on at least two electronic databases, reference list searching, and expert queries. Although EPCs could use this definition in most instances, they may need to tailor criteria for specific topics. For example, for assessing the quality of SRs that evaluate acupuncture, fully meeting the literature search criteria could require searching Asian-language databases.

For some criteria included in quality rating instruments, delineating objective definitions is difficult; EPCs then must apply subjective judgments. For example, AMSTAR includes the items "Was the scientific quality of the included studies used appropriately in formulating conclusions?" and "Were the methods used to combine the findings of studies appropriate?"[6] Assessing and rating quality using discrete categorical choices can make quality judgments appear more clear cut and objective than they really are. Operationalizing subjective qualifiers such as "appropriate" at the outset of each assessment, taking into consideration factors relevant to the specific topic at hand, could help. Having at least two independent reviewers from an EPC assess quality and reporting methods for resolving discrepancies is desirable.

Another limitation in applying quality rating instruments is that they are not designed to detect inconsistencies in application of inclusion criteria or errors in data abstraction. For example, an SR[16] of antidepressants for low back pain specified randomization as an inclusion criterion but included a nonrandomized clinical trial.[17] Among the included studies, this trial reported the highest estimate of benefit and may have affected the SR's conclusions.[16] Checking data from SRs against primary studies can reveal important discrepancies.[18,19]

Numerical summary scores (e.g., adding up the number of criteria that are adequately met) have been used to summarize the overall quality of SRs. Such scores can be misleading because reviews with different flaws may receive the same summary score. A summary score could not dissect the nature of the bias in the individual review. For example, an SR could meet nearly all methodological criteria and receive a near-perfect summary score, but one serious methodological shortcoming could invalidate its results; a summary score may well not reflect that important shortcoming.

We suggest that CER authors describe the implications of individual methodological flaws rather than rely on numerical summary scores. For example, exclusion of "grey literature" or non-English-language citations may or may not have important effects on estimates of benefits or harms.[20,21] If EPCs find no clear indication of publication bias in an SR and if stable and precise estimates are available for the outcome(s) of interest, excluding these types of literature is not likely to be a serious shortcoming. However, excluding "grey literature" or non-English language trials would be a serious shortcoming in an SR if large numbers of trials or important trials are known or suspected to exist in these literature types. As cases in point, medical device evaluations may rely on "grey literature,"[22] and alternative and complementary medicine evaluations may rely on foreign-language literature.[23]

Assigning categorical quality scores (such as "good," "fair," or "poor") may be appropriate after taking into account the number and seriousness of methodological shortcomings.[24] In general, good-quality SRs should be defined as those that have few or no methodological shortcomings and a low risk of bias. Fair-quality SRs have some methodological flaws but the EPC conducting the CER determined that the flaws will not seriously bias or invalidate the results. Poor-quality SRs contain a serious flaw or flaws that, in the judgment of the EPC conducting the CER, are highly likely to bias or invalidate the results.

CER Quality Assessment Recommendations

When EPCs assess the quality of an existing SR for a CER project, we recommend:

- At least two independent reviewers should assess SRs for quality.
- EPCs should report methods for resolving discrepancies between reviewers.
- EPCs should confirm the reproducibility of application for inclusion criteria and the accuracy of data abstraction in at least a sample of the studies. They should confirm that a nonbiased application of inclusion criteria was used.
- To have a common starting point, EPCs should use AMSTAR for quality evaluation for two reasons: (1) it was developed based on an SR of quality rating instruments and has undergone some construct and validity testing; and (2) it is becoming more widely used internationally.

AMSTAR assesses 11 criteria for quality and the choices are (Yes, No, Can't Answer, and Not Applicable).[6] We suggest supplementing the AMSTAR questions as deemed appropriate for the particular project or topic at hand. Table 1 summarizes the criteria with some additional considerations that EPCs may have for their CERs.

Table 1. AMSTAR quality criteria with considerations for Comparative Effectiveness Reviews

Number	Criterion	Considerations for Comparative Effectiveness Reviews
1	Was an a priori design provided?	—-
2	Was there duplicate study selection and data extraction?	Was there dual review for study selection and data extraction? After checking a sample of original studies: Was the application of inclusion/exclusion criteria unbiased? Were any discrepancies between data from primary papers and the published systematic review identified?
3	Was a comprehensive literature search performed?	Was the search strategy appropriate for the posed key questions? This should be consistent with the chapter on finding evidence in the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*.
4	Was the status of publication (e.g., grey literature) used as an inclusion criterion?	Some reviews do not restrict inclusion based on whether studies were peer reviewed or not. EPCs should state their criteria for inclusion/exclusion and justifications for the criteria (e.g., reasons for restriction to English language, excluding letters and abstracts, etc.)
5	Was a list of studies (included and excluded) provided?	—-
6	Were the characteristics of the included studies provided?	—-
7	Was the scientific quality of the included studies rated and documented?	Was individual study quality (such as sample size, study design, blinding, various biases and confounders, study subject attrition rate, etc.) assessed? This should be consistent with the chapter on assessing quality in the Methods Guide. Did the systematic review include high-quality primary studies? (No matter how well conducted a systematic review, its findings are limited by the quality of included primary studies.)
8	Was the scientific quality of the included studies used appropriately in formulating conclusions?	This item applies only if EPCs use the conclusions from the prior systematic review(s) in their CERs. Often EPCs will use only the results and formulate conclusions based on the data and analysis presented. This should be consistent with the chapter on grading the strength of a body of evidence in the Methods Guide.
9	Were the methods used to combine the findings of studies appropriate?	—-
10	Was the likelihood of publication bias assessed?	Publication bias can be assessed, in part, by assessing for editorials, letters to the editor, or comments elucidated in other peer-reviewed literature.
11	Was the conflict of interest stated?	Have the authors disclosed declared or known conflicts of interest? Examples include funding source for the project, consulting fees, and stock ownership.

Abbreviations: AMSTAR=Assessment of Multiple Systematic Reviews; EPC=Evidence-based Practice Center.

Checklists have been developed to improve the quality of reporting of meta-analyses evaluating therapeutic interventions (e.g., see previously mentioned PRISMA: www.prisma-statement.org/index.htm). These reporting checklists may not be directly applicable to individual patient data meta-analyses. Although these types of meta-analyses may not be comprehensive or systematic in construct, they may provide useful insight when answering certain types of key questions, such as questions regarding subpopulations.

Determining How To Use Existing Systematic Reviews

At this point in the process, we assume that EPCs have identified one or more existing SRs that are relevant to the CER and are of adequate quality. Now EPCs must determine the appropriate way to incorporate them into the CER (Figure 1, Step 4). Several possibilities are available (Figures 1 and 2), and they are not mutually exclusive.

- Incorporate already-summarized evidence from existing SRs into the CER.
- Incorporate summarized evidence from existing SRs into the CER but conduct de novo sensitivity analyses. In essence, use an existing SR to answer a key question but then conduct additional analyses using data from the original studies. For example, use an SR to answer a key question in a CER about whether or not to use coenzyme Q10 in heart failure, but then conduct de novo sensitivity analyses to determine the impact of publication date on the results.
- Utilize an SR's search strategy in lieu of a de novo process but then use de novo methods for analysis and synthesis. This would be possible if the search strategy was consistent with the chapter on finding evidence of the Methods Guide, but the quality of other processes were inadequate or could not be determined.
- Build on existing SRs by updating meta-analyses or qualitative syntheses.
- Address conflicting results of existing SRs with a de novo analysis.
- Use at least part of the comprehensive literature search strategy to identify trials or other studies for the CER.

The quality of each step of the existing review is likely to be a major factor in how the EPCs decide to incorporate existing SRs into a CER. The EPC may incorporate an existing SR in its entirety if its research questions are very similar to the CER's key question(s) and are of good quality at all steps of the review. They can also include an SR in part if only a portion is either of interest or relevant to a key question or questions within the CER. This may include incorporating summarized evidence within a specific population or for a specific intervention. In these cases, the methods used in the SR would have to be consistent with the chapters on finding evidence, assessing quality, grading the strength of a body of evidence, and principles in the Methods Guide, including issues of scientific independence and avoiding conflicts of interest.

Previous SRs are unlikely to be wholly sufficient to substitute for a CER because CER questions are identified by a process that assesses the redundancy of a topic with previously published SRs.[25] Moreover, other factors reduce the possibility that existing SRs will be able to answer all the key questions in a CER: the comprehensive and broad nature of many CERs; the need to evaluate efficacy, effectiveness, and harms; the inclusion of high-quality observational studies (often excluded in other SRs) in many CERs; and evaluations based on factors such as sex/gender, race, and/or ethnicity.

In cases where an EPC cannot determine the accuracy or validity of the result of an earlier SR, an EPC may decide to incorporate part of the existing SR, such as the search strategy, the list of included articles, or the data extraction tables, if these sections are felt to be of adequate quality. However, in cases of reporting deficiencies where SRs may not present results of individual trials, using summary findings without complete reporting may compromise transparency in the CER. Little is gained from incorporating full results of such an SR into a CER because EPCs could not update the meta-analyses or conclusions in the existing SR with

more recent trials or studies without obtaining the primary articles and repeating the data abstraction.

If EPCs find that several recent, relevant, and high-quality SRs are appropriate for a given CER, they then need to determine how best to proceed. One approach is to incorporate the single "best" existing SR (most relevant and least biased) into their own reports.[2] However, selecting a single review may pose the risk of introducing selection bias; EPCs must ensure transparency in their criteria for eligibility. Another approach is to conduct a meta-review (also known as an "umbrella review"), whereby they select all relevant, high-quality SRs that meet an a priori publication date threshold and then assess the consistency among them.[26,27] When using this approach, EPCs should provide summary tables with information about all the included SRs so as to maximize transparency. If the selected relevant, high-quality SRs have discordant findings, EPCs should explore the reasons for these disagreements. If EPCs cannot readily give reasons for the discordant findings, then they can regard this as an indication that they need to adopt a de novo approach to answer that key question.

Reporting Methods and Results

This chapter of the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* provides the recommended approach to use when locating existing SRs and assessing their relevance and quality, and it offers a strategy for dealing with multiple existing SRs that EPCs can use to replace a de novo process. We emphasize the need for both reproducibility and transparency when using an existing SR (Figure 1, Step 5). By specifying the targeted search databases and terms used to locate existing SRs and employing a flow diagram to demonstrate the disposition of the citations identified (Figure 2), EPCs can ensure that readers of the CER will be able to assess the process and, if desired, reproduce it. If EPCs decide to search for previous SRs within only a specific date range or to exclude citations based solely on the dates of the existing SR's literature search, then they should specify the rationale for using this cutoff date.

Providing a summary table that specifies the details of included existing SRs used to replace a de novo process is important.[28,29] Summary tables of existing SRs should document the volume, type, and quality of the primary research included. In comparing these previous SRs, ideally the table should address the overlap (or lack of overlap) in primary research in these SRs: e.g., what studies or types of studies were included in one review vs. another. Table 2 is an example. Documenting these points will help readers in assessing such factors and the magnitude of net benefits; it will also clarify how EPCs have graded the strength of a body of evidence.[2] Excluded existing SRs should also be cataloged in a table with the reason for their exclusion.

Table 2. Table template for included SRs

	Included studies (n)	Study types (n)	Total participants (n)	EPC assessment of the quality of primary literature	Overlapping studies (n)[a]	Comments
Reading 2005	7	RCTs, 5 OS, 2	RCTs, 1,175 OS, 2,756	Moderate	Referent	Inclusion criteria not restricted to RCTs.
Preakness 2005	6	RCTs, 6 OS, 0	RCTs, 1,464 OS, 0	High	5 of 7	One additional RCT included in this SR vs. Reading 2005. RCT included after contacting author for additional information.
Hung 2004	4	RCTs, 4 OS, 0	RCTs, 893 OS, 0	Moderate	4 of 7	All of the RCTs in this SR were included in Reading 2005 and Preakness 2005.

Number of overlapping studies using the most recent SR as the referent.
Abbreviations: EPC=Evidence-based Practice Center; OS=observational study; RCT=randomized controlled trial; SR=systematic review.

Discussion: Reiterate Justification for Using Existing Systematic Reviews

In the discussion section of a CER report, EPCs should restate the initial justification for using one or more earlier SRs instead of following a de novo process. They should discuss clearly any limitations arising from the use of existing SRs. Authors should comment on advantages and disadvantages identified through the process of creating the specific CER to help the conduct of future CERs.

Although not the focus of this paper, comparing findings from the CER with the findings from existing SRs is important because it helps health care decisionmakers understand how the CER in question relates to the existing SR literature. Authors can present similarities and differences and discuss potential reasons for any congruities or discrepancies that they have identified.

Future Directions

Many areas require further research to help determine how best to incorporate existing SRs into CERs. These include:

- Determining whether the targeted SR search strategy that has been proposed in this chapter consistently helps to identify the highest quality reviews with less resource allocation than a more broadly conducted search.
- Examining whether applying different relevance or quality criteria markedly changes the SRs that EPCs ultimately include in their CERs or the results derived from these SRs.
- In a situation involving several existing SRs with sufficient relevance and quality, investigating whether the conduct of a meta-review or selecting the best SR approach is the better strategy.
- Documenting savings or increases in time or resources (if any) that come from using an existing SR approach in place of a de novo process.

- Documenting the additional time or resources used in searching for and evaluating existing SRs when they are ultimately not used to replace a de novo process.
- Determining whether it is more efficient to search for an SR as part of the overall search strategy for a topic, or as a first step before searching for primary literature.
- Determining specific criteria to assess the quality of individual patient data meta-analyses.
- Determining if SRs evaluating diagnostic tests or harms require a different emphasis on certain quality criteria or if additional criteria might be warranted.
- Developing and validating criteria for categorizing quality of reviews into good/fair/poor metrics.

Author Affiliations

University of Connecticut/Hartford Hospital Evidence-based Practice Center, Hartford, CT, (CMW). Tufts Medical Center Evidence-based Practice Center, Boston, MA, (SI). Vanderbilt Evidence-based Practice Center, Nashville, TN, (MM). RTI/University of North Carolina Evidence-based Practice Center, Chapel Hill, NC, (TSC). Oregon Evidence-based Practice Center, Portland, OR, (RC). RTI International, Research Triangle Park, NC, (KNL). Johns Hopkins University Evidence-based Practice Center, Baltimore, MD, (KR). Stanford-University of California San Francisco Evidence-based Practice Center, Stanford, CA, (KM). Oregon Evidence-based Practice Center, Portland, OR (EW).

References

1. Moher D, Tetzlaff J, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007;4:e78.
2. Whitlock EP, Lin JS, Shekelle P, et al. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008;148:776–82.
3. Shojania KG, Sampson M, Ansari MT, et al. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007;147:273–4.
4. Rothwell PM. External validity of randomized controlled trials: "to whom do the results of this trial apply?" Lancet 2005;365:13–14.
5. West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. Research Triangle Institute-University of North Carolina Evidence-based Practice Center. AHRQ Publication No. 02-E016. 2002. Rockville, MD: Agency for Healthcare Research and Quality.
6. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.
7. Moher D, Cook DJ, Eastwood S, et al. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999;354:1896–1900.
8. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283:2008–12.
9. Shea BJ, Dube C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In: Egger M, Smith GD, Altman DG, eds. Systematic Reviews in Health Care: Meta-Analysis in Context. 2nd Edition. London: BMJ Publishing Group; 2001.
10. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;44:1271–78.
11. Jadad AR, McQuay HJ. Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. J Clin Epidemiol 1996;49:235–43.

12. Assendelft WJ, Koes BW, Knipschild PG, et al. The relationship between methodological quality and conclusions in reviews of spinal manipulation. JAMA 1995;274:1942–8.
13. Shea B, Dube C, Moher D. Assessing the quality of reports of systematic reviews: the QUORUM statement compared to other tools. In: Egger M, Smith GD, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. London, UK: BMJ Publishing Group, 2001:122–39.
14. Oxman AD, Schunemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 8. Synthesis and presentation of evidence. Health Research Policy and Systems 2006;4:20.
15. COMPUS Procedure. Evidence-based best practice recommendations. Available at: http://www.cadth.ca/media/compus/pdf/COMPUS_%20procedure_e.pdf. Accessed October 29, 2008.
16. Salerno SM, Browning R, Jackson JL. The effect of antidepressant treatment on chronic back pain. Arch Intern Med 2002;162:19–24.
17. Ward NG. Tricyclic antidepressants for chronic low-back pain. Spine 1986;11:661–5.
18. Gotzsche PC, Hrobjartsson A, Marie K, et al. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007;298:430–7.
19. Jones AP, Remmington T, Williamson PR, et al. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol 2005;58:741–2.
20. Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7:1–76.
21. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol 2000;53:964–72.
22. Hartling L, McAlister FA, Rowe BH, et al. Challenges in systematic reviews of therapeutic devices and procedures. Ann Intern Med 2005 Jun;142:1100–11.
23. Shekelle PG, Morton SC, Suttorp MJ, et al. Challenges in systematic reviews of complementary and alternative medicine topics. Ann Intern Med 2005 Jun;142:1042–7.
24. Drug Effectiveness Review Project. Quality assessment methods for drug class reviews for the Drug Effectiveness Review Project. Available at: http://www.ohsu.edu/ohsuedu/research/policycenter/DERP/about/upload/QualityAssessmentDERP-2.pdf. Accessed October 23, 2008.
25. Whitlock EP, Lopez SA, Chang S, et al. Identifying, selecting, and refining topics for research reviews: AHRQ and the Effective Health Care Program. JCE, Submitted.
26. Ruddy R, House A. Meta-review of high-quality systematic reviews of interventions in key areas of liaison psychiatry. Br J Psych 2005;187:109–20.
27. Moe RH, Haavardsholm EA, Christie A, et al. Effectiveness of nonpharmacological and nonsurgical interventions for hip osteoarthritis: an umbrella review of high-quality systematic reviews. Phys Ther 2007;87:1716–27.
28. Chou R, Huffman LH. American Pain Society. American College of Physicians. Nonpharmacologic therapies for acute and chronic low back pain: a review of the evidence for an American Pain Society/American College of Physicians clinical practice guideline. Ann Intern Med 2007;147:492–504.
29. Lorenz KA, Lynn J, Dy SM, et al. Evidence for improving palliative care at the end of life: a systematic review. Ann Intern Med 2008;148:147–59.

Chapter 12. Updating Comparative Effectiveness Reviews: Current Efforts in AHRQ's Effective Health Care Program

Alexander Tsertsvadze, Margaret Maglione, Roger Chou, Chantelle Garritty, Craig Coleman, Linda Lux, Eric Bass, Howard Balshem, David Moher

Key Points

- Comparative Effectiveness Reviews (CERs) need to be regularly updated as new evidence is produced. Lack of attention to updating may lead to outdated and sometimes misleading conclusions that compromise health care and policy decisions.
- The objective of this project was to review the current knowledge and efforts on updating systematic review (SRs) as applied to CERs.
- There is little information about what proportion of SRs needs updating. Similarly, there is no consensus on when to initiate updating and how best to carry it out.
- This paper outlines considerations for updating CERs by providing the following:
  - a definition of the updating process
  - when to update CERs
  - how to update CERs
  - how to present, report, and interpret results from updated CERs
  - current and future research efforts

Background

To maintain relevance, systematic reviews (SRs) need to be regularly updated as new evidence is produced.[1,2] The lack of attention to updating may lead to evidence-based conclusions becoming outdated and sometimes misleading, thus compromising health care and policy decisions. These problems could lead to a waste of resources, provision of redundant or ineffective health care, failure to implement more effective health care, and possibly cause harm. Disseminating the updated reviews will increase the awareness of new findings among relevant stakeholders and the likelihood that new evidence is incorporated into clinical practice. There is little information about what proportion of SRs are in need of updating at any given time, when to initiate updating, or how best to carry it out. Although the Cochrane Collaboration has invested substantial effort in preparing updates and keeping SRs up to date, other groups have published very few updates. One methodological survey,[3] based on 300 SRs indexed in MEDLINE during November 2004, reported that 37.6 percent of the 125 Cochrane SRs and 2.3 percent of the 88 non-Cochrane reviews were updates.

In the absence of a standard method to determine when or how to update any given SR, some organizations have made recommendations about the frequency with which the evidence base needs to be updated. The Cochrane Collaboration has an established policy that reviews be assessed and updated every 2 years, or that a commentary be added to explain why this is done less frequently.[4] Updating all SRs based on an arbitrarily defined time interval could result in inefficient use of resources, as SRs from diverse clinical areas will vary in how frequently they need to be updated depending on the pace of developments occurring in a given clinical area.

The U.S. Preventive Services Task Force (USPSTF) has addressed the issue of updating its clinical guideline recommendations.[5] Because of resource limitations, they set priorities and order in which updates are conducted. This process involves a review of clinical evidence often based on evidence from SRs. A committee determines updating priorities based on the public health importance of the topic (burden of suffering and expected effectiveness of preventive services to reduce that burden), the potential for a USPSTF recommendation to affect clinical practice (based on existing controversy or the belief that a gap exists between evidence and practice), and the availability of new evidence that has the potential to change prior recommendations.

