Methods Guide for
Comparative Effectiveness Reviews

Selecting Observational Studies for Comparing Medical Interventions

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

Comparative Effectiveness Reviews are systematic reviews of existing research on the effectiveness, comparative effectiveness, and harms of different health care interventions. They provide syntheses of relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. Strong methodologic approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a *Methods Guide for Comparative Effectiveness Reviews*. This Guide presents issues key to the development of Comparative Effectiveness Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The *Methods Guide for 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 Comparative Effectiveness Reviews* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

This document was written with support from the Effective Health Care Program at AHRQ

None of the authors has a financial interest in any of the products discussed in this document.

Suggested citation: Norris S, Atkins D, Bruening W, et al. Selecting observational studies for comparing medical interventions. In: Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews [posted June 2010]. Rockville, MD.
Available at:
http://www.effectivehealthcare.ahrq.gov/ehc/products/196/454/MethodsGuideNorris_06042010.pdf

Selecting Observational Studies for Comparing Medical Interventions

Authors:

Susan Norris, M.D., M.P.H.[a]

David Atkins, M.D., M.P.H.[b]

Wendy Bruening, Ph.D.[c]

Steven Fox, M.D., S.M., M.P.H.[d]

Eric Johnson, Ph.D.[e]

Robert Kane, M.D.[f]

Sally C. Morton, Ph.D.[g]

Mark Oremus, Ph.D.[h]

Maria Ospina, M.Sc.[i]

Gurvaneet Randhawa, M.D., M.P.H.[d]

Karen Schoelles M.D., S.M.[c]

Paul Shekelle, M.D., Ph.D., M.P.H.[j]

Meera Viswanathan, Ph.D.[g]

[a]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR.
[b]VA Quality Enhancement Research Initiative (QUERI), Washington, DC.
[c]ECRI Institute, Plymouth Meeting, PA.
[d]Agency for Healthcare Research and Quality, Rockville, MD.
[e]The Center for Health Research, Kaiser Permanente Northwest, and Oregon Evidence-based Practice Center, Portland, OR.
[f]Minnesota Evidence-Based Practice Center, Minneapolis, MN.
[g]RTI International, Triangle Park, NC.
[h]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON.
[i]University of Alberta Evidence-Based Practice Centre, Edmonton, AB.
[j]Southern California Evidence-Based Practice Center, RAND Corporation, Los Angeles, CA

Selecting Observational Studies for Comparing Medical Interventions

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.[8] 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 (Figure 1):

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 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.

In other cases, content experts and decisionmakers 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.

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.

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;169:1167-1170.
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):502-512.
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;330:1053-7.
5. Norris SL, Atkins D. Challenges in using nonrandomized studies in systematic reviews of treatment interventions. *Ann Intern Med* 2005;142:1112-1119.
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005;58:323-337.
7. Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet* 2004;363:1728-31.
8. Higgins JP and 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;63:484-490.
13. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-1218.
14. GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490-1498.
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):513-523.
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;59:1040-1048.

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;158:915-920.
28. Hutten JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *J R Stat Soc Ser C* 2000;49:359-370.
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;291:2457-2465.
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;171:735-740.
31. Furukawa TA, Watanabe N, Omori IM, et al. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. *JAMA* 2007;297:468-470.
32. Peters J, Mengersen K. Selective reporting of adjusted estimates in observational epidemiology studies: reasons and implications for meta-analyses. *Eval Health Prof* 2008;31:370-389.
33. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-926.

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

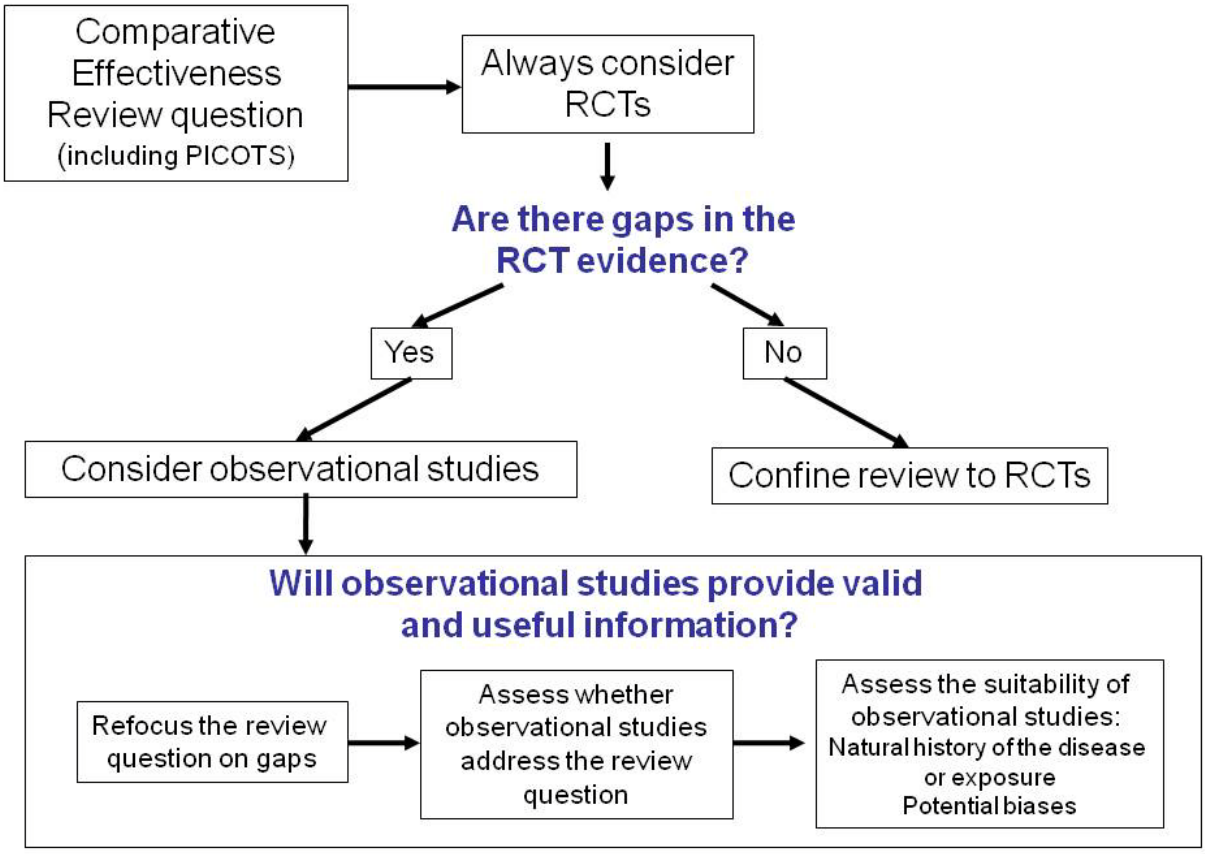
This table is adapted from the work of Owens and colleagues[15] and the work of the Methods Guide for Comparative Effectiveness Reviews: Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program.[16]

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.
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 of Healthcare Research and Quality; RCT=randomized controlled trial

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.