The Drug Effectiveness Review Project, the collaboration between the Oregon Evidence-based Practice Center (EPC) and the Center for Evidence-based Policy of Oregon established in 2003 (http://www.ohsu.edu/xd/research/centers-institutes/evidence-based-policy-center/derp/index.cfm), has conducted SRs of comparative effectiveness and safety for drugs of the same class. The updating process has included an annual scan of literature using the same search strategy as for the previous report, but limited to MEDLINE. After identified article abstracts are reviewed, a decision is made whether to update the report. If the decision is made to update the report, then key questions for potential modifications are assessed to accommodate new evidence (e.g., new drugs, safety alerts, and new indications). The incorporation of newly identified evidence follows the same methodology as one used for an original review report.

The U.S. Agency for Healthcare Research and Quality (AHRQ) faces a similar dilemma in relation to keeping their evidence synthesis research up to date. An important cornerstone of AHRQ's research is the Effective Health Care (EHC) Program of which one of its mandates is to produce Comparative Effectiveness Reviews (CERs). A CER is a type of SR that synthesizes the available scientific evidence on a specific topic, beyond 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.[6] CERs like other SRs are also susceptible to becoming out of date.

This paper reviews current knowledge and efforts on updating SRs as applied to CERs.

Why Update CERs?

Whether a CER needs to be updated depends on many factors, as several reasons may exist for undertaking an update. The most common reason is to include newly published studies or studies that have been updated with information not previously presented. Newly identified studies may report on newly emerged interventions, devices, technologies, diagnostic tests, procedures, harms, and efficacy outcomes. Updating may be conducted to include delayed publications to minimize the impact of time lag bias or to add missing or unpublished data obtained from authors of primary studies.[7] In some cases, the passage of time may bring about new understanding of disease mechanisms that may change the scope of key questions originally asked.

Updates may present a good opportunity to correct various errors or incorporate relevant older evidence in the original CER report, as studies may have been missed by the original searches because of inadequately conducted initial searches or incorrect application of study inclusion/exclusion criteria. In addition, subsequent publications of previously published studies may also provide relevant evidence not presented previously.

Definition of Update

The term "to update" means "to extend up to the present time" or "to include the latest information."[8] Moher and Tsertsvadze proposed a formal definition of update for SRs to mean a discrete event aiming to search for and identify "new evidence" to incorporate into a previously completed SR.[9] Central to updating is the effort to identify such "new evidence," irrespective of date of publication. We take this view to mean any relevant evidence not included in the previously completed review, not just new studies published since the last review. We believe this definition is appropriate given the purpose of CERs, and it is in keeping with the Cochrane Collaboration's definition.[4,10] The authors explain that a feature of an updated review distinguishing it from a new review is that during updating constituent elements of the originally formulated protocol (e.g., search strategy, eligibility criteria, and key questions) may be retained and sometimes extended/modified to accommodate newly identified evidence (e.g., new intervention, new outcome, or new subpopulation).[9]

When To Update CERs

The optimal timing for conducting an update for a CER depends on many factors: rapidity of scientific developments in a given clinical area, nature of the health condition in question, and public health importance. No standard methodology exists for assessing the need for updating a review at a given point in time.[11] Conducting periodic literature surveillance [12] and obtaining expert opinion[13,14] are helpful sources for efficiently identifying new relevant evidence to determine when to update.

Surveillance searching is one common technique to monitor emergence of new evidence for the purpose of updating. Although because of efficiency considerations, surveillance search strategies typically are not comprehensive, they are useful in flagging CERs in need of updating. Sampson and colleagues[12] tested and compared the feasibility and performance of five different surveillance search techniques alone or in combination for identifying relevant new evidence needed for updating SRs. The surveillance searches (i.e., related articles, clinical queries, CENTRAL, core clinical journals, citing article) were carried out for a cohort of 77 SRs. For each surveillance technique, the authors calculated recall (i.e., the proportion of identified relevant studies) and screening burden (i.e., the number of studies to be reviewed to identify relevant evidence for updating). The technique based on the combination of the PubMed-related articles search and subject searching with clinical queries was the most effective approach, yielding 71 new records per review with an inter-quartile range from 42 to 161. Identifying new evidence on harms warrants at least the same rigor in surveillance search as that for benefits; it should be an integral part of the updating process. The databases of peer-reviewed literature should be periodically searched for new studies reporting adverse events or SRs, meta-analyses and HTA reports focusing on harms to achieve greater efficiency with respect to time and resources spent. Drug warnings often based on adverse events data (e.g., case reports, case-series) reported by consumers or medical providers can be found in nationally licensed databases (e.g., U.S. Food and Drug Administration). Such case reports or case-series are not often submitted for journal publication, therefore to supplement searches of the peer-reviewed literature, we recommend searching such databases.[15]

Experts in the field are often aware of new developments before they become public. These developments include new controversies, drugs or devices in development, ongoing trials and observational studies, papers in submission or in press, and reports of adverse events (i.e., case reports). Expert opinion has been used in updating clinical practice guidelines.[16,17] While

reviewers are updating a CER, they may find expert opinion useful as a supplemental source for identifying new evidence.[13] The experts may be asked their opinion about whether the conclusion of any given review is still valid and whether or not they are aware of any new evidence that may change this conclusion.[14]

The body of empirical evidence indicating how frequently or when any given SR needs to be updated is small and inconsistent.[7] For example, findings reported in studies by French[18] and Shojania[19] convey conflicting messages regarding how frequently SRs need to be updated.

French and colleagues[18] surveyed and followed up 362 SRs in the Cochrane Database of SRs from their original publication in 1998 (Issue 2) to 2002 (Issue 2). The authors reported that 70 percent (254/362) of these reviews had been updated during the 4-year period. Of the updated SRs, only 9 percent (23/254) had changes in their conclusions.

Shojania and colleagues[19] proposed several quantitative and qualitative signals indicating when any given SR needs updating. They defined a quantitative signal as a change in statistical significance for an effect estimate using a conventional threshold of $\alpha=0.05$ or a relative change of $\geq 50\%$ in the magnitude of an effect. The authors defined a qualitative signal as a qualitatively different characterization of effectiveness that affects clinical decisionmaking (e.g., a new harm, a new alternative therapy, expansion of treatment to a new patient subgroup). The median time to a qualitative or quantitative signal for updating of 100 SRs was 5.5 years (95% CI: 4.6-7.6). Twenty-three percent of SRs had signals indicating the need for updating within 2 years, 15 percent within 1 year, and 7 percent at the time of publication. The odds of signals for updating were significantly higher for cardiovascular topics than for other topics. This work suggests the presence of several indicators that likely coexist to varying degrees, and it highlights the potential of signal detection in the updating process. The identification of a qualitative signal requires far fewer resources than determination of a quantitative signal.

In 2008, AHRQ asked the Southern California Evidence-based Practice Center (SCEPC) to determine whether 11 AHRQ-funded CERs representing different clinical areas and published since 2005 needed updating.[14] To assess the need for updating for specific CERs, SCEPC applied a modification of a method proposed by Shekelle and colleagues,[16] which is a combination of abbreviated literature review of several preselected, high-impact generalist, and specialty peer-reviewed journals for each clinical area, expert opinion, and the review of U.S. Food and Drug Administration (FDA) Web site. For each CER, the recommendations for updating (e.g., needs updating now, may need updating in future, no need for updating now) were based on changes in four indicators: (a) evidence on the benefits and harms of existing interventions, (b) available interventions, (c) outcomes considered important, and d) evidence that current practice is optimal. Of the 11 CERs published in 2005 or later, 4 were recommended for current updating and 4 for future updating, and the remaining 3 were deemed not in need of updating for some time.

How To Update CERs

If new studies are published, new harms have emerged, a new more effective intervention(s) is introduced, or existing (or new) interventions are extended to new patient groups, the question of updating for an individual EPC moves from “when to update,” which may be based on priorities and available resources, to “how to update.”

The updating process for any given CER can be viewed as a continuum stretching over a wide range of activities from a single update search to a comprehensive expanded search including old and new searches and incorporating new evidence across all sections of a CER.

Moreover, the updating process may be different for CERs with and without meta-analysis in terms of updating scope, methodology, and amount of needed resources.

Therefore, the rational choice of the scope for an update search will depend largely on where a given investigator stands along the continuum of updating process and available resources allocated to updating.[20]

Assessment of Key Questions and Constituent Elements for an Update

Because medical disciplines are constantly evolving through emergence of new evidence, it is recommended that reviewers assess the key question(s) of the original CER at the initial stage of updating. Specifically, they should determine the extent to which the constituent elements of the key research question(s) denoting Population, Intervention, Comparator, and Outcome (PICO) may have changed. If an update search does not identify any relevant evidence, the key question(s) and CER section(s) of the original report will not be modified. However, the status of the CER will be registered as 'updated' by including information on the search dates and time-periods covered by the search.

When newly identified evidence does not entail the modification of any PICO elements of a key question (e.g., no new subpopulation, no new intervention, or no new outcome was identified), the update process will consist of only incorporating this evidence into relevant sections of the report (e.g., Results and Conclusion). However, if newly identified evidence includes a new PICO element (e.g., new harm and/or new subpopulation was identified), the inclusion/exclusion criteria will need to be extended and the key question(s) modified with respect to the given PICO element in order to accommodate this evidence in relevant sections of the updated CER (e.g., Methods, Results, and Conclusion). The identification of evidence on the same intervention, comparator, and outcome as specified in a key question of the original CER, but for people with a newly identified health condition, would not be an update of the previous CER, since it entails the exploration of a new key question.

The assessment process of the updating scope and corresponding modifications are depicted in Table 1.

Table 1. Scope of updating and corresponding actions using original or modified search strategy

Scope of Newly Identified Evidence Warranting an Action to Update	Action for a Key Question	Changes After Updating (Updated vs. Original CER)
Search performed but no evidence	None	No change in the CER or KQ KQ status = updated
Evidence from new studies (without identification of a new PICO element)	Update Results and Conclusion sections	No change in KQ Updated Results and Conclusions sections
New evidence from already included studies (without identification of a new PICO element)	Update Results and Conclusion sections	No change in KQ Updated Results and Conclusions sections
Identification of a new PICO element New subpopulation(s) only New intervention(s) only New comparator(s) only New outcome(s) only	Update Methods, Results and Conclusion sections Extend the inclusion/exclusion criteria for the population the intervention the comparator the outcome	Modify KQ with respect to a new PICO element (population, intervention, comparator, or outcome) Updated Methods, Results and Conclusions sections

Abbreviations: CER=comparative effectiveness review; PICO=Population/Intervention/Comparator/Outcome; KQ=key question

General Search Strategies for Updating CERs

Once a decision has been made to conduct an update of a CER, it is important to perform comprehensive searches that adhere to the general principles for conducting a systematic search as recommended in the AHRQ methods guide.[15] This includes searches of multiple literature sources (e.g., SRs, bibliographic databases, Web sites, allied health professional databases, pharmacoepidemiologic databases, governmental regulatory cites, scientific information packets, and miscellaneous resources). The guide recommends searching several major bibliographic databases such as MEDLINE, EMBASE, CINAHL, Cochrane CENTRAL, and PsycInfo.[15] Some authors suggest the search of other supplemental sources such as reference lists of key citations.[13]

Moreover, there are some specific approaches to searching listed below that are particularly relevant to the process of updating. During any given update, the original search strategy can frequently be carried over to the update. Investigators should also use the opportunity to review the search strategy and modify search terms, databases and other sources searched, if necessary, and have it peer-reviewed, if not previously done.[21] For example, use of governmental and nongovernmental clinical trials registries has expanded; their inclusion could provide useful information on in-progress or unpublished trials as well as unpublished outcomes.[22,23] Investigators should also consider previous decisions regarding the inclusion/exclusion of grey literature, non-English language literature, or other sources of evidence.[24,25] Additional information worth considering in updating may be requested through contacting manufacturers of pharmaceutical or biotechnical products.

To limit the number of citations to review, one strategy is to limit the start date for update searches. However, delays between publication in journals and indexing in MEDLINE and other electronic databases occur and are variable in duration.[26] Therefore, we recommend that reviewers use a start date at least 1 year before the end date of the original search. Searches could be based on the "entry date" (date the publication was added to MEDLINE) rather than the publication year.[27] This search technique results in more complete retrieval of relevant records, including those that have become available since the date of the last search, thereby minimizing publication bias.

When newly identified evidence through an update includes a new PICO element (e.g., new harm, new subpopulation), resulting in corresponding modifications to the key question(s), it is recommended that a repeated search covering the start date of search for the original CER be conducted to ensure there are no missed studies reporting the new PICO element.

Statistical Methods Relevant to Updating Meta-Analyses

Updating or assessing the need for updating a meta-analysis as a part of any given CER will necessitate the use of statistical method(s). A recent SR surveyed and appraised various methods and/or strategies describing the process of updating SRs.[7] This review identified two statistical methods (cumulative meta-analysis and identifying null meta-analyses ripe for updating).[28-31]

Cumulative meta-analysis (CMA) is a statistical procedure in which the combined effect estimate is sequentially updated by incorporating results from each newly available study.[29-31] This technique documents trends in a treatment effect over time and provides up-to-date information. When done prospectively, it may be useful in identifying the earliest time at which the statistical evidence that an intervention is effective or harmful is sufficient.[30] However, CMA can be costly and time consuming, and it may pose the potential for an inflated rate of type-I

error arising from repeated hypothesis testing.[32] Moreover, the use of this procedure is limited only to instances when all PICO elements of the key question remain constant over time. In one extension of CMA proposed by Mullen and colleagues,[33] a least-squares regression line is fitted to points corresponding to the effect size for each successive cumulatively added study. The slope of this line helps reviewers to gauge the stability of effect size (including no effect) more objectively than through visual inspection. The cumulative slope is a useful tool in determining when the updating process should stop to avoid waste of resources in the absence or presence of effect for any given health intervention.

Barrowman and colleagues[28] proposed a method to assess whether the amount of new evidence that has accrued is sufficient to turn a statistically nonsignificant meta-analytic result into a significant one, thereby rendering the meta-analysis in question "ripe for updating." Thus, this approach helps to identify meta-analyses with negative results (i.e., non-significant pooled estimate) in need of updating. It requires searching, screening, and only partial data extraction (i.e., number of newly identified additional participants), rather than a complete updating implemented through addition of each new study. Depending on the configuration of computer simulation, this approach was shown to classify correctly whether a statistically nonsignificant result of a meta-analysis was outdated with a sensitivity ranging from 49 percent to 62 percent and a specificity ranging from 80 percent to 90 percent.

Evolution of Methods When Conducting an Update

Methods used to conduct CERs (e.g., methods for pooling, assessing the risk of bias, grading the strength of evidence) continue to evolve. If some methods have changed between the original and the to-be-updated CERs, we recommend that investigators compare the methods used in the original CER with the newly developed methods. If the new methodology is an obvious improvement over the older one, the CER team should ideally rereview (e.g., appraise, grade) all previously and newly included studies using the new methodology for sake of consistency between the assessments and conclusions of the original and updated review.

Moreover, critical feedback obtained on the original review can provide useful information regarding correct choices for the analyses the reviewers might consider conducting in an updated CER. For example, if a CER is criticized for its use of a fixed-effect over random-effects model for pooling results of individual studies, conducting sensitivity analyses using both pooling methods (or only random-effects model, if deemed appropriate) in the update might be reasonable.

Incorporating New Evidence and Reporting an Update

After reviewers identify new evidence, they must incorporate it into the update. The amount of resources, complexity of methods, and logistic efforts needed for incorporation of an update in a CER will depend on the amount of newly identified evidence (e.g., number of new studies) and the degree of consistency of evidence-based findings in the original versus the updated CER.

One commonly used approach is to incorporate the new evidence into the previous review by updating results (i.e. search yield, number of studies, quality assessments, effect estimates, and conclusions) and other respective sections of the review as appropriate. The reviewers can summarize the updated evidence in a distinct section at the end of the review (i.e., "summary of update results and discussion" sections).

To make updates most useful to readers, reviewers need to describe clearly the purpose of the update, the methods used to conduct it, and the results. Reviewers should explicitly note any changes in the scope, methods, and understanding of the mechanism of an intervention's action on a disease for the key question in the updated versus the original review. The rationale for introducing any new methodology or different conceptual framework in the updated report compared to the original one also needs to be described. Important elements to focus on include the search strategy (including sources, search terms, the start and end dates covered by searches), the yield of the searches, important characteristics of new evidence (number, type, size, and quality of studies; study participants; outcomes), and main results, including how the conclusions of the update differ from those of the original review. Evidence that has the most impact on the conclusions of the update should be emphasized and described in detail. If reviewers have not identified new evidence for part of the review, they should still update the report by including all the details of last search (see above), results of search yield (e.g., no new studies), and the currency of the conclusions (i.e., no change and still judged to be accurate). When incorporating evidence on a new intervention, outcome or subpopulation group, we suggest adding a new section in the Results chapter of the CER report.

For more efficient presentation of update results, we suggest including a summary table (Table 2, given as an example) and the PRISMA study flow diagram[34] in the CER report. Currently, the SCEPC is developing the recommended format of the summary table.

The updating process will have optimal credibility if it is conducted and reported transparently. To ensure continued transparency, the EHC Program should publish the titles of CERs selected for updating. Updated CERs should include a description of how they were updated. There should be adequate opportunity provided for public comment on both the CERs chosen for updating as well as subsequent updated draft reports. Posting a list of key questions for CERs that will be updated will ensure that a broad range of stakeholders (e.g., biopharmaceutical and device manufacturers, governmental agencies, academic institutions) have the opportunity to provide relevant new evidence that the project team might consider as informative to the decisionmaking process.

Table 2. Example of a summary table for an update of key questions within comparative effectiveness review

Comparison (Design)	2001 Report			2009 Update				Did the conclusion for KQ change?
	Outcome (binary) and population	N studies	Summary result	N new studies	Summary Result	New PICO element(s)	Conclusion	
‘A’ vs. ‘No Tx’ (RCTs)	Outcome-1 (e.g., efficacy) Sub-population-1 (e.g., males)	5	1.5 (1.1, 1.7)£ N=5	2	1.4 (1.2, 1.6) N=7	None	‘A’ more effective than ‘No Tx’ in males	No
	—	—	—	1	1.6 (1.2, 2.0)	Outcome-2 (e.g., new harm) in subpopulation-1 (e.g., males)	‘A’ more harmful than ‘No Tx’ in males	KQ may need modification to accommodate new results
	—	—	—	2	1.7 (1.1, 2.3) N=2	Outcome-1 (e.g., efficacy) in subpopulation-2 (e.g., females)	‘A’ more effective than ‘No Tx’ in females	
	—	—	—	1	1.1 (0.7, 1.3)	Outcome-2 (e.g., new harm) in subpopulation-2 (e.g., females)	No evidence that ‘A’ is more harmful than ‘No Tx’ in females	
‘A’ vs. ‘PL’ (RCTs)	Outcome-1 (e.g., efficacy) Sub-population-1 (e.g., males)	3	0.9 (0.8, 1.4) N=3	0	0.9 (0.8, 1.4) N=3	None	No evidence of difference in efficacy between ‘A’ and ‘PL’ in males	No
‘A’ vs. ‘B’ (Non-RCTs) µ	Outcome-1 (e.g., efficacy) Sub-population-1 (e.g., males)	2	2.3 (1.5, 3.4) 1.2 (0.7, 1.9)	2	1.6 (1.1, 3.0) 2.0 (1.2, 3.3) 2.3 (1.5, 3.4) 1.2 (0.7, 1.9)	None	Some evidence that ‘A’ more effective than ‘B’ in males	Yes
‘A’ vs. ‘C’ (RCTs)	—	—	—	3	1.1 (0.9, 2.2) N=3	New treatment ‘C’ for outcome-1 (e.g., efficacy) in subpopulation-1 (e.g., males)	No evidence of difference in efficacy between ‘A’ and ‘C’ in males	KQ may need modification to accommodate new results

Abbreviations: N=number; PL=placebo; Tx=treatment; RCT=randomized controlled trial; KQ=key question

µ Trials could not be pooled due to heterogeneity in methodology of their conduct

F Bold and not bolded fonts denote pooled and individual study point estimates of relative risk (95percent confidence interval), respectively

Issues of Authorship and Challenges of Updating CER

Ideally, the original CER authors should be asked to conduct the update. But this approach may be problematic for many reasons. Over time, authors may be working on new topics, may have changed institutions or affiliations, or may not be interested in updating already published CER. Garritty and colleagues found that of the health care agencies and organizations involved in conducting SRs that were surveyed, only 54 percent (56/103) were able to draw on the same authors of the original review for updating.[11] This phenomenon poses significant problems for the cost, time, and practicality of an update. Naturally, new reviewers would require additional time to become familiar with a CER. In addition, knowledge of project history would be diminished or perhaps lost, and issues of replication and transparency could arise if the original CER was not well reported. These factors combined would add to costs and jeopardize the feasibility of updating.

If an update involves new authors, it is important to discuss author issues as early in the updating process as possible. One objective would be to ascertain the level of involvement and authorship of the original CER team in the update. These discussions can be informed by examining current international policies and guidance on authorship suggested by the International Committee of Medical Journal Editors (http://www.icmje.org) and contributions of authors.[35]

Current and Future Research Efforts

In the near future, a standardized guideline for updating of CERs applicable across EPCs across the range of health care interventions and treatment modalities (e.g., devices, pharmaceutical products, surgery, diagnostic tests, and other procedures) is needed. This guideline could incorporate a step-wise use of selected updating strategies and methods that have been empirically shown as valid, reliable, and resource-efficient. Ideally, such a guideline would include specific recommendations on three important dimensions: (1) setting updating priorities based on factors such as public health burden, severity of health condition, number of outdated key questions for a given CER; (2) clarifying the responsibilities and authorship (especially when authors of the original report change their institutional affiliations or are difficult to locate) for updating CERs; and (3) implementing the updating process (e.g., triggers for updating, timing and sources for evidence surveillance).

To date, there has been insufficient research to inform which strategy or method used for updating is most reliable, applicable, and cost effective.[7] Future research should compare different approaches used for updating evidence to help to identify most robust and efficient strategies and methods to carry out updating. Furthermore, methods developed in other fields (e.g., health economics, bibliography) need to be considered to inform when and how to update CERs. For example, value-of-information analysis may determine a benefit for making a decision to update a CER in terms of reduced uncertainty even if conclusions of the original CER are unchanged.[36]

As an ongoing effort, the EPCs of Tufts Medical Center, Southern California, and University of Ottawa have jointly piloted and elaborated the process of assessing the need of updating for selected CERs by comparing two methods developed at the SCEPC-based Research and Development corporation (the RAND method)[14] and University of Ottawa (the Ottawa method).[19] The RAND method is based on the combination of external domain expert opinion, an abbreviated search, and determination of the validity of conclusions in the original

CER. The Ottawa method relies on the identification of qualitative and quantitative signals through literature search used in the original report but limited to five major general-interest medical journals, supplemented with a small number of specialty journals. If the original report includes a meta-analysis, a quantitative signal is considered.

Based on the previous work,[14,19] the EPCs of Southern California (RAND), University of Ottawa, and Emergency Care Research Institute initiated a joint collaboration to develop and implement a system of ongoing literature surveillance to identify triggers (or signals) for updating systematic reviews within the EPC program of the AHRQ. This project is being coordinated across the three participating centers to ensure consistency in application of methods.

This joint collaboration emphasizes the importance and usefulness of international harmonization of the updating process for maintaining, modifying, and disseminating the updated findings of CERs in future.

Author Affiliations

University of Ottawa Evidence-based Practice Center, Ottawa, Ontario, Canada (AT, CG, DM). RAND Corporation–Southern California Evidence-based Practice Center, Santa Monica, CA (MM). Oregon Evidence-based Practice Center, Portland, OR (RC, HB). University of Connecticut Evidence-based Practice Center, Hartford, CT (CC). RTI–University of North Carolina Evidence-based Practice Center, Research Triangle Park, NC (LL). Johns Hopkins Bloomberg School of Public Health Evidence-based Practice Center, Baltimore, MD (EB).

This paper has also been published in edited form: Tsertsvadze A, Maglione M, Chou, R, et al. updating Compartive Effectiveness Reviews: Current Efforts in AHRQ's Effective Health Care Program. *J Clin Epidemiol* 2011 Nov;64(11):1208-1215. PMID: 2168114.

References

1. Chalmers I, Enkin M, Keirse MJ. Preparing and updating systematic reviews of randomized controlled trials of health care. Milbank Q 1993;71(3):411–37.
2. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. BMJ 1994;309(6958):862–5.
3. Moher D, Tetzlaff J, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007;4(3):e78.
4. Higgins JPT, Green S, Scholten RJPM. Chapter 3. Maintaining reviews: updates, amendments and feedback. In: Higgins JPT, Green S, editors. Cochrane Handbook For Systematic Reviews of Interventions Version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available at: http://www.cochrane-handbook.org. Accessed May 15, 2011.
5. Guirguis-Blake J, Calonge N, Miller T, et al. Current processes of the U.S. Preventive Services Task Force: refining evidence-based recommendation development. Ann Intern Med 2007;147(2):117–22.
6. Agency for Healthcare Research and Quality. Effective Health Care Program. 2010. Available at: http://www.effectivehealthcare.ahrq.gov. Accessed February 23, 2011.
7. Moher D, Tsertsvadze A, Tricco AC, et al. A systematic review identified few methods and strategies describing when and how to update systematic reviews. J Clin Epidemiol 2007;60(11):1095–1104.
8. Merriam-Webster's Collegiate Dictionary. 10th ed. Springfield, Massachusetts: Merriam-Webster. 1996.
9. Moher D, Tsertsvadze A. Systematic reviews: when is an update an update? Lancet 2006;367(9514):881–3.

10. Higgins JPT, Green S. editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available at: www.cochrane-handbook.org.
11. Garritty C, Tsertsvadze A, Tricco AC,et al. Updating systematic reviews: an international survey. PloS one 2010;5(4):e9914.
12. Sampson M, Shojania KG, McGowan J, et al. Surveillance search techniques identified the need to update systematic reviews. J Clin Epidemiol 2008;61(8):755–62.
13. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005;331(7524):1064–5.
14. Shekelle P, Newberry S, Maglione M, et al. Assessment of the Need to Update Comparative Effectiveness Reviews: Report of an Initial Rapid Program Assessment (2005-2009). Rockville, MD: Agency for Healthcare Research and Quality. 2009.
15. Relevo R, Balshem H. Finding Evidence for Comparing Medical Interventions. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC021-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2011. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=605. Accessed May 15, 2011.
16. Shekelle P, Eccles MP, Grimshaw JM,et al. When should clinical guidelines be updated? BMJ 2001;323(7305):155–7.
17. Gartlehner G, West SL, Lohr KN, et al. Assessing the need to update prevention guidelines: a comparison of two methods. Int J Qual Health Care 2004;16(5):399–406.
18. French SD, McDonald S, McKenzie JE,et al. Investing in updating: how do conclusions change when Cochrane systematic reviews are updated? BMC Med Res Methodol 2005;5:33.
19. Shojania KG, Sampson M, Ansari MT, et al. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007;147(4):224–33.
20. Garritty C, Tricco A, Sampson M, et al. Updating Systematic Reviews: the Policies and Practices of Health Care Organizations Involved in Evidence Synthesis. [MSc thesis]. University of Toronto; 2009.
21. Sampson M, McGowan J, Cogo E, et al. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009;62(9):944–52.
22. DeAngelis CD, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 2004;292(11):1363–4.
23. Manheimer E, Anderson D. Survey of public information about ongoing clinical trials funded by industry: evaluation of completeness and accessibility. BMJ 2002;325(7363):528–31.
24. Bennett DA, Jull A. FDA: untapped source of unpublished trials. Lancet 2003;361(9367):1402–3.
25. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol 2000;53(9):964–72.
26. McAuley L, Pham B, Tugwell P, et al. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000;356(9237):1228–31.
27. Bergerhoff K, Ebrahim S, Paletta G. Do we need to consider 'in process citations' for search strategies? 12th Cochrane Colloquium. October 26, 2004; Ottawa, Ontario, Canada.
28. Barrowman NJ, Fang M, Sampson M, et al. Identifying null meta-analyses that are ripe for updating. BMC Med Res Methodol 2003;3(1):13.
29. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 1992;327(4):248–54.
30. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol 1995;48(1):45–57.
31. Baum ML, Anish DS, Chalmers TC, et al. A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. N Engl J Med 1981;305(14):795–9.
32. Chalmers T. Problems induced by meta-analyses. Stat Med 1991;10(6):971–9.
33. Mullen B, Muerllereile P, Bryant B. Cumulative meta-analysis: a consideration of indicators of sufficiency and stability. Pers Soc Psychol Bull 2001;27:1450–62.

34. Moher D, Liberati A, Tetzlaff J, et al. The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 2009;6(7):e1000097.
35. Rennie D, Flanagin A, Yank V. The contributions of authors. JAMA 2000;284(1):89-91.
36. Claxton K, Ginnelly L, Sculpher M, et al. A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme. Health Technol Assess 2004;8(31):1–103, iii.

Chapter 13. Avoiding Bias in Selecting Studies

Marian McDonagh, Kim Peterson, Parminder Raina, Stephanie Chang, Paul Shekelle

Key Points

- One hypothesis-testing study and numerous case examples indicate that operational criteria guiding the selection of studies into a systematic review (SR) or meta-analysis can influence the conclusions.
- Assessments of how this source of bias can be reduced, or even the magnitude of the bias, are not available.
- In the absence of conclusive evidence about how to reduce this potential for bias, we recommend that inclusion criteria be clearly described in detail sufficient to avoid inconsistent application in study selection and that inclusion criteria be documented in a protocol.
- We propose hypothetical examples that illustrate how selection of inclusion and exclusion criteria may introduce bias.
- Experience suggests that dual review can identify inclusion criteria that are not sufficiently clear and occasions where subjective judgment may differ. Gray literature (e.g., U.S. Food and Drug Administration [FDA] documents, trial registry reports) can help identify and possibly reduce bias from publication bias or selective outcome reporting.

Background

Much has been written about the importance of various aspects of the conduct of a SR: how to best search computerized databases; whether or not reviewers should be masked to the authors and journals and outcomes of studies being reviewed; how to assess studies for the risk of bias; and the strengths and weaknesses of various different methods of statistically combining the results. The Methods Guide for the Agency for Healthcare Research and Quality (AHRQ)

Evidence-based Practice Center (EPC) Program has chapters summarizing the literature and best-practices advice on numerous such aspects of a SR.[1]

We are concerned here with the potential for bias at a point upstream in the SR process—namely what is the effect of going from the initial question of interest ("what is the effect of intervention X on condition Y?") to the operational aspects of the review (such as selecting inclusion/exclusion criteria). For example, in a recent Comparative Effectiveness Review on drugs to treat low bone density, the EPC identified nine prior meta-analyses evaluating the antifracture efficacy of alendronate compared with placebo or no treatment.[2] The meta-analyses were published between 1997 and 2009, and included between them 17 randomized controlled trials (RCTs) published between 1994 and 2004. One might expect that all the trials included in earlier meta-analyses would be included in later meta-analyses, but this is not the case. One meta-analysis published in 2002 included 10 trials, while another published in 2004 included only 5: 4 were among the 10 trials in the 2002 meta-analyses, but 1 trial (published in 1998) was not. Some of the differences in trial inclusion could be explained by whether data were included

on vertebral and nonvertebral fractures; whether nonvertebral fractures were treated as a general group; whether nonvertebral fractures were split out into fractures of the hip or wrist; or whether patient populations were considered as secondary prevention or as primary prevention. These differences in which trials were included led to differences in conclusions. In one meta-analysis,[3] the conclusion was that the decrease in nonvertebral fractures was not statistically significant. In another meta-analysis[4] published 3 years earlier, the conclusion was that the beneficial effect of alendronate compared with placebo on nonvertebral fractures was statistically significant. All EPCs can tell similar stories.

Conflicting conclusions confuse decisionmakers, especially if all reviews purported to answer the same question and the differences in the applicability of the evidence are not clearly denoted. Bias results from systematic alteration from the truth. Although we do not know the exact truth, different conclusions lead readers to believe that alternate inclusion and exclusion criteria result in biased conclusions. In order to investigate the potential for this source of bias and identify methods studies that investigate how best to reduce it, we searched for studies that examined two or more SRs of the same topic, evaluating the impact of variation in study inclusion.

We found a very small number of relevant studies (Table 1).[5-8] The most relevant example was a prospective study designed to examine reproducibility between two review groups (on different continents) commissioned to review evidence on the same question, using a common methods specification manual.[8] While the manual outlined the important features of inclusion criteria, the specific criteria used by each group are not reported. Search terms were specified a priori, and the groups were instructed to find and include all study designs, including non- English language, case series, ecological, cross-sectional, case-control, cohort, and intervention studies. Both review groups agreed on including 166 articles, but disagreed on 72 articles (Center A included 52 papers not included by Center B, and Center B included 20 papers not included by Center A). Sixty-three of the 72 discrepancies occurred in screening title and abstract; 9 of the 72 discrepancies occurred during review of full-text articles. Other similar retrospective studies also found differences in their lists of included studies and sometimes different conclusions (Table 1). Although the amount of evidence is small to confirm the presence of bias, the potential for bias is possibly quite large.

Table 1. Studies evaluating reasons for discrepancies in included studies among systematic reviews

Study	Study Aims	Evaluation
Hopayian K and Mugford M (1999) Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed?[6]	The aim of this study was to find the reasons for the discordance between two reviews focusing on use of epidural steroid injection for treatment of low back pain and sciatica and to draw conclusions for users of these reviews.	Each review excluded two papers that the other included, both of which supported the ultimate conclusions of the review that included them. One of these studies was published in a non-English language journal and was excluded by one review. The other papers, however, were published in well-known journals. One of these papers was excluded from one review due to problems with extracting the data, while the other review was qualitative and did not require these data to come to a conclusion. The outcome measures included, and inclusion of non-English language papers account for at least some of the differences.
Peinemann F, McGauran N, et al (2008). Disagreement in primary study selection between systematic reviews on negative pressure wound therapy.[7]	The objective of this study was to compare systematic reviews on negative pressure wound therapy with regard to their agreement in inclusion of primary studies.	The authors conclude that the reviews differed in inclusion of studies, primarily the inclusion of studies other than nonrandomized controlled trials. They indicate that the differences arise from differences in methodology, classification of study design, and style of reporting excluded studies. Our analysis of this example showed that included study designs varied among reviews. However, only one of the five reviews concluded that evidence supported the use of the treatment, while the others consistently found that the evidence was insufficient, largely due to concerns over quality. The review that found treatment to be effective had the broadest inclusion criteria with respect to study design and ultimately included 25 papers, compared with 14, 6, 6, and 7 included in the other reviews.
Cook DH, Reeve BK, et al. (1996) Stress Ulcer Prophylaxis in Critically ill patients: resolving discordant meta-analyses.[5]	This study aimed to resolve discrepancies in four previous systematic reviews and provide estimates of the effect of stress ulcer prophylaxis on gastrointestinal bleeding, pneumonia, and mortality in critically ill patients.	From abstract: "The source of discrepancies between prior meta-analyses included incomplete identification of relevant studies, differential inclusion of non-English language and nonrandomized trials, different definitions of bleeding, provision of additional information through direct correspondence with authors, and different statistical methods." Our analysis of these reviews focused on the prevention of stress ulcer bleeding, as this outcome was common across the reviews. The definition of bleeding differed among reviews. Two more recent reviews came to very different conclusions that can be directly related to the inclusion criteria. One review included both randomized and "quasi-randomized controlled trials," while the other review included randomized controlled trials with at least 10 subjects per arm published in a variety of languages. In this example, the difference in conclusions in appears to be related largely to inclusion of non-English language articles in one but not the other.

Other authors have addressed reasons for discrepant results from meta-analyses on the (seemingly) same topics.[9,10] Ioannidis has examined multiple such scenarios and concluded that the reasons for discrepancy are typically multifactorial, but include differing study questions and inclusion criteria as well as differences in the process of applying the criteria in study selection. He gives examples of situations where inclusion criteria for meta-analyses were apparently specified in way that would obtain results that supported the viewpoints of the authors rather than reflecting questions of clinical uncertainty.[9]

As part of the EPC Methods Guide, we intend that this paper will guide EPCs when selecting studies for inclusion in an SR. Guidance is intended to reduce inconsistencies and risk of bias. Unfortunately, because there are no available studies to guide us how best to reduce this variation, what follows is based on fundamental principles of SRs and the experience of the EPC program.

Inconsistencies and bias can certainly occur during the development of key questions, which define the scope of the review and details the population(s), intervention(s), comparator(s), outcome(s), timing, and setting (PICOTS), and sometimes even the study designs or study characteristics of interest. The methods used by the EPC program at this earlier stage are discussed elsewhere.[11] Likewise, we recognize that bias can also be introduced during the searching stage,[12] or in how reviewers handle assessment of reporting biases,[13] and guidance on these methods are provided elsewhere.[11,12] This paper focuses on what to do with the literature once it is identified. We first describe the types of bias then stratify the guidance on addressing these biases into sections: Setting Inclusion Criteria to Avoid Bias in Selecting Studies, Study Selection Process, and Using Gray Literature to Assess and Reduce Bias.

Types of Potential Biases in Selecting Studies

Spectrum Bias

The inclusion or exclusion of a specific population can have a dramatic impact on the conclusions for the effectiveness of a treatment. For example, while one meta-analysis found no significant benefit of the invasive treatment for coronary artery disease over conservative treatment, a subsequent meta-analysis by invasive cardiologists found significant benefit with invasive treatment when they included patients with unstable angina, a population in which invasive management is known to be more beneficial.[9]

Publication bias and outcome reporting bias can have implications for the conclusions of a review. Bias in selection of studies may overlap with these biases, but methods for avoiding them are addressed in other chapters.[13,14]

Random Error

Even when reviewers have a common understanding of the selection criteria, random error or mistakes may result from individual errors in reading and reviewing studies.

Guidance for Setting Inclusion Criteria To Avoid Bias in Selecting Studies

Although setting inclusion criteria based on key questions may seem straightforward, the experience in the AHRQ EPC program has shown that this is often not the case. The AHRQ EPC

program has an explicit process of systematic review development called Topic Refinement. Its goal is the development of inclusion criteria based on the Key Questions via a process that involves the review team and technical expert panel input.

One of the main goals in developing inclusion criteria is to minimize ambiguity. Greater ambiguity in inclusion criteria increases the possibility of poor reproducibility due to many subjective decisions regarding what to include, potentially resulting in at least random error in study selection.

The criteria should be set a priori and based on the analytic framework or conceptual model using a protocol.[15-17] The benefits of using a protocol specific to SRs include improving transparency and rigor of SRs, and important to this chapter, reducing bias in study selection decisions. Requirements for SR protocols for reviews conducted by EPCs are currently undergoing further development in coordination with other organizations (e.g., Institute of Medicine and PROSPERO). The protocol should be based on a standard set of elements, publicly available, ideally through a SR Registry, (e.g. PROSPERO, www.crd.york.ac.uk/prospero/).

However, there is a balance to be struck between making the inclusion criteria so narrow that it is unlikely that eligible evidence will be found and so loosely defined that it increases the possibility of poor reproducibility due to many subjective decisions regarding what to include. EPCs should attempt to strike this balance, but recognize that there will be times when their initial attempt is not working and changes need to be made. All eligibility criteria decisions should be reported transparently in the published SR.

Selecting PICOTS Criteria

In addition to random error from ambiguous definition of criteria, the selection of PICOTS inclusion or exclusion criteria can introduce systematic bias. A systematic review starts with a broad comprehensive search and the choice of which studies to include can directly influence the resulting conclusions. The EPC should carefully consider whether PICOTS criteria are effect modifiers and how inclusion and exclusion criteria may potentially skew the studies and thus results reported in the review.

Table 2 below suggests potential implications or biases that may result from specific hypothetical examples of inclusion and exclusion criteria.

Table 2. Hypothetical examples of potential for bias based on inadequately defined PICOTs

PICOTS Criterion	Inclusion Criterion	Potential for Bias in Selecting Studies for Review	Possible Biased Result
Population	Population is described as patients with heart failure	The reviewer may have to decide which classes of heart failure the question was meant to whether these different severities are meant to be combined or evaluated separately.	Reviewer chooses to include only Class III and IV heart failure and finds that the intervention is effective, whereas conclusions on effectiveness may have been diluted if all severity classes had been included.
Intervention	Intervention described as anticoagulants	Reviewer must make the decision on which interventions are considered anticoagulants; e.g., may combine oral and injectable anticoagulants.	Combining oral and injectable anticoagulants may be inappropriate for short term effectiveness and harms and may overestimate benefits for oral anticoagulants and underestimate harms for short term effects.

Table 2. Hypothetical examples of potential for bias based on inadequately defined PICOTs (continued)

PICOTS Criterion	Inclusion Criterion	Potential for Bias in Selecting Studies for Review	Possible Biased Result
Comparator	Not defined	Reviewer makes choice among other interventions include in review, interventions excluded from the review, and how to handle placebo, or no treatment, groups.	Reviewer includes only placebo or no treatment groups and concludes that the intervention is effective, whereas it may be less effective in comparison to existing interventions.
Outcome	Described as effectiveness outcomes	Reviewers determine whether specific outcomes are in fact effectiveness. For example, cognitive testing using laboratory settings.	Reviewers report information on intermediate or surrogate outcomes and fail to report lack of effectiveness outcomes, thus making the intervention seem more effective than if clinical outcomes are considered.
Timeframe	Not defined	Reviewers may report whatever is available in the literature, which may be short-term studies.	Without prespecifying that long term outcomes are essential and only reporting short term outcomes, reviewers may overestimate effectiveness of treatment. Also secular trends may mean that older studies may either over or under estimate the effect of an intervention depending on changes in standard of care, technology, or disease epidemiology.
Setting	Described as outpatient	Reviewers must decide whether various settings are in fact outpatient, such as residential treatment programs.	Patients in residential treatment programs may be patients with more severe symptoms or other comorbidities in which the intervention may be more or less effective.
Study Designs or Study Characteristics	Randomization or allocation of treatment (RCT vs. observational studies)	Reviewer decides to include RCTs only.	Limitation to RCTs may be more likely to exclude certain types of interventions such as procedures or dietary/nutritional interventions, as well as studies reporting long term outcomes or harms.
	Quality or risk of bias of individual	Reviewer decides to exclude low quality studies or those at high risk	Studies conducted in nonacademic centers or with a null effect may be more likely to rate as “low quality” due to rejection from high impact journals. Exclusion of all low quality studies or large body of consistent studies that may yield valuable information on benefits or harms.
	Study size	Reviewer decides to exclude RCTs less than 50 participants or observational studies less than 1000 patients.	Exclusion of small studies may exclude valuable information. Exclusion of small studies introduce bias such as by excluding studies conducted in nonacademic or urban populations which may have higher severity of disease, and overestimate effectiveness.
	English language	Reviewer decides to exclude non-English studies.	Exclusion of non-English studies may exclude studies that found a null effect and thus overestimate effectiveness.
	Inclusion of necessary information	Reviewer may exclude studies that do not report the primary outcomes listed.	Studies may have measured outcomes, but not reported them in the studies due to null findings. Exclusion of these studies may overestimate effectiveness.

Abbreviations: PICOTS = population(s), intervention(s), comparator(s), outcome(s), timing, and setting; RCT = randomized controlled trial

Population

Inclusion criteria for the population(s) of interest should be defined in terms of relevant demographic variables, disease variables (i.e., variations in diagnostic criteria, disease stage, type, or severity), risk factors for disease, cointerventions, and coexisting conditions.[18] For example, if an SR is focusing only on adult populations, then the inclusion criteria should specify the age range of interest. Ambiguity in population inclusion criteria increases the risk that inclusion decisions could be influenced by differing viewpoints about potential relationships between particular demographic or disease factors and outcome. Table 2 illustrates one such example of how inadequate description of inclusion criteria for a heart failure population may bias the results of SR. Inclusion criteria for population subgroups of interest should also be defined with similar specificity.

Intervention and Comparators

Although the Key Questions may frame the interventions in broad terms such as "anticoagulants," it is essential for the inclusion criteria to specify exactly which individual interventions are of interest, including their duration and intensity. Otherwise, reviewers may end up missing important interventions and thus overestimate or underestimate the effectiveness or harms of an intervention. This is particularly important in reviews of health care delivery programs that are less clearly defined. A review may examine a specific program as a whole, the component parts of a program, or the theoretical mechanism of action of a component part. Defining an intervention too narrowly may increase the confidence in effectiveness, but reduce the relevance of the finding for implementation in other settings.

To enhance readability, key questions may not always define the comparison, which may introduce both random and systematic error. Without specifying the comparator, one reviewer may compare the effectiveness of anticoagulants to compression stockings, another may compare them to early walking, and yet another may compare it to other anticoagulants. Selection of a comparison of known poor effectiveness may systematically bias the effectiveness of the intervention away from the null, whereas poor specification and thus inappropriate combination of comparisons may result in an uninterpretable result.

Outcomes

Regardless of the topic, SRs should focus on assessing a range of patient-centered outcomes, including both benefits and harms. The scope of included outcomes should address both effectiveness and harms on which strength of the evidence will be graded.[19] If intermediate outcomes are included they should be presented in context of how they relate to the clinically important harms and benefits (e.g., via an analytic framework) as outlined in the chapter of grading the strength of the evidence.[19] When there are a large number of outcomes included, EPCs should specify a priori which clinically important outcomes they will grade the strength of evidence. Despite the temptation to exclude studies that only report a specific outcome (e.g., mortality), EPCs should be cautious since this may augment the risk of identifying studies that have selectively published only outcomes with positive results (selective outcome reporting bias).

In order to reduce variation in study selection related to outcomes, we recommend that the inclusion criteria clearly identify and describe outcomes, outline any restrictions on

measurement methods or timing of outcome measurement, and provide guidance for handling of composite outcomes. For clinical areas (such has pain and psychological functioning) that are notoriously characterized by variability in outcome measurement methods and a multitude of scales and instruments, the risk is greater for inconsistency in study selection. In these cases, it is especially important to consider how to handle this variation early in the SR process. The EPC may choose to restrict to specific measurement methods (i.e., only including studies that used measurement scales that have been published or validated), but need to consider what studies they will be eliminating and what effect this may have on the review. Study investigators that do not use the most commonly validated instruments may be systematically different from those that do. For example, investigators from different communities may use different instruments and systematic exclusion of these studies may exclude specific populations such as rural or small communities or nonacademic populations.

Lack of specificity on other aspects of outcome measurement may also bias SR conclusions. For example where study reports include multiple time points for outcome measurement, but the SR inclusion criteria are not adequately specific about the relative importance of different time points, the choice of which to include or to emphasize is left to the reviewers. This scenario could lead to important differences in conclusions depending on which outcome-time point pair are selected for inclusion, particularly in a meta-analysis.[10]

Finally, it is ideal to consider individual outcome separately, rather than using composite outcomes. Composite endpoints are often difficult to interpret and may exaggerate the magnitude of treatment effect.[20] EPC reviewers should consider specifying whether composite outcomes are of interest and, if so, whether there is a need to place any restrictions on which combinations of outcomes are acceptable (e.g. those with similar importance to patients and magnitude of treatment effect). Otherwise, there may be variation in selection of studies that, for example, do not separately report mortality and cardiovascular events. EPC review teams should rely on empiric research when available to form the basis of any decisions to limit study selection based on outcomes.

Timeframe and Setting

Setting inclusion criteria for timeframe (duration of study, years of study conduct, etc.) and setting may not apply to all clinical questions. Reviewers should identify the expected time period of study that would be needed to identify effectiveness on patient-important outcomes and harms. Lack of specification for the need for long-term studies may overestimate the effect on short term outcomes, while under-reporting the effect on long term outcomes. EPCs should clearly specify any decision to limit studies based on followup duration and define a priori the most relevant time periods for the interventions, populations, and outcomes of interest. When the focus of a SR is confined to a particular setting, such as a nursing home environment or residential treatment center, the inclusion criteria should include guidance for considering eligibility of studies that include commingled or ill-defined settings. Reviewers should consider how interventions may be different in settings such as nursing homes or other long-term care settings compared with general inpatient or outpatient settings and how inclusion or exclusion of these settings may systematically bias the conclusions. The criterion for study setting may also be considered when setting the selection criteria for population.

Study Designs or Study Characteristics

Due to time, budget, or resource constraints as well as concerns about the validity and relevance of the studies, reviewers often make decisions about excluding studies based on study design features (randomization or nonallocation of treatment), study conduct (quality or risk of bias of individual study), language of publication, study size, or reporting of relevant data.

Observational studies make up the bulk of the published literature. EPCs should refer to the Methods guidance for when to include observational studies.[21,22] However after deciding to include observational studies, EPCs need to take special care in developing and testing criteria for determining eligibility.[4] Because of the lack of consensus on any single taxonomy for assigning labels to specific types of observational study designs,[23] EPC teams should define study designs with sufficient clarity so that their reviewers can consistently and correctly determine if a given study is eligible. Exclusion of observational studies without careful consideration about whether these studies may provide information that would not be available from RCTs (i.e., long-term outcomes or harms and representative populations) may bias the review conclusions.

Reviewers often include other study design or reporting characteristics as eligibility criteria. Reviewers may decide to restrict study inclusion based on sample size (e.g.,> 1,000 patients) or publication language (e.g., English language only). However, smaller studies or non-English studies may be systematically different from larger studies or English-language studies and limiting by these characteristics for convenience may introduce a systematic bias as well. For example, in a review of surgical and pharmaceutical interventions, studies on surgical interventions may be smaller than studies on pharmaceuticals, thus biasing a review that excludes small studies to find evidence on drugs but insufficient evidence on surgical interventions.

Typically such decisions are taken for reasons of time-efficiency. The assumption is that not employing such limits would yield a very large number of studies that would significantly increase workload without providing additional value in terms of high-quality evidence. Without empirical evidence relative to the topic area under review, it is not possible to rule out systematic bias. For example, the decision to use only English-language publications may be set because the review team does not have the ability to read other languages but the time and cost of translation are not feasible within the report timeline and budget. Studies of language restrictions in SRs have had variable results, from significant impact to very little impact, sometimes depending on the specific topic being studied.[24-34]

The way that high risk of bias studies are handled in SRs also varies and may introduce bias. Once a study has been determined to have high risk of bias, options include outright exclusion; inclusion in evidence tables with or without inclusion in a narrative description of the evidence (possibly depending on whether the study constitutes the only evidence for a given intervention and/or outcome); or inclusion in quantitative analyses using weighting based on quality or sensitivity analysis. Including studies with a high risk of bias without appropriate weighting for their risk of bias may introduce bias in the SR. However, because assessments of risk of bias are never based entirely on empirical evidence, and are subjective by nature, outright exclusion of studies with high risk of bias may also introduce bias. Additionally, weighting in meta-analysis based on risk of bias assessments may introduce bias and has been shown to result in inconsistency.[35] EPCs should be explicit about how such studies will be handled, a priori. If studies with high risk of bias are to be excluded in any way, they should be clearly identified in

the text or in an appendix. Such transparency improves the likelihood that erroneous ratings of studies with high risk of bias can be identified.

Study Selection Process

Even with clear, precise inclusion criteria, elements of subjectivity and potential for human error in study selection still exist. For example, inclusion judgments may be influenced by personal knowledge and understanding of the clinical area or study design (or lack thereof).

The study selection process is typically done in two stages; the first stage involves a preliminary assessment of only the titles and abstracts of the search results. The purpose of this step is to eliminate efficiently all obviously ineligible publications. The second stage involves a careful review of the full-text publications.

Dual review—having two reviewers independently assess citations for inclusion—is one method of reducing the risk of biased decisions on study inclusion, as is recommended in the Institute of Medicine's "What works in healthcare: standards for systematic reviews."[36] Some form of dual review should be done at each stage to reduce the potential for random errors and bias. Reviewers compare decisions and resolve differences through discussion, consulting a third party when consensus cannot be reached. The third party should be an experienced senior reviewer. The two stages of assessment are discussed in more detail below. Dual review can help identify misunderstandings of the criteria and resolve them such that the studies included will truly fulfill the intended criteria.

At the title and abstract stage, one alternative to 100 percent dual review is to have one reviewer accept the citations that appear to meet inclusion criteria and send them on to full-text review, with a second reviewer assessing only those citations and abstracts that the first reviewer deemed ineligible. Although there is currently no empiric evidence to support this method, we speculate that the sensitivity of the process is increased although the specificity may be somewhat reduced; the tradeoff is a potentially larger pool of full-text articles to review but a lower chance of having missed an eligible study. Additionally there is a risk of reviewer bias, with the second reviewer's knowledge that the first reviewer had deemed the studies ineligible. A second reasonable alternative is to conduct dual review on a small percentage of the citations, insuring reliability of assessments before going on to have the remainder of citations assessed by a single reviewer. In this situation, we recommend that review teams start with a pilot phase, using screening forms based on the eligibility criteria, to screen a small number of studies (e.g., 10 to 20 percent), followed by discussion such that variation in interpretation of how the inclusion criteria should be applied can be resolved early on. For this calibration process we suggest pairing a methodologist with a clinical expert if possible. For the stage of reviewing of full-text articles we recommend that EPCs undertake a complete independent dual review.

Some experts assert that reviewers' knowledge of the identity of the study authors, institution, or journal, or year of publication may influence their decisions and that masking of these factors might be useful.[37,38] These assertions may be based on the findings of a randomized study conducted by Berlin, et al., where there was considerable disagreement between blinded and unblinded reviewers in selecting studies for meta-analysis in where reviewers were using the same inclusion criteria.[39] However, the conclusions of this study were that masking "during study selection and data extraction had neither a clinically nor a statistically significant effect on the summary odds ratio" and that masking required 1.3 hours per paper. Hence, masking of reviewers to manuscript details is not routinely recommended.

Testing of inter- or intra-rater reliability, using the kappa statistic is sometimes suggested as a necessary component of the dual review strategy. However, because the goal is to include the "right" studies and not necessarily to achieve perfect agreement, and using the usual dual review process should obviate the need for such testing, this approach is not generally recommended.

Documenting and reporting all decisions made in the study selection process at the full-text level provides transparency that is essential in allowing independent assessment of the potential for bias by readers of SRs. SRs should include the numbers of studies screened, assessed for eligibility, and included in the review, ideally in the form of a flow diagram as recommended in the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) statement.[17]

As a part of this transparency, SRs should include a listing of excluded studies, along with respective reasons for exclusion. The list of excluded studies is meant to document the reason that specific studies reviewed at the full-text level were excluded when a reader may reasonably think they might have been included. An example would be studies in which the population and interventions meet eligibility, but the study design or comparator does not.

Using Gray Literature To Assess and Reduce Bias

In reviewing gray literature documents, reviewers are seeking to identify unpublished studies and unpublished data supplemental to published studies. Just as excluding studies can cause systematic variation, different approaches to finding and including or using grey literature can also affect the studies included and thus the conclusions of a review. While there may be variation in definitions of gray literature in general, EPC guidance outlines the best practices for identifying gray literature from regulatory data (e.g., the FDA), manufacturers, and other unpublished information such as abstracts or trial registries (see Table 3 for descriptions).[12] At a minimum, knowledge of unpublished studies may lead the EPC to reduce their assessment of the strength of the body of evidence in the review because of the existence of grey literature may suggest evidence of publication bias.[40] There is a risk that the gray literature identified has a high risk of bias; that the reason for lack of publication was due to flaws in the study rather than negative results. In some cases, enough information may be available for the reviewers to assess study quality and include the study in the SR.

A review of original protocols (i.e., registered with clinicaltrials.gov) may identify selective reporting in the published literature for outcomes in which there is a positive result. Comprehensive searches for protocols and identification of selective outcome reporting may lead a reviewer to reduce their confidence in a positive finding. EPC reviewers should be alert to the possibility that the study measured and analyzed the outcome of interest, but did not report the finding due to a negative result. Gray literature helps to provide some fuzzy information on areas that were previously a blind spot in SRs of only published literature.

Reviewing gray literature may be resource intensive, and it is not yet clear if or when the effort required is worth the potential benefit. Despite these limitations, the risk for selective and biased publication of studies makes the inclusion of gray literature a necessary component of high quality SRs until empirical evidence is available to provide further guidance. Given the complexity of gray literature and the likelihood that a given review may not be able to fully search and include all gray literature, we recommend that the review protocol define, a priori, the sources of gray literature (Table 3), and the eligibility criteria applied to them. The following are

our recommendations for how to approach selecting studies from gray literature documents in a way that will minimize potential bias in selection of studies:

1. Identify studies for the SR using standard search techniques first and become familiar with these studies before reviewing gray literature documents.
2. Assess studies in gray literature documents for eligibility in the SR using the key questions and inclusion criteria as discussed above.
3. As some sources of gray literature will have overlap with published literature, for example, FDA documents and trial registries, reviewers should match studies in gray literature documents based on characteristics such as unique study identifies, sample size (by group), and study duration, to those found in published literature to remove any duplicates. This information is sometimes readily available, but often matching is difficult.
4. As with assessment of other types of evidence, dual review is a good way to guard against potentially biased inclusion decisions. Reporting on the inclusion of unpublished studies or data is important to ensure transparency and to identify areas about which EPCs have less confidence that the reporting is unbiased because the included information had not been published and, therefore, had not yet been vetted through a peer review process.
5. If gray literature search uncovers studies that were not included in the published literature, EPC must consider whether the studies have sufficient data and are of sufficient quality to be included in the analysis. If not, then consider whether the presence of such studies suggests that the published literature is biased and should be "downgraded" for publication bias in assessing the strength of evidence.

Table 3. Sources of unpublished information for comparative effectiveness reviews

Source	Description
FDA Documents	Documents from the FDA are the reports written by FDA professional staff assigned to review a New Drug Application submitted by a pharmaceutical manufacturer when applying for FDA approval of a drug for a specific indication or set of related indications. Although FDA review documents have multiple parts, the two most relevant sections for the EPC review team are the medical reviewers' and statistical reviewers' reports. By reviewing these sections, the EPC may identify studies that they did not find through their published literature search and that may indicate the presence of publication or outcome reporting bias. Many of the FDA documents currently available are only scanned originals, meaning that EPCs cannot use software search functions on them; moreover, in some sections, the FDA may have redacted some material; finally, in addition to potentially relevant trials, these documents may also include studies that are not relevant to a SR (e.g., studies in healthy subjects). Nonetheless, they can provide data and analyses of Phase 2 and 3 trials that may be more extensive than are available in published manuscripts.
Scientific Information Packets	Through the SIPs,[12] manufacturers may submit published and unpublished data from RCTs and observational studies relevant to clinical outcomes. For unpublished studies, manufacturers are asked to provide a summary that includes study number, study period, design, methodology, indication and diagnosis, drug dose and duration, inclusion and exclusion criteria, primary and secondary outcomes, baseline characteristics, numbers of patients screened/eligible/enrolled/lost to withdrawn/follow-up/analyzed, and effectiveness/efficacy and safety results. For studies registered with ClinicalTrials.gov, the ClinicalTrials.gov identifier, condition, and intervention are also requested.
Trial Registries	Trial registries that contain results from trials registered, such as the ClinicalTrials.gov and Clinicalstudyresults.org, can be useful sources of information for reviewers. Because the study is registered at the beginning of the study, the intended primary outcome measures, sample size, and other trial characteristics are known prior to reading reports of results. While this can be very useful in identifying potential outcome reporting biases, these registries are also useful in identifying completed studies that have not yet been published, and data on outcomes that may not have been reported in the publications of the trial.

Abbreviations: EPC = Evidence-based Practice Center; FDA = U.S. Food and Drug Administration; RCT = randomized controlled trial; SIP = scientific information packet; SR = systematic review

Because the studies in the FDA documents and trial registries are referred to by codes and because the publications of these studies may or may not also list these numbers, EPCs must often match up the studies using study characteristics (e.g., numbers of included patients, duration of study). Doing so allows reviewers to identify relevant unpublished studies or additional outcomes or and statistical analyses examined in a known study that had not been not reported in the published literature. This process, although lengthy, can help EPCs identify the full body of evidence that is relevant to the question and better identify or reduce bias in selection of studies. Comparing these documents to published manuscripts of the trials may also uncover changes in the definition the primary outcome or misrepresentation of the primary outcome.[41] Dual review of gray literature documents is recommended when assessing relevance for potential inclusion into the review.

EPCs may determine that unpublished, supplemental data from the documents in the scientific information packets (SIPs) pertaining to studies with publications may be appropriate for inclusion into their review. For example, subgroup analyses may be reported in SIPS that had not appeared in the published manuscript(s); however, EPCs do need to view these data with caution. EPC reviewers should have discussed and established a priori guidance on when to include specific types of unpublished data and how to handle such data when they are included. With respect to subgroup data or analyses, for example, the review team should define the clinically relevant subgroup populations (e.g., characterized by comorbidities and drug co-administration) during topic development and document them a priori in the inclusion criteria document. If SIPs present data on populations other than those identified as clinically relevant, then EPCs would not include them or include them only as hypothesis generating; alternatively,EPCs may consider formally amending the inclusion criteria if clinical expertise indicates that noninclusion of these subgroups was an oversight.

Discussion

Our review of the literature indicates that systematic bias and random error can potentially occur in the selection of studies for SRs. Methods exist to reduce the likelihood of both problems, as described in this chapter. Some aspects of potential bias in study selection overlap with considerations to reduce bias when defining the key questions (discussed in further detail by Whitlock, et al.[11]). Table 2 highlights some potential sources of bias that reviewers should consider when selecting inclusion and exclusion criteria. However these are only potential sources of bias and need further research to establish which may be more likely to introduce systematic bias into a review. Further, as this is likely topic specific, reviewers need to have a careful and considered approach in selecting inclusion and exclusion criteria. After selection of inclusion and exclusion criteria, reviewers should track the reasons for exclusions of studies and consider at the end whether exclusion of studies due to the reasons identified in Table 2 may have biased the study. The potential effect of excluding or combining studies on the results should be highlighted as a potential limitation in the Discussion section of the SR.

A potential source of bias that was not addressed in this paper is the assessment and management of conflict of interest for authors, funders, and others with input into the SR process, including technical experts, key informants, and peer reviewers. The possible impact of conflicts is unknown at this time, but is the subject of future research, and is addressed in the Institute of Medicine's Standards for Systematic Reviews.[15] EPCs must be aware of not only the possibility of outcome reporting bias of individual studies, but also their own presentation of

outcomes and how that may be introduce bias into the interpretation of findings. While some of these issues have been touched on in this paper, they are the subject of future research as well.

EPC reviewers should explicitly consider how they handle the concept of "best evidence" in both inclusion and synthesis of studies. Even when studies technically meet all eligibility criteria, and are correctly identified for inclusion using rigorous assessment procedures, the level of contribution each eligible study will make to the body of evidence can vary importantly. Depending on the availability of the best possible evidence, EPCs may differ in the extent to which they use lower-strength evidence for a given SR.

For example, when the evidence from randomized controlled trials that directly compare interventions has no obvious gaps, then the value of lower-strength evidence from observational studies, indirect comparisons from placebo-controlled trials, and pooled analyses of only a select number of studies is lower than it would be if the EPC reviewers did encounter such gaps. Thus, when gaps exist in the best possible evidence, the value of lower-strength evidence is greater. Reviewers must rely on their expert judgment as to what constitutes a gap in the best possible evidence and to what extent to report the lower-strength evidence. Systematic bias or random error can occur when EPCs do not clearly establish decision rules for utilizing lower-strength evidence.[22]

Conclusion

In summary, EPCs should write the key questions and inclusion criteria in a way that provides their reviewers with detail sufficient to minimize variation in interpretation. Discussion, dual review, and practice will aid in reducing potential bias by establishing consistent interpretation of the criteria. EPCs should disclose the studies evaluated at the full-text level and determined to be ineligible and provide brief reasons for those exclusions.

Reporting the steps taken to avoid bias in selecting studies, such as conducting dual review, tracing the resulting flow of studies through the review (e.g., PRISMA diagram), and reporting potentially relevant studies that were excluded (with reasons for their exclusion) in the SR is essential for transparency. Gray literature can provide evidence on publication bias and outcomes reporting bias; EPCs should use processes similar to those used with published literature in reviewing gray literature to avoid potential bias in selecting unpublished studies or data. Depending on the experience levels of the SR team members, the complexity of the clinical area, the size of the SR, and other factors, the exact approach to operationalizing the study selection process may vary somewhat from SR to SR. Below are some summary points to minimize various types of study selection bias.

- Define inclusion and exclusion criteria by PICOTS clearly and in a protocol. Reduce ambiguity as much as possible.
- Consider the risk of introducing spectrum bias when selecting populations.
- Define interventions with specificity such that they are applicable to the intended user of the review.
- Be cautious about excluding studies based on reporting of outcomes of interest.
- Dual review can help reduce random error in applying inclusion and exclusion criteria

Examine grey literature for evidence of unpublished data or studies that may indicate the presence of publication bias or selective outcome reporting bias. Consider the risk of bias of this information before using the information in the review or to adjust the strength of evidence of the review.

Author Affiliations

Oregon Health and Science University Evidence-based Practice Center (MM, KP). McMaster University Evidence-based Practice Center (PR). Agency for Healthcare Research and Quality (SC). Southern California Evidence-based Practice Center (PS).

References

1. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2011. www.effectivehealthcare. ahrq.gov/ehc/products/60/318/MethodsGuide_Prepublica tion-Draft_20120523.pdf.
2. MacLean C, Alexander A, Carter J, et al. Comparative Effectiveness of Treatments To Prevent Fractures in Men and Women With Low Bone Density or Osteoporosis. Comparative Effectiveness Review No. 12. (Prepared by Southern California/RAND Evidence-based Practice Center under Contract No. 290-02-0003). Rockville, MD: Agency for Healthcare Research and Quality; 2007. www.effective healthcare.ahrq.gov/reports/ final.cfm.
3. Sawka A, Papaioannou A, Adachi J. Does alendronate reduce the risk of fracture in men? A meta-analysis incorporating prior knowledge of anti-fracture efficacy in women. BMC Musculoskelet Disord. 2005;6:39. PMID: 16008835.
4. Cranney A, Wells G, Willan A, et al. Meta-analyses of therapies for postmenopausal osteoporosis. II. Meta-analysis of alendronate for the treatment of postmenopausal women. Endocr Rev. 2002;23(4):508-16. PMID: 12202465.
5. Cook DJ, Reeve BK, Guyatt GH, et al. Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta- analyses. JAMA. 1996 Jan 24-31;275(4):308-14. PMID: 8544272.
6. Hopayian K, Mugford M. Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed? Br J Gen Pract. 1999 Jan;49(438):57- 61. PMID: 10622020.
7. Peinemann F, McGauran N, Sauerland S, et al. Disagreement in primary study selection between systematic reviews on negative pressure wound therapy. BMC Med Res Methodol. 2008 Jun 26;8(1):41. PMID: 18582373.
8. Thompson R, Bandera E, Burley V, et al. Reproducibility of systematic literature reviews on food, nutrition, physical activity and endometrial cancer. Public Health Nutr. 2008 Oct;11(10):1006-4. PMID: 18053295.
9. Ioannidis JPA. Meta-research: The art of getting it wrong. Research Synthesis Methods. 2011;1:169-84.
10. Tendal B, Higgins JP, Juni P, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. BMJ. 2009;339:b3128. PMID: 19679616.
11. Whitlock EP, Lopez SA, Chang S, et al. AHRQ series paper 3: identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):491-501. PMID: 19540721.
12. Relevo R, Balshem H. Finding Evidence for Comparing Medical Interventions. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11- EHC021-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2011. http://effectivehealthcare.ahrq.gov.
13. Norris S, Holmer H, Ogden L, et al. Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness. (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-2007-10057-I.) AHRQ Publication No. 12-EHC110-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012. www.effectivehealthcare.ahrq.gov/reports/ final.cfm.
14. Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 12-EHC047-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012. www.effectivehealthcare.ahrq.gov/ehc/products/322/998/MethodsGuideforCERs_Viswanathan_IndividualStudies.pdf.

15. Institute of Medicine. Finding What Works In Health Care: Standards for Systematic Reviews. Washington, DC: National Academies Press; 2011.
16. Clarke M, Stewart L. PROSPERO–the new international prospective register of systematic reviews. Cochrane Methods. Cochrane Database of Systematic Reviews 2011;Suppl 1:1-40.
17. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336-41. PMID: 20171303.
18. West S, Gartlehner G, Mansfield AJ, et al. Comparative Effectiveness Review Methods: Clinical Heterogeneity. Methods Research Paper. AHRQ Publication No. 10- EHC070-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2010. http://effectivehealthcare.ahrq.gov.
19. Owens DK, Lohr KN, Atkins D, et al. Chapter 10. Grading the Strength of a Body of Evidence When Comparing Medical Intervention. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2011. Chapters available at: www.effectivehealthcare.ahrq.gov
20. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol. 2007 Jul;60(7):651-7; discussion 8-62. PMID: 17573977.
21. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):502-12. PMID: 18823754.
22. Norris S, Atkins D, Bruening W, et al. Selecting Observational Studies for Comparing Medical Interventions. In: Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2010. www.effectivehealthcare.ahrq. gov/ehc/products/196/454/MethodsGuideNorris_ 06 042010.pdf.
23. Hartling L, Bond K, Harvey K, et al. Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures. Methods Research Report. Rockville, MD: Agency for Healthcare Research and Quality; 2010. AHRQ Publication No. 11-EHC-007. http://effectivehealthcare.ahrq.gov.
24. Nylenna M, Riis P, Karlsson Y. Multiple blinded reviews of the same two manuscripts. Effects of referee characteristics and publication language. JAMA. 1994 Jul 13;272(2):149-51. PMID: 8015129.
25. Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol. 1995 Jan;48(1):159-63. PMID: 7853041.
26. Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet. 1996 Feb 10;347(8998):363-6. PMID: 8598702.
27. Egger M, Zellweger-Zahner T, Schneider M, et al. Language bias in randomised controlled trials published in English and German. Lancet. 1997 Aug 2;350(9074):326-9. PMID: 9251637.
28. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta- analyses? J Clin Epidemiol. 2000 Sep;53(9):964-72. PMID: 11004423.
29. Juni P, Holenstein F, Sterne J, et al. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002 Feb;31(1):115- 23. PMID: 11914306.
30. Moher D, Pham B, Lawson ML, et al. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technol Assess. 2003;7(41):1-106. PMID: 14670218.
31. Pham B, Klassen TP, Lawson ML, et al. Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. J Clin Epidemiol. 2005 Aug;58(8):769-76. PMID: 16086467.

32. Pilkington K, Boshnakova A, Clarke M, et al. "No language restrictions" in database searches: what does this really mean? J Altern Complement Med. 2005 Feb;11(1):205-7. PMID: 15750383.
33. Baussano I, Brzoska P, Fedeli U, et al. Does language matter? A case study of epidemiological and public health journals, databases and professional education in French, German and Italian. Emerg Themes Epidemiol. 2008;5:16. PMID: 18826570.
34. Morrison A, Moulton K, Clark M, et al. English-language restriction when conducting systematic review-based meta- analyses: systematic review of published studies. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2009. p. 1-17.
35. Greenland S, O'Rourk K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001;2:463-71. PMID: 12933636.
36. Institute of Medicine. Knowing what works in health care: A roadmap for the nation. In: Eden J, Wheatley B, McNeil B, et al., eds. Washington, DC: National Academies Press; 2008.
37. Systematic reviews: CRD's guidance for undertaking reviews in health care. York: Centre for Reviews and Dissemination, University of York, UK; 2009.
38. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration. 2009.www.cochrane- handbook.org.
39. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet. 1997 Jul 19;350(9072):185-6. PMID: 9250191.
40. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol. May;63(5):513-23. PMID: 19595577.
41. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008 Jan 17;358(3):252-60. PMID: 18199864.

Chapter 14. Handling Continuous Outcomes in Quantitative Synthesis

Rongwei Fu, Benjamin W. Vandermeer, Tatyana A. Shamliyan, Maya E. O'Neil, Fatemeh Yazdi, Steven H. Fox, Sally C. Morton

Introduction

In quantitative synthesis of randomized clinical trials (RCTs) for a comparative effectiveness review, continuous outcomes are usually less straightforward to analyze than binary outcomes. Continuous outcomes are often measured at both baseline and followup time points. Results of continuous data can be reported as means, mean differences, or differences in change score from baseline, and measures of precision are reported as standard deviation (SD), standard error (SE), or confidence intervals. The distribution of the data is not always symmetric, and journal publications may not report all of the information required for meta-analysis.

The original quantitative synthesis chapter of the "Methods Guide for Effectiveness and Comparative Effectiveness Reviews" has a very brief continuous outcomes section that provides limited guidance on using mean difference versus standardized mean difference, but the section does not provide guidance on a number of other issues relating to meta-analysis of continuous outcomes. To fill this gap, this report updates the guidance on quantitative synthesis of continuous outcomes measured in RCTs.

Accordingly, we address the following topics applicable to quantitative synthesis of continuous outcomes measured in RCTs: choice of effect measures of continuous outcomes, choice of estimates for mean difference and baseline imbalance; calculation of SD and SE, how to handle missing data and skewed data, use and interpretation of the standardized mean difference (SMD) and of the ratio of means (RoM) as an alternative measure, and dichotomization of continuous outcomes in meta-analyses.

For each of the topics related to quantitative synthesis of continuous outcomes, we searched for relevant methodological or applied methodological papers in the Effective Health Care Program Methods Library and in Ovid Medline, Current Index to Statistics, and Scopus databases (Appendix A). Recommendations for each topic were then developed based on current knowledge of the literature along with group discussion and consensus. A draft report of the workgroup's key conclusions and recommendations was circulated for comment to peer reviewers and Agency for Healthcare Research and Quality officers, and those comments were considered by the team in preparing this report. The summary of final key points and recommendations are presented in Table 2 at the end of this chapter.

Effect Measures for Continuous Outcomes

The two effect measures most often used for continuous outcomes are mean difference and standardized mean difference (SMD). The choice of effect measure is determined primarily by the scale of the available data: Investigators can combine mean differences if multiple trials report results using the same or similar scales, but SMD is typically used when the outcome is measured using different scales. RoM,[1,2] a recently proposed measure, is an alternative to SMD for outcomes measured using different scales and allows evaluation of the percentage change of a

continuous outcome. This section and the next focus on different estimates of mean difference and choice of estimates for mean difference related to baseline imbalance. SMD and RoM are discussed in detail in subsequent sections.

There are several ways to calculate mean difference for continuous outcomes measured at both baseline and followup in randomized clinical trials:

1. Use the followup score only to calculate a mean difference between intervention groups.
2. Calculate the mean change score from baseline to followup for each intervention group and use the difference in the mean change scores between the intervention groups as the effect measure.
3. Use the followup score as the dependent variable in an analysis of covariance (ANCOVA) model to estimate the difference between the intervention groups as the effect measure.
4. Use the change score from baseline to followup as the dependent variable in an ANCOVA model to estimate the difference between the intervention groups as the effect measure.

In both options 3 and 4, the variable for the intervention groups is an independent variable in the ANCOVA model, and the baseline score enters the model as a covariate. The coefficient for the variable of the intervention groups provides the estimate for the effect measure, that is, the difference between the two intervention groups. Options 3 and 4 are equivalent statistically in terms of estimating the effect measure. When the variance of the baseline score equals the variance of the followup score, an ANCOVA estimate is the weighted sum of the two estimates from options 1 and 2, and the weight is the correlation between baseline and the followup score.[3] If the correlation is greater than 0.5, the difference in change in score from option 2 has more weight; otherwise, the difference between followup scores has more weight. Note that the correlation between baseline and the followup score is generally positive.

It is possible that the observed variance at baseline is very different from the variance of the followup score, and an ANCOVA estimate is not exactly a weighted sum of the two measures; however, the ANCOVA estimate usually lies between the estimates from options 1 and 2. For example, in a study evaluating glycemic control in patients with type 2 diabetes,[4] patients randomized to the metformin group have a mean level of hemoglobin A1c of 6.79 percent, and the mean level for the patients randomized to the metformin plus glimepiride group is 6.42 percent. After 20 weeks, the mean level of hemoglobin A1c is 6.86 percent in the metformin group, and 5.68 percent in the metformin plus glimepiride group. For the mean difference between the two groups, options 1 and 2 provide an estimate of 1.18 percent and 0.81 percent, respectively; the ANCOVA estimate is 0.92 percent, located between the above two estimates. The correlation between baseline and the followup score is about 0.6.

Choice of Estimate for Mean Difference and Baseline Imbalance

For an adequately randomized RCT, on average, distribution of baseline characteristics should be similar among intervention groups. However, baseline imbalance often occurs for one or more characteristics. This imbalance could be due to chance, especially in small trials,[5] or due to selection bias, often caused by inadequate randomization concealment.[6]

Assessment of Baseline Balance

Should Investigators Assess Baseline Balance of Included Trials in Quantitative Synthesis?

In the process of quality rating, the balance of baseline scores is one of the factors usually assessed to check the adequacy of randomization, but little attention has been paid to baseline balance in quantitative synthesis. A meta-analysis may have different results depending on whether we adjust for baseline imbalance.[7] Here we distinguish between two types of baseline variables. The first reflects the usual patient characteristics and important prognostic factors for the medical condition under study, and the second type reflects the baseline measurements of continuous variables that are specified as outcomes. Both types should be incorporated in quality rating, but the second is more relevant in quantitative synthesis. Quality should be downgraded if the balance of important prognostic factors and outcome variables is not achieved and this imbalance is not addressed in the included studies.

For the second type of baseline variables, investigators should also assess the baseline balance for each continuous outcome and take any imbalance into consideration when conducting quantitative synthesis.

How To Assess Whether the Baseline Scores Are Balanced

Though alternative opinion exists,[8] for both types of baseline variables the use of statistical testing for baseline difference is generally not recommended for individual studies.[9-14] Some argue that such statistical testing "is a test of a null hypothesis that is known to be true,"[14] and that it "assesses the probability of something having occurred by chance when we know that it did occur by chance."[12] Even if the statistical tests are not significant, imbalance of important prognostic factors could affect results, and the unadjusted estimates could be biased.

Current practices of using statistical testing for baseline difference vary. In a study of published RCTs in leading medical journals, unadjusted estimates of treatment effects were reported more frequently than adjusted estimates.[15] Of the 110 included RCTs, 42 used statistical testing to compare baseline differences. In a systematic review, investigators should base assessments of the baseline distribution on the potential clinical importance of the actual differences between groups and the direction of the imbalance, not on the p-values of tests. An imbalance that favors the control group may have less serious consequences than an imbalance favoring the treatment group. When the decision is not clear cut, we recommend that the investigators take a conservative approach and consider the baseline scores to be imbalanced.

If the baseline scores of the continuous variables specified as outcomes are not reported, investigators should not assume they are comparable even if they consider reported baseline patient characteristics and important prognostic factors to be comparable. If possible, investigators should also consider how attrition may impact imbalances in continuous outcome variables for the subsample with outcome data. For trials with high attrition, the baseline balance may not be maintained in the subsample with outcome data.[16] If baseline scores are not reported with sufficient detail to judge whether they are comparable, the investigators should not assume that they are comparable, and this should be appropriately accounted for in quality rating.

If the baseline score imbalance is only by chance, meta-analysis of baseline score differences between treatment groups of included studies should provide a combined estimate close to zero (given no publication bias).[7] Investigators are encouraged to do such an analysis.

Choice of Estimate for Mean Difference

When the baseline scores are balanced, options 1, 2, or 3 would provide unbiased estimates of mean difference. The ANCOVA approach (option 3) provides a more efficient estimator with more precision.[10,17,18] When the baseline scores are imbalanced, options 1 and 2 produce biased effect estimates of mean difference—option 1 simply ignores baseline imbalance, and option 2, contrary to common belief, does not control for the baseline imbalance. The change score is negatively associated with the baseline score and patients with a worse baseline score are more likely to experience a high change score (regression to the mean). For instance, suppose that a trial has an intervention and a placebo group and the intervention group has a worse baseline score. The treatment effect size from the intervention will be underestimated using option 1 and overestimated using option 2.[19] When baseline imbalance occurs by chance, the ANCOVA has been shown to be a better method to control for this imbalance, and the estimates from ANCOVA are less biased. When baseline scores are correlated to followup scores, adjusting for baseline using ANCOVA has been shown to remove conditional bias in treatment group comparisons due to chance imbalances[11] and to improve efficiency over unadjusted comparisons.[11,18]

Choice of Estimate for Mean Difference When There is No or Only Minimal Baseline Imbalance

Estimates from options 1, 2, or 3 could be combined in one single meta-analysis to obtain a combined mean difference. When there is little or no baseline imbalance, we recommend the following for the choice of estimates for mean difference:

1. If reported, use an ANCOVA estimate—it is an unbiased and more efficient estimator. When a study does not report ANCOVA estimates, it is possible to calculate them if the studies report: (1) means and SDs at baseline and followup for both intervention and control groups, (2) means and SDs of change for both intervention and control groups, and (3) sample size of both intervention and control groups. However, we recognize that studies rarely report such detailed data and calculating ANCOVA estimates is not usually a practical option.
2. If an ANCOVA estimate is not reported and the study directly reported the mean difference or reported enough data to calculate mean difference based on both options 1 and 2, use the estimate with the smaller SE. Option 2, difference in change score, produces a small SE when correlation between baseline and post treatment is high (> 0.5 when variance is equal at baseline and post intervention). Otherwise, option 1, difference between post scores, produces a small SE. There is evidence to show that the correlation between baseline and post score is often greater than 0.5.[20] This correlation is often not reported, and Section "Dealing with Missing Data" provides more information on handling the missing correlation.
3. If the study reported neither the mean difference nor enough data to calculate the mean difference based on both options 1 and 2, use either the reported estimate or whichever estimate can be calculated from the reported data. Sometimes data needed to include the study in the meta-analysis are missing from the report but can be calculated or imputed from the reported data. For more guidance on handling such situations, see the sections "Calculating Standard Deviation and Standard Error When They Are Not Directly Reported" and "Dealing With Missing Data," below.

4. Since all options provide unbiased estimates, it is appropriate for investigators to use the same estimate across trials. In practice, this advice is limited to options 1 and 2, since ANCOVA estimates are usually not reported consistently. In such cases, some assumptions about missing data are usually needed to obtain an estimate of the same effect measure for all trials. For example, when the change score between baseline and followup needs to be calculated, the correlation between baseline and the followup score is often not known and an assumption about the correlation is needed in order to calculate the SE of change score. For more information about handling such situations, see "Calculating Standard Deviation and Standard Error When They Are Not Directly Reported" and "Dealing With Missing Data," below.

Choice of Estimate for Mean Difference When There is Baseline Imbalance

When there is baseline imbalance, ANCOVA estimates are preferred over other options as they provide the least biased estimate with more precision. Options 1 and 2 would provide biased estimates. However, trials that are otherwise appropriate for inclusion but lack ANCOVA estimates should not be excluded from the quantitative synthesis, since they still provide valuable information about the study effect. For the choice of estimates for mean difference for each study, we recommend:

1. Use ANCOVA estimates if reported (more precision and less bias).
2. If ANCOVA estimates are not reported, conduct analyses using both estimates from options 1 and 2 and report the more conservative combined estimate, usually the one with a smaller absolute effect size. Since ANCOVA estimates lie between the estimates from options 1 and 2, the more conservative combined estimate is likely an underestimate compared with the ANCOVA estimate and therefore a better choice for guarding against type I error. If the results from the two estimates do not agree, investigators may also present both combined estimates and clearly explain that the combined estimates are sensitive to the choice of estimate for mean difference. A meta-regression approach[7] has been suggested to adjust for baseline imbalance, though its performance has not been fully studied. Investigators may choose this approach as an additional sensitivity analysis.
3. If enough trials in a meta-analysis report ANCOVA estimates, investigators are encouraged to conduct subgroup analyses to compare results from ANCOVA versus non-ANCOVA estimates as an additional sensitivity analysis.

Calculating Standard Deviation and Standard Error When They Are Not Directly Reported

Commonly used meta-analysis packages (e.g., Review Manager [RevMan], Stata) require three parameters from each of the intervention groups in order to calculate a weighted mean difference: the mean, the SD, and the sample size. The mean could be the mean change score from baseline or the mean score at followup based on the choice of estimate for mean difference. If any of these are missing, the study will be omitted from the meta-analysis. Alternatively, investigators could use the mean difference between the intervention groups and its associated SE directly in meta-analysis.

Frequently, precision parameters such as SD and SE are not reported directly but may be calculated from other reported statistics. Investigators should always look for reported data that could be used to conduct exact algebraic calculation of these parameters. In this section, we

present formulas for calculating SD and SE using other reported statistics. We also briefly discuss the issue of incorporating correlation into calculation of SD for crossover and cluster randomization trials.

Calculation of Standard Deviation and Standard Error Using Available Data

When SD is not directly reported, it can be computed (assuming both mean and sample size are given) from other reported data: SEs, confidence intervals, z- or t-statistics, or exact parametric p-values using available formulas.[21] These other reported data could be available for either the mean between baseline and followup from each intervention group or for the mean difference between two intervention groups.

Available Data for One Intervention Group and the Change Scores

In this section, all calculations apply to obtaining the SD for the change score (i.e., the difference between baseline and followup from any one intervention group) when conducting a meta-analysis using three parameters from each intervention group.

If given an SE of the mean change score of one intervention group in a trial of sample size *n*, the SD for that group can be computed as:

$$SD = SE\sqrt{n}\ ) \qquad (1)$$

If given a 95% normal confidence interval in the form of (lower confidence bound [LCB], upper confidence bound [UCB]) around the mean, we can compute the SE using the formula:

$$SE = \frac{UCB - LCB}{3.92} \qquad (2)$$

Formula (1) can then be used to compute SD. If a 90% confidence interval is given rather than a 95% confidence interval, the divisor in formula (2) should be changed to 3.29. If the 95% confidence interval was based on *t*-distribution, the denominator in the formula must be replaced with the appropriate inverse percentile of the *t*-distribution multiplied by 2. This could easily be done in Microsoft Excel® by typing in any cell "= tinv(0.05,*n*-1)" where *n* is the sample size of the intervention group. If the confidence interval is 90% instead of 95%, replace 0.05 with 0.1.

If given a *z*-statistic or a t-statistic, for the instance of the change score from baseline in each intervention group, the SE can be computed using the change score:

$$SE = \frac{|mean\ change\ score|}{z} \text{ or } SE = \frac{|mean\ change\ score|}{t} \qquad (3)$$

Again, formula (1) can then be used to determine the SD.

If an exact *p*-value is reported for testing whether the followup score is significantly different from baseline in each intervention group, the *p*-value can be converted to a *z*-statistic first, using the inverse normal value. The easiest way to obtain the *z*-statistic is by entering "= normsinv(1- *p*/2)" in any cell, where *p* is the reported *p*-value. For example, if the given *p*-value is 0.03, enter "=normsinv(0.985)", which returns the *z*-statistic of 2.17. If the sample size is small and the study obtained the *p*-value using a paired *t*-test, then the *t*-statistic could be

obtained by entering "=tinv (p,df)", where p is the reported p-value and df is the degree of freedom for the t-test and equals n-1, where n is the sample size of the intervention group. Then formula (3) can be used to calculate SE.

If an upper-bound p-value (e.g., $p<0.05$) is given, then this upper bound can be used with the same formulas to obtain a conservative estimate of the SD.

For calculating SD for the change score, if the SD at baseline (SD_b) and followup (SD_f) are reported, SD for the change score can also be calculated as:

$$SD = \sqrt{SD_b^2 + SD_f^2 - 2 * r * SD_b * SD_f} \qquad (4)$$

where r is the correlation between baseline and the followup score. Information about r is often not available and needs to be imputed. For more information on handling missing data for r, see the section "Dealing with Missing Data."

Available Data for the Mean Difference between Two Groups

If a confidence interval, a z-statistic, or a t-statistic is given for the difference of means between two intervention groups, variations on formulas (2) and (3) can be used to calculate the SE for the mean difference between groups. For formula (3), replace the change score with the mean difference. If an exact p-value for a mean difference is given, it can be converted to a z-statistic using the same Excel "normsinv(1-p/2)" function. If the sample size is small and the study obtained the p-value using a two-sample t-test, then the t-statistic could be obtained by using the Excel function "tinv(p,df)" where p is the reported p-value, but df equals $n_1 + n_2 -2$ in this case, where n_1 and n_2 are the sample size of each intervention group. If an upper-bound p-value (e.g., $p<0.05$) is given, then the same formulas can be used to obtain a conservative estimate of the SE for mean difference.

In some cases, when the SDs for each intervention group (SD_T and SD_C for treatment and control groups, respectively) are reported, SE for the mean difference between intervention and control can be calculated as:

$$SE = \sqrt{\frac{SD_T^2}{n_T} + \frac{SD_C^2}{n_C}}, \qquad (5)$$

where n_T and n_C are the sample sizes of the two intervention groups. If the estimates of SD_T and SD_C are similar, one can also use:

$$SE = \sqrt{\frac{(n_T - 1)SD_T^2 + (n_C - 1)SD_C^2}{n_T + n_C - 2}\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}. \qquad (6)$$

Unlike formula (4), there is no need to consider correlation since the intervention groups are independent in a parallel design.

If the individual standard deviations are not given but the SE of the mean difference is presented, this SE can be used directly in the meta-analysis. While this SE is sufficient to determine the precision of the mean difference, some meta-analysis software packages (e.g., RevMan) require the user to input the individual standard deviations. In this case, the simplifying assumption could be made that treatment SD is equal to the control SD, and this computed SD

can then be used for both intervention and control groups. This assumption will not affect the final result since the precision of the estimate is determined solely by the given SE, and the estimated SD is only used to re-compute this given SE for the specific software package. The common SD can be estimated as:

$$SD = SE\sqrt{\frac{n_T n_C}{n_T + n_C}}. \qquad (7)$$

Direct use of the SE of the difference in means between groups (and the mean difference) in the meta-analysis or computing the SD of each of the trial group will give the same result. Usually the choice of method depends on the type of data reported in the included trials and the meta-analysis package used.

Occasionally trial authors may confuse standard deviation and standard error. The formulas in this section can be used to verify the values if the study has reported confidence intervals or p- values in addition to the SDs or SEs. In a meta-analysis, if one study has an SD that is much smaller than that of all the other trials and has a disproportionally high weight in the meta-analysis, this can be a red flag that an SE was misreported as an SD.

A Worked Example

Suppose that a parallel study with 15 patients in each group reports the following: "The mean systolic blood pressure in the treatment group was 122.4 mmHG while in the control group it was 134.5 mmHG. This difference was not statistically significant (p=0.24)." If this p-value was computed from a z-statistic, how would we compute the SD?

- Mean difference = 134.5 – 122.4 = 12.1.
- 1-p/2 = 1-0.24/2 = 0.88. Entering "=normsinv(0.88)" in an Excel cell gives a z-statistic of
- 1.175. Note: If the t-distribution had been used, then the t-statistic = tinv(0.24, 28) = 1.201 where 28 = 15+15-2.
- SE = 12.1/1.175 = 10.298. This number could be used directly in the meta-analysis, or if one is using a software package that requires the SD in each group, it can be computed from this SE:

$$SD = SE\sqrt{\frac{n_T n_C}{n_T + n_C}} = 10.298\sqrt{\frac{15*15}{15+15}} = 28.2$$

- This SD can be entered for *both* treatment and control groups.

Crossover Trials

For trials with a parallel design, the intervention groups are independent of each other, and there is no need to consider correlation between intervention groups when calculating SE for mean difference. A crossover design is one where the participants, in sequence, receive both the intervention and the control and thus all patients are included in both arms of the trial. When a crossover trial is included in a meta-analysis, in most cases, using the methods of a parallel design to calculate SE for mean difference will give an SE that is too large because the positive correlation associated with using the same patients in both the treatment and control groups lowers the variance of the mean difference. The formula to compute the pooled SE for a crossover trial is:

$$SE_d = \sqrt{SE_T^2 + SE_C^2 - 2rSE_TSE_C} \qquad (8)$$

where r is the within-patient correlation coefficient and SE_d, SE_T, and SE_C are the difference, treatment, and control SEs respectively. For a parallel trial the value of r is always 0, thus the last term becomes 0. For a crossover study, however, the value of r is usually not reported from the trial and needs to be estimated in order to properly compute the correct SE. See Section "Dealing With Missing Data" for methods for calculating or imputing r.

Cluster Randomized Trials

Cluster randomized trials are similar to crossover trials in that formula (5) or (6) will not provide the correct SE for mean difference. Data among patients within a cluster are usually positively correlated. However, unlike in crossover trials, ignoring this correlation in cluster randomized trials will produce an SE of the mean difference between intervention groups that is too small. If a cluster randomized trial reported an SE that failed to account for this correlation, the simplest way to account for this discrepancy is to compute a design effect (DE) as:

$$\mathrm{DE} = 1 + (m-1)\mathrm{ICC} \qquad (9)$$

where m is the average cluster size and ICC is the intra-class correlation coefficient. The ICC is defined as the proportion of the total variance (the within-cluster variance plus the between-cluster variance) that is attributed to the between-cluster variance. The square root of the design effect can then be multiplied by the standard error of the regular mean difference (computed as if it were parallel) to produce the adjusted SE. This new adjusted variance will appropriately reflect the loss of precision due to the cluster randomization design.

A Worked Example

For a cluster randomized trial, suppose that the SE of the mean difference is calculated to be 2.4 using the methods for a parallel design. If the average cluster size was 10 and the ICC was estimated to be 0.03, we can adjust the SE for the design effect as:

$$\mathrm{DE} = 1 + (10-1) * 0.03 = 1.27$$

$$SE_{adj} = \mathrm{sqrt(DE)} * SE = \mathrm{sqrt}(1.27) * 2.4 = 2.7$$

Therefore, 2.7 is the standard error that should be used in the meta-analysis.

The ICC will generally be quite low (less than 0.1) in cluster randomized trials, but it can still have a fairly large effect on the trial variance, particularly when the average cluster size is quite large. Usually this ICC is not reported from the published trials and the investigators need to assume a plausible value to calculate the SE. Investigators should always conduct sensitivity analyses by assuming several values of ICC and checking how robust the results are in comparison with the assumed ICC values. In addition, databases for ICC estimates are available for some outcomes,[22-25] and investigators may refer to the relevant literature to check whether the typical magnitudes of ICC for the type of outcome under study have been reported and make assumptions around the typical estimates.

Dealing With Missing Data

Missing data is a common issue in meta-analysis and often leads to biased estimates. Missing data can take many forms: missing studies, missing outcomes, missing summary data, missing individual, and missing study-level characteristics. Missing studies and missing outcomes are complex issues that are not specific to continuous data and will not be discussed here. This section focuses on the issue of missing summary data, which is most relevant to continuous data in the meta-analysis. The issues of missing individuals and missing study-level data will be discussed briefly.

How Are the Missing Data Distributed?

Missing data can be categorized into one of three types based on missing mechanism: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).[26] Data are said to be MCAR if being missing does not depend on observed or unobserved measurements. MAR means that, given the observed data, the reason data are missing does not depend on *unobserved* data. Data are MNAR if they are neither MCAR nor MAR. Missing data that are MCAR or the more reasonable MAR are considered ignorable in a systematic review. There is no bias in simply performing the meta-analysis without the missing data, and the combined estimate only suffers from less precision.[27] Unfortunately, missing data are usually suspected to be MNAR and must be considered. Simply omitting trials with data that are MNAR will lead to biased results.[26]

Missing Summary Data

If a study is missing data elements that are required in a meta-analysis and these data cannot be calculated from reported data, it is often a good idea to contact the authors to obtain the missing values before conducting the analysis. If it is not possible to obtain the missing values, investigators need to either exclude the study or impute the missing data in some way. Both omitting a study and imputing for missing values can result in bias and under-precision, but it is generally accepted that omitting studies should be avoided when possible.

Standard deviation is the most commonly missing parameter. We recommend that studies missing only SDs should not be excluded, as this could lead to a biased combined estimate. For example, studies with nonsignificant results were more likely to omit standard deviations.

Imputation of Standard Deviation

If the data are not available in an alternative form that allow direct calculation, imputation of missing values is often recommended, based on results from simulation studies.[28] Several simple methods have been suggested for directly imputing missing SDs, including direct substitution using the largest SD of the included studies, arithmetic means,[29] linear regression,[30] coefficient of variation,[31] and imputation from correlation.[28] We demonstrate some of these methods using the following example, taken from a review comparing asthma patients using long-acting beta agonist (LABA) and inhaled corticosteroid (ICS) in combination versus using ICS alone.[32] The outcome is pulmonary function in L/min.

The studies labeled Strand and SAM40036 are missing their SD and are not counted in the final meta-analysis (Figure 1).[32] A direct substitution of the largest SD shows that the largest SD in the LABA/ICS group is 52.14 and in the ICS group is 49.64 (Figure 2).

Figure 1. Results of meta-analysis of pulmonary function without including studies with missing data[32]

Study or Subgroup	LABA/ICS Mean [L/min]	LABA/ICS SD [L/min]	LABA/ICS Total	ICS Mean [L/min]	ICS SD [L/min]	ICS Total	Weight	Mean Difference IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	6.3%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	6.9%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	6.4%	18.90 [5.45, 32.35]	2004
Strand 2004	40	0	78	14	0	72		Not estimable	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	6.1%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	5.7%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	0	288	40.1	0	289		Not estimable	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	11.1%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	22.2%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	16.1%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	19.1%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**19.56 [16.16, 22.95]**	

Heterogeneity: Tau² = 0.00; Chi² = 2.28, df = 8 (P = 0.97); I² = 0%
Test for overall effect: Z = 11.29 (P < 0.00001)

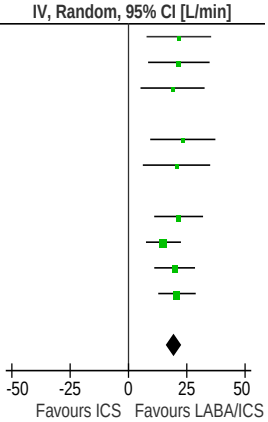
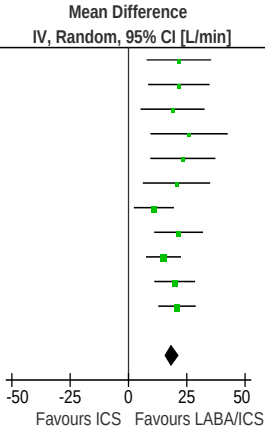


Figure 2. Results of meta-analysis of pulmonary function, imputing missed data using direct substitution.*

Study or Subgroup	LABA/ICS Mean [L/min]	LABA/ICS SD [L/min]	LABA/ICS Total	ICS Mean [L/min]	ICS SD [L/min]	ICS Total	Weight	Mean Difference IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	5.2%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	5.7%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	5.3%	18.90 [5.45, 32.35]	2004
Strand 2004	40	52.14	78	14	49.64	72	3.6%	26.00 [9.71, 42.29]	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	5.1%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	4.7%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	52.14	288	40.1	49.64	289	13.8%	10.90 [2.59, 19.21]	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	9.2%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	18.3%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	13.3%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	15.8%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**18.59 [15.51, 21.68]**	

Heterogeneity: Tau² = 0.00; Chi² = 6.68, df = 10 (P = 0.76); I² = 0%
Test for overall effect: Z = 11.81 (P < 0.00001)



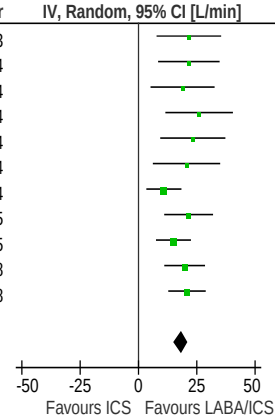
*The two studies with imputed SDs are indicated in boxes.

Alternatively, investigators could use the arithmetic means of the SDs in each group. That is, for the LABA/ICS group, take (46.2 + 51.14 + 45.8 + … + 40.58)/9 = 45.28. This results in 43.47 for the ICS group. Using these values for the two missing studies yields similar results to imputing using the maximum (Figure 3).

Figure 3. Results of meta-analysis of pulmonary function, imputing missed data using arithmetic means.*

Study or Subgroup	LABA/ICS Mean [L/min]	LABA/ICS SD [L/min]	LABA/ICS Total	ICS Mean [L/min]	ICS SD [L/min]	ICS Total	Weight	Mean Difference IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	4.9%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	5.4%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	5.0%	18.90 [5.45, 32.35]	2004
Strand 2004	40	45.28	78	14	43.47	72	4.5%	26.00 [11.79, 40.21]	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	4.8%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	4.5%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	45.28	288	40.1	43.47	289	17.2%	10.90 [3.66, 18.14]	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	8.7%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	17.4%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	12.6%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	15.0%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**18.36 [15.35, 21.36]**	

Heterogeneity: Tau² = 0.00; Chi² = 7.94, df = 10 (P = 0.63); I² = 0%
Test for overall effect: Z = 11.97 (P < 0.00001)



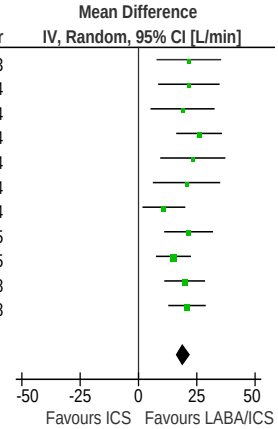
*The two studies with imputed SDs are indicated in boxes.

To use average coefficient of variation (CV) to impute, investigators need to first calculate a CV for each study. CV is defined as SD/mean. For example, for the Nelson study, CV=46.2/51.5 = 0.897. Computing CV for each study and then taking the average gives 0.921 for the LABA/ICS group and 1.527 for the ICS group. To estimate the SD for studies with a missing SD, use these values and the formula SD = CV*mean. In this case, for the Strand study, the mean is 40 in the LABA/ICS group, and the estimate of SD is 40*0.921 = 36.84. Using this method gives similar results to the previous two methods (Figure 4).

Figure 4. Results of meta-analysis of pulmonary function, imputing missed data using coefficient of variation.*

Study or Subgroup	LABA/ICS Mean [L/min]	LABA/ICS SD [L/min]	LABA/ICS Total	ICS Mean [L/min]	ICS SD [L/min]	ICS Total	Weight	Mean Difference IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	4.9%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	5.4%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	5.0%	18.90 [5.45, 32.35]	2004
Strand 2004	40	36.84	78	14	21.38	72	9.9%	26.00 [16.45, 35.55]	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	4.8%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	4.5%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	46.97	288	40.1	61.23	289	11.4%	10.90 [2.00, 19.80]	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	8.7%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	17.4%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	12.7%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	15.0%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**19.21 [16.20, 22.22]**	

Heterogeneity: Tau² = 0.00; Chi² = 7.61, df = 10 (P = 0.67); I² = 0%
Test for overall effect: Z = 12.50 (P < 0.00001)



*The two studies with imputed SDs are indicated in boxes.

More complex methods for calculating a weighted mean difference directly in the presence of missing SD data include sample size weights,[33] bootstrap methods,[34] multiple imputation methods,[35,36] the interval method,[37] and the prognostic method.[37] These methods are complex and don't permit the creation of a standard forest plot. While these methods may yield more accurate accounting of the true variance in the meta-analysis, this has yet to be fully evaluated. Other work has been done taking into account the uncertainty of the SD when it is imputed.[38,39] A full accounting of these methods is beyond the scope of this paper and

investigators are encouraged to look more into each of these methods themselves. There is not yet enough evidence to indicate the relative performance of the various approaches, though there is some evidence that the method chosen may not make a meaningful difference.[21,40]
To summarize, investigators should always try to contact authors to request exact estimates. Studies missing only SDs should not be excluded as this may lead to a biased combined estimate when studies with nonsignificant findings are more likely to omit SDs. If exact estimates cannot be obtained, imputation using one of the methods described above should be conducted. Direct substitution using the largest SD is the simplest method and the most likely to lead to a conservative estimate. However, if one is comfortable with one of the more complex methods, using it may lead to a more accurate estimate of precision parameter and is encouraged. No method has been shown to be absolutely superior to any other, so it is most important that the reviewer choose a valid method with which they are comfortable. Investigators may choose to use alternative imputation methods in a sensitivity analysis to determine how robust the results are with respect to the different imputation methods. It is also recommended that investigators report which studies had imputed SDs and which method(s) was used to perform the imputations.

Missing Correlations

To calculate the SD for change from baseline when meta-analyzing change from baseline scores, the correlation between baseline and followup scores is required in addition to the SDs for baseline and followup scores. This information is often not available from trial reports and has to be imputed.

The first option for imputation is to use estimates of correlation from other similar studies included in the same meta-analysis. If a study gives the SDs for both individual scores as well as for the change score, one can compute the correlation (r) using the following formula (which is a rearrangement of formula [4]):

$$r = \frac{SD_b^2 + SD_f^2 - SD^2}{2SD_b SD_f} \qquad (10)$$

where SD_b, SD_f, and SD represent the SDs for baseline, followup, and change scores, respectively. This correlation can be used as an estimate of the correlation in studies where the SD for change scores is not available but the SDs for baseline and followup scores are available.

If it is not possible to compute a correlation from any of the included studies, one can either estimate it from historical data or use an approximate value. In the latter case, the most common value to use is 0.5.[29] This can be considered a conservative estimate when using the change scores from baseline. A recent study[20] showed that the median correlation for change from baseline among trials included in systematic reviews was 0.59 (interquartile range [IQR]: 0.40, 0.81). A correlation less than 0.5 would make using followup scores generally more efficient than using the change scores from baseline. Thus if a trial author used the change scores from baseline, we can assume the correlation was at least 0.5. As in the case of missing SDs, investigators can always conduct sensitivity analyses by assuming several values of correlation.

The methods described here can also be used for dealing with crossover studies, in which case r would be calculated by rearranging formula (8).

Missing Individuals and Missing Study-Level Characteristics

Individuals missing from a study due to withdrawal and other reasons create an issue at the study level more than at the meta-analysis level. While missing individuals will also affect the results of meta-analysis, it is very difficult to deal with at the meta-analysis level without access to individual patient data. Nevertheless, three methods have been proposed to account for missing patient data: reweighting by completion rate, incorporation of the completion rate into a Bayesian random-effects model, and inference based on a Bayesian shared-parameter model (including the completion rate).[41]

Missing study-level characteristics will not affect the overall meta-analysis but can affect or even prevent subgroup analysis and meta-regression. Bayesian methods have been suggested to account for missing study-level data during meta-regression,[42] but these issues are complex and do not specifically pertain to continuous data. No particular methods are recommended, and investigators may try the methods outlined above for exploratory purposes.

Dealing With Skewed Data

Most meta-analytic techniques for continuous data are based on the mean of the variable of interest, for example, a clinical outcome and a measure of dispersion. If the variable's distribution is asymmetric, then the data are classified as skewed.

Meta-analytic methods based on means provide correct inference when the individual studies have sufficiently large sample sizes regardless of the variable's distribution due to the Central Limit Theorem, or when the variable of interest is at least approximately normally distributed.[43] However, if neither the sample size is sufficient nor the variable of interest is approximately normal, ignoring variable skewness or treating skewness inadequately can result in misleading conclusions. We know of no comprehensive survey or simulation study addressing the range of possible results of ignoring skewness. However, several examples are available that demonstrate the effects. For example, Ziguras et al.[44] compared two meta-analyses of interventions to reduce alcohol consumption, one of which excluded skewed data and one of which did not. The difference in handling skewed data was discussed as one of the reasons that the two analyses produced different results. Shen et al.[45] provided an example regarding the relationship between hospital ownership and financial performance in which disregarding skewness produced misleading results.

Typically, an individual study would report nonparametric summaries such as the median and interquartile range if the variable's distribution is not symmetric. However, the variable of interest may be suspected to be skewed and yet an individual study will report parametric summaries, that is, the mean and SD (or SE or variance). Alternatively, for variables with a skewed distribution, an individual study may transform the data and present either summary statistics on the transformed scale or different statistics, for example, the geometric mean, on the raw (original) scale.

Assessing Skewness

When nonparametric summaries are reported in individual studies, the study authors often have evidence of skewness in the data. Thus, prior to beginning analysis, we recommend that the meta-analyst carefully consider the distribution of each variable of interest and assess whether the distribution may be skewed. This assessment should be based on substantive knowledge of the variable and prior data, if available. For example, utilization and cost variables

are often skewed due to a subpopulation of users with no use, and thus no cost, and a few individuals with very high use and hence high cost. When median (or mean) with IQR or range are reported, some idea about the distribution usually can be gained. The two end points of IQR and range are not symmetric around median (or mean) if the distribution of the data is skewed. Altman and Bland[46] also provide two useful checks for skewness. If the mean is smaller than twice the SD in each intervention group, the data are likely to be skewed. If there are data from several groups of individuals, and the SD increases as the mean increases across these groups, this indicates that the data are positively skewed. However, data needed for the second check for skewness often may not be reported in the individual studies.

Using Nonparametric Summaries Assuming Symmetry

If symmetry is assumed, nonparametric statistics like medians, ranges, and interquartile ranges can be used to estimate both means and SDs. These nonparametric summaries are only estimates of the true parameters, unlike the direct calculations in the section "Calculating Standard Deviation and Standard Error When They Are Not Directly Reported." Depending on sample size, different nonparametric summary methods have been used to obtain means from either the median or the range and SDs from either the range or the interquartile range.[21, 23, 47]

The median is similar to the mean when the variable distribution is symmetric. Thus, if an individual study reports the median for a variable of interest, the median can be used in place of the mean to calculate the mean difference. Most past analyses have used a simple direct substitution of median, but there is a recent study [47] showing that if the range (i.e., the minimum [a] and maximum [b] values) are given, a better estimate of the mean for sample sizes less than 25 is:

$$\overline{X} = \frac{a + 2m + b}{4} \qquad (11)$$

while the median itself remains the best estimator for sample sizes greater than 25.

For estimating SD, the most common practice has been to simply compute it from the range or IQR. IQR indicates the length of the interval between the 25th percentile and 75th percentile in which the central 50 percent of the sample values of the variable lie. In these situations, SD can be estimated as IQR/1.35 or as range/4. Hozo[47] suggested that range/4 should be used for sample sizes between 15 and 70, while range/6 should be used for sample sizes greater than 70. For sample sizes smaller than 15, the formula below can be used to estimate SD:

$$SD = \sqrt{\frac{1}{12}\left\{\frac{(a - 2m + b)^2}{4} + (b - a)^2\right\}} \qquad (12)$$

Since range is inherently dependent upon sample size, Wiebe[21] suggests that the table below reproduced from Pearson[48] (see Table 1) should be used to impute SD from range. The SD can be determined simply by dividing the range by the given divisor (which represents the percentage limit for the distribution of the range in a normal population).

Look up the sample size on Table 1 and use the given divisor. For example, if the sample size is 22, then SD could be estimated as range/3.819. It should be noted that Table 1 assumes that the sample data is drawn from a normal distribution. Investigators should use it only when the distribution of data is at least symmetric.

Dealing With Skewness

If skewness is suspected, and individual studies present nonparametric summaries, one can estimate the mean and SD and proceed with usual meta-analysis methods using the resulting estimates. This approach works if the skewness is at most moderate, for example, when the variable of interest has a symmetric distribution in most included studies but shows some skewness in others. However, in the case of significant skewness, for example, when the distribution of the variable of interest is consistently skewed across studies, we recommend transforming the summary statistics of the variable of interest to reduce skewness. An additional advantage of such a transformation can be increased clinical interpretability.[43] Generally a logarithmic transformation is used, particularly when the data are economic in nature. Some studies may report summaries on the logarithmic scale; the antilog of the mean of the log data is the geometric mean. Alternatively, study may present the geometric mean on the raw (original) scale alongside its confidence interval or SE. Investigators cannot combine summaries on the raw scale with summaries on the transformed scale. Higgins et al.[43] present methods for transforming between different scales which allow the meta-analyst to determine whether to conduct the meta-analysis on the raw scale or on the log-transformed scale as appropriate. Issues to take into consideration when choosing the scale include, for example, which scale was most commonly used across the individual studies.

Some recent research focuses on conducting nonparametric meta-analysis. For example, Ma et al.[37] discussed a nonparametric method that utilizes U-statistic theory. Such nonparametric approaches would obviate the need for distributional assumptions, be they normality or symmetry, but may be statistically inefficient. Other authors propose using a ratio of geometric means to analyze skewed continuous data;[49] however, the lack of clinician experience with geometric means may make such methods difficult to implement. Investigators may choose to explore these methods and compare them with the results of their primary analysis.

Standardized Mean Difference

For continuous outcomes, different studies in a meta-analysis may use a variety of instruments on different scales to assess the same outcome. For example, included trials might use the Beck Depression Inventory, the Geriatric Depression Scale, and the Center for Epidemiologic Studies Depression scale to measure depression. If these instruments are sufficiently similar to suggest that they are truly measuring the same outcome, standardized mean difference (SMD), a measure of effect size, could be used to combine the studies using different scales. In this section, we discuss the choice and interpretation of SMD estimates and offer caveats on using SMD.

Choice of Standardized Mean Difference

Commonly used estimates of SMD include Cohen's *d*, Hedges' *g*, and Glass'Δ,[50,51] which are all calculated by dividing the mean difference by the SD. The difference between the effect measures lies in the denominator: Glass' Δ uses the estimate of the SD from the control group:

$$\Delta = \frac{\overline{X}_T - \overline{X}_C}{SD_{Control}}, \qquad (13)$$

and the estimated variance for Glass' Δ is given by

$$\hat{Var}(\Delta) = \frac{n_T + n_C}{n_T n_C} + \frac{\Delta^2}{2(n_C - 1)}. \qquad (14)$$

Cohen's *d* divides by the maximum likelihood estimate of the common population SD, calculated as:

$$d = \frac{\bar{X}_T - \bar{X}_C}{S_P} \quad \text{where} \quad S_P = \sqrt{\frac{(n_T - 1)SD_T^2 + (n_C - 1)SD_C^2}{n_T + n_C}} \qquad (15)$$

where $\bar{X}_T - \bar{X}_C$ is the mean difference between the two intervention groups and SD_T and SD_C are the standard deviations of the two intervention groups.

Hedges' *g* uses the pooled sample SD, calculated as:

$$g = \frac{\bar{X}_T - \bar{X}_C}{S_{Pooled}} \quad \text{where} \quad S_P = \sqrt{\frac{(n_T - 1)SD_T^2 + (n_C - 1)SD_C^2}{n_T + n_C - 2}}. \qquad (16)$$

The estimated variance for Cohen's *d* and Hedges' *g* is given by

$$\hat{Var}(d) = \frac{n_T + n_C}{n_T n_C} + \frac{d^2}{2(n_T + n_C - 2)} \qquad (17)$$

and

$$\hat{Var}(g) = \frac{n_T + n_C}{n_T n_C} + \frac{g^2}{2(n_T + n_C - 2)}, \qquad (18)$$

respectively.

All three effect measures are biased to estimate the population standardized mean difference, and the bias can be more than trivial when the sample sizes of both intervention groups are small. Durlak[51] suggested that the positive bias "amounts to a 4 percent reduction in effect when the total sample size is 20 and around 2 percent when $N = 50$." Hedges [52] provided a formula to correct for this small sample bias for Hedges' *g* and to serve as an unbiased estimator of the population SMD:

$$g_{adj} = g * \frac{\Gamma(\frac{n_T + n_C - 2}{2})}{\sqrt{\frac{n_T + n_C - 2}{2}}\Gamma(\frac{n_T + n_C - 3}{2})} \qquad (19)$$

where $\Gamma(.)$ is the gamma function. The estimated variance for Hedges' g_{adj} is given by

$$\hat{Var}(g_{adj}) = \frac{n_T + n_C}{n_T n_C} + \frac{g_{adj}^2}{2(n_T + n_C)}. \qquad (20)$$

Under the equal variance assumption, Cohen's *d* and Hedges' *g* are more precise estimators than Glass' Δ, and Hedges' *g* has smaller sample variance than Cohen's *d*.

Hedges' unbiased estimator should be used whenever possible, especially when the sample sizes are smaller than 20. For sample sizes greater than or equal to 20, Hedges' *g* is generally preferred over Cohen's *d* or Glass' Δ. When sample size is large, the difference

between Hedges' *g* and Cohen's *d* is small and they can be used interchangeably. When variance across the groups differs and the control group may be a more accurate estimate of true population variance, Glass' Δ is preferable. Sensitivity analyses are recommended to check how the results differ between using Hedges' *g* and Glass' Δ.

Interpreting Values of Standard Mean Difference

In theory, SMD can be any number, positive or negative. SMDs of 0.2, 0.5, and 0.8 are suggested corresponding to small, medium, and large effects [53] and widely used, although they are not defined in meaningful clinical contexts. Conclusions about clinical importance of the differences are often not clear using SMDs.

We recommend that investigators consider back-transforming the combined SMD to the original scale to facilitate assessing the clinical importance of combined SMDs and to aid decision making. Back-transforming can be done by multiplying the SMDs by the SD of the original scale derived from the population representative studies. Since data from more than one scale are combined, investigators need to choose an SD for each scale they plan to back-transform. The standard deviation chosen for the back-transformation could be pooled from the individual studies included in the meta-analysis as long as they all use the same original scale, or from representative studies using the same scale. Whichever approach is taken, researchers are cautioned that back-transformation should only occur for the summary estimate of effect size and not for effect size results from individual studies, due to possible differences in variability across studies (Chapter 12, section 6).[27] The back-transformed mean difference should be evaluated for clinical importance according to evidence-based definitions of minimum clinically important differences.

A Worked Example To Illustrate Back-Transformation of the Pooled SMD

In a CER looking at the effectiveness of treatment in preschoolers at risk of attention deficit hyperactivity disorder (ADHD),[54] a meta-analysis was conducted to summarize the benefit of parent behavior training (PBT) for disruptive behavior disorder (DBD) in eight "good" quality studies. The outcome was the measured change in parent-rated child behavior, and scales used to measure the child disruptive behavior included the Eyberg child behavior inventory (ECBI), parental account of childhood symptoms (PACS), and reports of ADHD symptoms. The meta-analysis yielded a combined SMD of -0.68 (95% CI -0.88, -0.47), which corresponded to a medium effect size and indicated that PBT improved parent-rated child behavior in preschoolers. The original CER did not do back-transformation of SMD.

Four studies included in the meta-analysis used (the intensity subscale of) ECBI, and the SDs for the mean difference between PBT and the control groups were similar across studies, ranging from 33.0 to 36.8. To back-transform the combined SMD to the ECBI scale, as discussed above, the SD could be pooled from these four studies or from a representative study. If we take the second approach and consider the largest study, which has a SD of 36.8, to be a representative study then the back-transformed mean difference is -25.0 (95% CI -32.4, -17.3) on the ECBI scale.

Two studies included in the meta-analysis used PACS. One study had a sample size of 50 with a SD of 6.07 for the mean difference, and the other study had a sample size of 30 with a SD of 7.53 for the mean difference. If we use the pooled SD from the two studies to back-transform

the combined SMD, the pooled SD could be calculated as $\sqrt{\frac{6.07^2 * 50 + 7.53^2 * 30}{30 + 50}} = 6.65$, and the back-transformed mean difference is -4.5 (95% CI -5.9, -3.1) on the PACS scale.

Caveats on Using Standard Mean Difference

Synthesis of multiple scales adds complexity to the use and interpretation of SMD. Here are a few caveats investigators should consider when using SMD.

Sample variance heterogeneity. Some studies have identified bias associated with using SMD in heterogeneous studies and studies with large SD.[55] Inverse variance weighted SMD could produce a biased estimate of the mean SMD since the weight is a function of the observed SMD. Because the SMD is greatly influenced by the SD, factors affecting the SD will affect the SMD. Though SDs are not directly comparable when different measurement scales are used, if there are meaningful differences in variance across studies due to factors such as different inclusion criteria (e.g., one study includes only severely depressed participants, while another includes participants with mild, moderate, and severe depression), especially for the subset of studies using the same scale, then these differences in variance due to populations will affect the SMD.

The bias associated with the use of SMD is small when the true variance is small relative to the effect being estimated.[55] However, investigators should examine sample variance heterogeneity when combining SMDs across studies and evaluate how these differences could affect the meta-analysis. In studies using the same scale, this can be accomplished by doing subgroup analyses based on the magnitude of the SD. Subgroup analyses can also be done by grouping studies according to inclusion criteria. For example, in each subgroup, only SMDs from homogeneous populations should be combined (e.g., combining all studies limited to severely depressed participants, and comparing results to those from studies including mildly or moderately depressed samples). If subgroup analyses suggest that results differ, then SMDs should not be combined across all studies with heterogeneous populations.

Covariates. Studies may account for the effect of covariates. When combining SMDs, SMDs calculated using the unadjusted mean difference[56] should not be combined with SMDs adjusted for covariates if there is heterogeneity between the two sets of SMDs. For SMDs calculated from mean difference adjusted for covariates, investigators should consider combining only results with a similar degree of adjustment (e.g., adjusted for similar covariates) to ensure comparable effect size across studies. Otherwise, the combined estimate may be biased. If a study uses balanced groups based on important covariates (e.g., if it has achieved balance through adequate randomization), and another study adjusts for these same covariates, these two studies could be considered as having a similar degree of adjustment and could be combined in a meta-analysis.

Directionality. Note that the direction of the scale must be consistent across the scales used in the included studies. For example, if in one study a high score indicates depression and in another study a low score indicates depression, then one of the scores must be reverse-coded to account for scale direction differences. Investigators should assure that scales are converted to a consistent direction of effect across all studies when calculating SMD.

Missing standard deviation. Information from the SD is required when calculating SMD. When the SD is missing, investigators can use imputed SD; Furukawa et al.[57] showed that studies using

imputed SDs produced similar results to studies using known SD values. Furukawa et al. also discussed how imputing SD applies to SMD, and more information on imputing SD is provided in the above section "Dealing with Missing Data."

Multiplicity of data. Studies often report data from outcomes based on multiple measures from multiple time points, an important source of possible bias in meta-analysis.[58] For example, one trial may assess an outcome using five measures assessed at three time points; the results may be published in four separate articles. Investigators should establish *a priori* inclusion criteria regarding which outcomes and time points should be used in a meta-analysis and make sure that all outcome measures meeting inclusion criteria are included. Outcome measures should not be excluded on the basis of statistical significance, direction of effect, or magnitude of effect, since such exclusions would result in selection bias. Investigators must also make sure that only one outcome measure is included in a single meta-analysis. Sensitivity analyses may be conducted to assess the impact of the different measures (for the same outcome) on the combined estimate. In addition, investigators should note that the multiplicity of data is a potential issue for all continuous outcomes. This applies to other effect measures, including mean difference and RoM.

Ratio of Means

Mean difference or SMD have been the most commonly used measures in meta-analysis for continuous outcomes. Recently, RoM[1,2] was proposed as an alternative. This measure offers the advantage that it can be used regardless of the units used in the individual trials. As with SMD, RoM can be used to combine outcomes that are measured using different scales. RoM can be interpreted in terms of the percentage change of the intervention group from the control group.

The RoM is calculated by dividing the mean outcome value from the intervention (or treatment) group ($\bar{X}_T$) by the mean outcome value from the control group ($\bar{X}_C$). For meta-analysis, the natural logarithm of each trial's RoM and its SE are calculated using the mean values, number of participants (n), and SD in each group[2] as:

$$\log(\text{RoM}) = \log\left(\frac{\bar{X}_T}{\bar{X}_C}\right) \qquad (21)$$

$$SE\left(\log(\text{RoM})\right) = \sqrt{\frac{1}{n_T}\left(\frac{SD_T}{\bar{X}_T}\right)^2 + \frac{1}{n_C}\left(\frac{SD_C}{\bar{X}_C}\right)^2} \qquad (22)$$

Then the natural logarithm transformed ratios are combined across studies using the standard inverse variance method. A combined ratio and its 95% confidence interval could be obtained by back-transforming the combined log-transformed ratio and its 95% confidence interval:

$$\text{RoM} = \exp(\log(\text{RoM})_{pooled}) \qquad (23)$$

$$\text{95\% Confidence Interval: } \exp\left(\log(\text{RoM})_{pooled} \pm 1.96 \times SE(\log(\text{RoM})_{pooled})\right) \qquad (24)$$

This method can be employed using a free meta-analysis software package called COMPARE2.[59]

RoM has a straightforward interpretation and expresses the percentage change in the mean value of the intervention group relative to the control group. The results are in a relative

form similar to the risk ratio: For example, if the combined RoM is 1.15, it means that the mean of the intervention group is 15 percent higher than the control group; if the combined RoM is 0.85, then the mean of the intervention group is 15 percent lower than the control group.

In simulation studies,[2] RoM has shown comparable statistical performance to mean difference methods in terms of bias, coverage probability, and statistical power. Overall, the data suggest that RoM is a reasonable alternative. Further data from an empirical analysis of 232 clinically diverse published meta-analyses[1] have confirmed the findings of simulated data, and this study suggests that, on average, RoM produces similar effect estimates, and SMDs of 0.2, 0.5, and 0.8 corresponded to increases in mean of 8, 22, and 37 percent, respectively. There was less heterogeneity in meta-analyses using RoM compared with mean difference but more compared with SMD.

Several meta-analyses have used RoM.[60-63] One study[62] utilized the RoM method when included studies reported various units of dosing for analgesics for a meta-analysis of total analgesic used within a postoperative period. Traditional methods would require standardizing all analgesic doses (i.e., conversion to "morphine equivalent"), which was not possible in all cases since not all analgesics have a reliable equivalence ratio. The treatment effect of cumulative analgesics used was therefore expressed as RoM in the experimental versus the control groups.

In summary, RoM appears to be a reasonable alternative to the traditional effect measures of continuous outcomes based on empirical evidence. Therefore, investigators may choose RoM as an effect measure when appropriate. When the outcome is assessed using different scales, RoM is easier to interpret than SMD. RoM has no units and allows for pooling of the studies expressed in different units; RoM also facilitates comparisons regarding relative effect sizes across different interventions. On the other hand, investigators should note that RoM can only be used in scenarios where the mean values of the intervention and control groups are both positive or both negative. Caution is warranted when RoM is used for small trials with large SDs and large effect sizes. Similar to the limitation of SMD for small trials, the combined estimate of RoM biases towards no effect, and this bias is accentuated by high heterogeneity.

Dichotomizing Continuous Outcomes in Meta-Analyses

For some continuous outcomes, a meaningful clinically important change is often defined, and patients achieving such change are considered as "responders."[64] There are methods developed to convert effect measures for continuous outcomes to effect measures of binary outcomes;[65,66] however, understanding the relationship between continuous effect measures and proportion of "response" is not straightforward. The assumptions used to assess such relationships are usually difficult to verify,[66] and the results could be sensitive to underlying assumptions.[65] Further research is necessary, and we currently recommend against inferring response rate from a combined mean difference.

Conclusion

In this report, we have provided recommendations on relevant topics applicable to quantitative synthesis of continuous outcomes measured in RCTs. The key points and recommendations for each topic are summarized in Table 2. Investigators are encouraged to follow these recommendations to improve the quality, transparency, and consistency of quantitative synthesis. The recommendations will be updated with the development of new research and methods, and new topics will be added when needs arise.

Table 2. Summary of key points and recommendations for quantitative synthesis of continuous outcomes in comparative effectiveness reviews

Methods for Quantitative Synthesis of Continuous Outcomes	Key Points and Recommendations
Inclusion of continuous outcomes	• Investigators should establish a priori inclusion criteria regarding which outcomes and time points should be used in a meta-analysis and make sure that all outcome measures meeting inclusion criteria are included. Outcome measures should not be excluded on the basis of statistical significance, direction of effect, or magnitude of effect.
Mean difference	• Mean difference should be used if results are reported using the same or similar scales. • There are three major estimates for mean difference: (1) mean difference of followup score, (2) mean difference of the change score, and (3) the ANCOVA estimate. • Estimates from options 1, 2, or 3 could be combined in one single meta-analysis.
Assessment of baseline imbalance	• Investigators should assess baseline balance of included trials in quantitative synthesis. • Assessing baseline balance based on statistical testing of homogeneity among treatment groups for individual trials is not generally recommended. • There are no concrete criteria to determine balanced versus imbalanced distribution and the decision could be subjective. The actual differences between baseline measurements, clinically important differences, and the direction of the imbalance are important considerations. • When the decision is not readily clear cut, the investigators should conservatively consider the baseline scores to be imbalanced.
Choice of estimates for mean difference under no or minimal baseline imbalance	• Estimates from options 1, 2, or 3 are all unbiased and appropriate to use. • The investigators should first use an ANCOVA estimate. If it is not reported and investigators could obtain the mean difference based on both options 1 and 2 (see Mean difference above), use the estimate that has a smaller SE. Otherwise, use either option 1 or 2 based on the available reported data of the included study. • The investigators may choose to use the same estimate across studies in one meta-analysis. • Data on standard deviation or standard error may not be reported but often can be calculated or imputed.
Choice of estimates for mean difference under baseline imbalance	• The ANCOVA estimates are least biased with more precision, and they are preferred. Options 1 and 2 provide biased estimates. • The investigators should first use ANCOVA estimates, and if they are not reported, the investigators should conduct analyses using both estimates and report the more conservative combined estimate, which is usually the one with a smaller absolute effect size. • If enough trials in a meta-analysis reported ANCOVA estimates, the investigators are encouraged to conduct subgroup analyses to compare results from ANCOVA versus non-ANCOVA estimates as sensitivity analyses.

Table 2. Summary of key points and recommendations for quantitative synthesis of continuous outcomes in comparative effectiveness reviews (continued)

Methods for Quantitative Synthesis of Continuous Outcomes	Key Points and Recommendations
Calculation of standard deviation and standard error	• Depending on the software package used, either standard deviation or standard error will be required from each study in order to be included in the meta-analysis. These quantities are often not given directly, but can be easily computed from confidence intervals, exact p-values, z-statistics, and t-statistics. • Studies with a crossover design or a cluster-randomized design have design effects that must be taken into account when computing their standard errors. Ignoring this design effect will tend to overestimate standard error for crossover studies and underestimate it for cluster randomized studies.
Dealing with missing data	• In general, studies containing information on point estimate but missing data on standard deviation or standard error should be included in a meta-analysis using imputed standard deviation or standard error. • Whenever possible, as a first recourse, contact study authors to obtain missing data. • If authors cannot provide information on missing data, investigators should perform imputation of standard deviation. • There is no consensus as to which method of imputation is best, and most methods tend to give similar results. Sensitivity analyses can be performed to check the robustness of results in regards to the choice of imputation methods.
Dealing with skewed data	• Assess whether a variable may be skewed, based on substantive knowledge of the variable and any available data. If possible, the approach described in Altman and Bland[46] should be applied. • If approximate symmetry could be assumed for a variable and nonparametric summaries are reported in the included trials (e.g., median, interquartile range, range), estimate the mean and standard deviation from nonparametric summaries for use in meta-analysis. • If a variable is skewed, transform the data to reduce skewness, for example, via a logarithmic transformation, and conduct the meta-analysis on the transformed scale. • Conduct sensitivity analysis to assess how robust conclusions are in regards to different transformations and other methodological choices.
Standardized mean difference	• Standardized mean difference should be used if included studies use different continuous scales to measure the same outcome. • Hedges' unbiased estimator and Hedges' *g* are generally preferred. When variance across the groups differs and the control group may be a more accurate estimate of true population variance, Glass' Δ is preferable. • SMDs of 0.2, 0.5, and 0.8 correspond to small, medium, and large effects. • Investigators should back-transform the pooled SMD to the original scale to facilitate assessing the clinical importance of the combined estimate. • Investigators should consider the impact of sample variance heterogeneity and degree of covariate adjustment when combining SMD. • Investigators need to make sure that the directions of the included scales are consistent. • When SD is missing, investigators could use imputed SD.
Ratio of means	• Investigators could choose RoM as an alternative option for meta-analyzing continuous variables assessed using different scales in the same direction. • RoM should be used with caution for small trials with large standard deviations and larger effect size.
Dichotomizing continuous outcomes in meta-analyses	• We currently recommend against inferring response rate from a combined mean difference.

Abbreviations: RoM = ratio of means, SD = standard deviation, SE = standard error, SMD = standard mean difference.

Abbreviations

ADHD	Attention deficit hyperactivity disorder
ANCOVA	Analysis of covariance
DBD	Disruptive behavior disorder
ECBI	Eyberg child behavior inventory
EPC	Evidence-based Practice Center
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
PACS	Parental account of childhood symptoms
PBT	Parent behavior training
RCT	Randomized clinical trial
RoM	Ratio of means
SD	Standard deviation
SE	Standard error
SMD	Standardized mean difference

References

1. Friedrich JO, Adhikari NK, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. J Clin Epidemiol. 2011 May;64(5):556-64. PMID: 21447428.
2. Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Med Res Methodol. 2008;8:32. PMID: 18492289.
3. Senn S. Baseline distribution and conditional size. J Biopharm Stat. 1993 Sep;3(2):265-76. PMID: 8220409.
4. Charpentier G, Fleury F, Kabir M, et al. Improved glycaemic control by addition of glimepiride to metformin monotherapy in type 2 diabetic patients. Diabet Med. 2001 Oct;18(10):828-34. PMID: 11678974.
5. Rosenberger W LJ, ed. Randomization in Clinical Trials: Theory and Practice: New York: Wiley; 2002.
6. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306):614-8. PMID: 11867132.
7. Trowman R, Dumville JC, Torgerson DJ, et al. The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. J Clin Epidemiol. 2007 Dec;60(12):1229-33. PMID: 17998076.
8. Berger VW, Weinstein S. Ensuring the comparability of comparison groups: is randomization enough? Control Clin Trials. 2004 Oct;25(5):515-24. PMID: 15465620.
9. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ. 1999 Jul 17;319(7203):185. PMID: 10406763.
10. Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994 Sep 15;13(17):1715-26. PMID: 7997705.
11. Senn SJ. Covariate imbalance and random allocation in clinical trials. Stat Med. 1989 Apr;8(4):467-75. PMID: 2727470.
12. Altman DG. Comparability of randomised groups. Statistician. 1985;34:125-36.
13. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet. 1990;335(8682):149-53. PMID: 1967441.
14. Begg CB. Suspended judgment. Significance tests of covariate imbalance in clinical trials. Control Clin Trials. 1990(11):223-5. PMID: 2171874.
15. Austin PC, Manca A, Zwarenstein M, et al. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol. 2010 Feb;63(2):142-53. PMID: 19716262.

16. Hewitt CE, Kumaravel B, Dumville JC, et al. Assessing the impact of attrition in randomized controlled trials. J Clin Epidemiol. 2010 Nov;63(11):1264-70. PMID: 20573482.
17. Senn S. Change from baseline and analysis of covariance revisited. Stat Med. 2006 Dec 30;25(24):4334-44. PMID: 16921578.
18. Crager MR. Analysis of covariance in parallel-group clinical trials with pretreatment baselines. Biometrics. 1987 Dec;43(4):895-901. PMID: 3427174.
19. Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323:1123-4. PMID: 11701584.
20. Balk EM, Earley A, Patel K, et al. Empirical Assessment of Within-Arm Correlation Imputation in Trials of Continuous Outcomes. Methods Research Report. AHRQ Publication No. 12(13)-EHC141-EF. Rockville, MD: Agency for Healthcare Research and Quality; November 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.
21. Wiebe N, Vandermeer B, Platt RW, et al. A systematic review identifies a lack of standardization in methods for handling missing variance data. J Clin Epidemiol. 2006 Apr;59(4):342-53. PMID: 16549255.
22. Health Services Research Unit. Database of ICCs. University of Aberdeen Chief Scientist Office. www.abdn.ac.uk/hsru/research/delivery/behaviour/methodological-research. Accessed April 18, 2013.
23. Cook JA, Bruckner T, MacLennan GS, et al. Clustering in surgical trials--database of intracluster correlations. Trials. 2012;13:2. PMID: 22217216.
24. Ukoumunne OC, Gulliford MC, Chinn S, et al. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technol Assess. 1999;3(5):iii-92. PMID: 10982317.
25. Taljaard M, Donner A, Villar J, et al. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. Paediatr Perinat Epidemiol. 2008 Mar;22(2):117-25. PMID: 18298685.
26. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ: Wiley; 1987.
27. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [Updated March 2011]. In: Higgins JPT, Green S, eds.: The Cochrane Collaboration; 2011.
28. Idris N, Robertson C. The effects of imputing the missing standard deviations on the standard error of meta analysis estimates. Communiations in Statistics - Simulaiton and Computation. 2009;38(3):513-26.
29. Follmann D, Elliott P, Suh I, et al. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol. 1992 Jul;45(7):769-73. PMID: 1619456.
30. Pigott T. Methods for handling missing data in research synthesis. In: Cooper H, Hedges LV, eds. Handbook of Research Synthesis. New York: Sage Publications; 1994:163-76.
31. Bracken M. Statistical methods for analysis of effects of treatment in overviews of randomized trials. In: Sinclair JD, Bracken MD, eds. Effective Care of the Newborn Infant. New York: Oxford University Press; 1992:13-20.
32. Bond K, Coyle D, O'Gorman K, et al. Long-Acting Beta2-Agonist and Inhaled Corticosteroid Combination Therapy for Adult Persistent Asthma: Systematic Review of Clinical Outcomes and Economic Evaluation. [Technology report number 122]. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2009. www.cadth.ca/en/products/health-technology-assessment/publication/941.
33. Sanchez-Meca J, Marin-Martinez F. Weighting by inverse variance or by sample size in meta-analysis: a simulation study. Educ Psychol Meas. 1998;58(2):211-20.
34. Zhu W. Making bootstrap statistical inferences: a tutorial. Res Q Exerc Sport. 1997 Mar;68(1):44-55. PMID: 9094762.
35. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med. 1991 Apr;10(4):585-98. PMID: 2057657.
36. Stevens JW. A note on dealing with missing standard errors in meta-analyses of continuous outcome measures in WinBUGS. Pharm Stat. 2011 Jul-Aug;10(4):374-8. PMID: 21394888.
37. Ma Y, Mazumdar M. Multivariate meta-analysis: a robust approach based on the theory of U-statistic. Stat Med. 2011 Oct 30;30(24):2911-29. PMID: 21830230.

38. White IR, Higgins JP, Wood AM. Allowing for uncertainty due to missing data in meta-analysis--part 1: two-stage methods. Stat Med. 2008 Feb 28;27(5):711-27. PMID: 17703496.
39. White IR, Welton NJ, Wood AM, et al. Allowing for uncertainty due to missing data in meta-analysis--part 2: hierarchical models. Stat Med. 2008 Feb 28;27(5):728-45. PMID: 17703502.
40. Thiessen Philbrook H, Barrowman N, Garg AX. Imputing variance estimates do not alter the conclusions of a meta-analysis with continuous outcomes: a case study of changes in renal function after living kidney donation. J Clin Epidemiol. 2007 Mar;60(3):228-40. PMID: 17292016.
41. Yuan Y, Little RJ. Meta-analysis of studies with missing data. Biometrics. 2009 Jun;65(2):487-96. PMID: 18565168.
42. Hemming K, Hutton JL, Maguire MG, et al. Meta-regression with partial information on summary trial or patient characteristics. Stat Med. 2010 May 30;29(12):1312-24. PMID: 20087842.
43. Higgins JP, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. Stat Med. 2008 Dec 20;27(29):6072-92. PMID: 18800342.
44. Ziguras SJ, Stuart GW, Jackson AC. Assessing the evidence on case management. Br J Psychiatry. 2002 Jul;181:17-21. PMID: 12091258.
45. Shen YC, Eggleston K, Lau J, et al. Hospital ownership and financial performance: what explains the different findings in the empirical literature? Inquiry. 2007 Spring;44(1):41-68. PMID: 17583261.
46. Altman DG, Bland JM. Detecting skewness from summary information. BMJ. 1996 Nov 9;313(7066):1200. PMID: 8916759.
47. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol. 2005;5:13. PMID: 15840177.
48. Pearson ES. The percentage limits for the distribution of range in samples from a normal population (n less than or equal to 100). Biometrika. 1932;24(3-4):404-17.
49. Friedrich JO, Adhikari NK, Beyene J. Ratio of geometric means to analyze continuous outcomes in meta-analysis: comparison to mean differences and ratio of arithmetic means using empiric data and simulation. Stat Med. 2012 Jul 30;31(17):1857-86. PMID: 22438170.
50. Card NA. Applied Meta-analysis for Social Science Research. 1st ed. New York: The Guilford Press; 2012.
51. Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol. 2009 Oct;34(9):917-28. PMID: 19223279.
52. Hedges LV. Estimation of effect size from a series of independent experiments. Psychological Bulletin. 1982 September;92(2):490-9.
53. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
54. Charach A, Dashti B, Carson P, et al. Attention Deficit Hyperactivity Disorder: Effectiveness of Treatment in At-Risk Preschoolers; Long-Term Effectiveness in All Ages; and Variability in Prevalence, Diagnosis, and Treatment. Comparative Effectivess Review No. 44. AHRQ Publication No. 12-EHC003-EF. Rockville, MD: Agency for Healthcare Research and Quality; October 2011. www.effectivehealthcare.ahrq.gov/reports/final.cfm.
55. Van Den Noortgate W, Onghena P. Estimating the mean effect size in meta-analysis: bias, precision, and mean squared error of different weighting methods. Behav Res Methods Instrum Comput. 2003 Nov;35(4):504-11. PMID: 14748494.
56. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007 Nov;82(4):591-605. PMID: 17944619.
57. Furukawa TA, Barbui C, Cipriani A, et al. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol. 2006 Jan;59(1):7-10. PMID: 16360555.
58. Tendal B, Nuesch E, Higgins JP, et al. Multiplicity of data in trial reports and the reliability of meta-analyses: empirical study. BMJ. 2011;343:d4829. PMID: 21878462.
59. WINPEPI Program COMPARE2 [Internet]. www.brixtonhealth.com/pepi4windows.html. Accessed April 18, 2013.

60. Adhikari NK, Burns KE, Friedrich JO, et al. Effect of nitric oxide on oxygenation and mortality in acute lung injury: systematic review and meta-analysis. BMJ. 2007 Apr 14;334(7597):779. PMID: 17383982.
61. Kunz R, Friedrich C, Wolbers M, et al. Meta-analysis: effect of monotherapy and combination therapy with inhibitors of the renin angiotensin system on proteinuria in renal disease. Ann Intern Med. 2008 Jan 1;148(1):30-48. PMID: 17984482.
62. Peng PW, Wijeysundera DN, Li CC. Use of gabapentin for perioperative pain control -- a meta-analysis. Pain Res Manag. 2007 Summer;12(2):85-92. PMID: 17505569.
63. Sud S, Sud M, Friedrich JO, et al. High frequency oscillation in patients with acute lung injury and acute respiratory distress syndrome (ARDS): systematic review and meta-analysis. BMJ. 2010;340:c2327. PMID: 20483951.
64. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis. 2005 Jan;64(1):34-7. PMID: 15130902.
65. Anzures-Cabrera J, Sarpatwari A, Higgins JP. Expressing findings from meta-analyses of continuous outcomes in terms of risks. Stat Med. 2011 Nov 10;30(25):2967-85. PMID: 21826697.
66. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med. 2000 Nov 30;19(22):3127-31. PMID: 11113947.

Chapter 14 Appendix A. Search Strategies

Standardized Mean Difference

Ovid Medline (Date Searched 3/8/2012)

1	(standardized adj1 mean adj1 difference).ti,ab.	532
2	meta-analysis as topic/	12130
3	meta-analys$.ti,ab.	41814
4	exp statistics as topic/	1697404
5	meta-analysis.sh.	33853
6	2 or 3 or 5	59871
7	1 and 4 and 6	79

Current Index to Statistics (Date Searched 2/22/2012)

Keyword search using combinations of standardized mean difference

Baseline Imbalances

Ovid Medline (Date Searched 2/22/2012)

1	((imbalance* or balance* or distribution) and (pre-treatment or pretreatment or baseline or pre-intervention or preintervention or covariat*)).ti,ab.	18981
2	exp clinical trials as topic/	255550
3	meta-analysis as topic/	12130
4	"review literature as topic"/	4314
5	exp "bias (epidemiology)"/	45684
6	exp "analysis of variance"/	237153
7	((analys$ adj3 covarian$) or ANCOVA).ti,ab.	8690
8	data interpretation, statistical/	42335
9	3 or 4 or 5 or 6 or 7 or 8	338233
10	1 and 2 and 9	210

Current Index to Statistics (Date Searched 2/22/2012)

Keyword search using combinations of (imbalance* or balance* or distribution) and (pre-treatment or pretreatment or baseline or pre-intervention or preintervention or covariat*)

Scopus

Pearling search to identify additional relevant citations from relevant articles already identified.

Meta-analysis of Skewed Data

Ovid Medline (Date Searched: 3/8-20/2012), Current Index to Statistics, Scopus

We took the Higgins article (Higgins, White and Anzures-Cabrera, "Meta-analysis of skewed data: combining results reported on log-transformed or raw scales." Stats in Med 2008; 27:6072-6092.) as a starting point but were unable to define a subject search that worked, so we did a combination of keyword and pearling searches in Ovid Medline, Current Index to Statistics, and Scopus.

Means Ratios in Pooled Analyses and Categorizing for Continuous Outcomes

We searched Ovid MEDLINE(R) <1946 to January Week 4 2012> and PubMed on March 1st 2012 for (Dichotomis* or Dichotomiz*) limited to: Humans, Meta-Analysis, and English. We searched Web of Science for articles citing either of 2 known studies:

1. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. Journal of Clinical Epidemiology. 2011;64(11):1187-97.
2. Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Med Res Methodol. 2008;8:32. PMID: 18492289.) in combination with a known author/expert (Friedrich, JO). Experts and reviewers also recommended references based on experience and reference list checking.

U.S. Department of Health and Human Services
Agency for Healthcare Research and Quality
www.ahrq.gov

Agency for Healthcare
Research and Quality

AHRQ Pub. No. 10(13)-EHC063-EF
September 2013