Methods Research Report

Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank

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

Methods Research Report

Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10056-I

Prepared by:
RTI International–University of North Carolina at Chapel Hill Evidence-based Practice Center, Research Triangle Park, NC
University of Alberta Evidence-based Practice Center
Edmonton, Canada

Investigators:
Meera Viswanathan, Ph.D.
Nancy D. Berkman, Ph.D.
Donna M. Dryden, Ph.D.
Lisa Hartling, Ph.D.

AHRQ Publication No. 13-EHC106-EF
August 2013

This report is based on research conducted by the RTI–UNC Evidence-based Practice Center, (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10056-I). 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. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

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

Suggested citation: Viswanathan M, Berkman ND, Dryden DM, Hartling L. Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank. Methods Research Report. (Prepared by RTI–UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13-EHC106-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.gov.

Carolyn M. Clancy, M.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 and Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Investigator Affiliations

Meera Viswanathan, Ph.D. [a]
Nancy D. Berkman, Ph.D. [a]
Donna M. Dryden, Ph.D. [b]
Lisa Hartling, Ph.D. [b]

[a] RTI International–University of North Carolina at Chapel Hill Evidence-based Practice Center, Research Triangle Park, NC
[b] The University of Alberta Evidence-based Practice Center, Edmonton, Canada

Acknowledgments

We gratefully acknowledge the contributions of the associate editor, Mark Helfand, M.D., M.S., M.P.H, the EPC editor, Jennifer Drolet, M.A., Sharon Barrell, M.A., research assistant, Priyanka Sista, B.A., and the EPC publications specialist, Loraine Monroe.

Workgroup Members

Mohammed T. Ansari, M.B.B.S., M.Med.Sc., M.Phil.
University of Ottawa Evidence-based Practice Center
Ottawa, Canada

Nancy Dreyer, M.P.H., Ph.D.
Outcome Sciences Inc.
Cambridge, MA

Susan Norris, M.D., M.Sc., M.P.H.
Oregon Health and Science University
Portland, OR,

Ida Sim, Ph.D., M.D.
University of California San Francisco School of Medicine
San Francisco, CA

Tatyana Shamliyan, M.D.
University of Minnesota Evidence-based Practice Center
Minneapolis, MN

Jan Vandenbroucke, M.D., Ph.D.
Dept. of Clinical Epidemiology, Leiden University Hospital
Leiden, Netherlands

Peer Reviewers

Ethan M. Balk, M.D., M.P.H.
Tufts Medical Center
Boston, MA

Donna Ciliska, R.N., Ph.D.
McMaster University Faculty of Health Sciences
Hamilton, Ontario, Canada

Kathleen Gans-Brangs
AstraZeneca Canada, Inc.
Wilmington, DE

Ian Shrier, M.D., Ph.D.
McGill University
Montreal, Quebec, Canada

James Restom, Ph.D.
ECRI Institute Evidence-based Practice Center
Plymouth Meeting, PA

Karen Robinson, Ph.D.
Johns Hopkins Evidence based Practice Center
Baltimore, MD

Barbara Mauger Rothenberg, Ph.D.
Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-based Practice Center
Chicago, IL

Alric Rüther, M.D.
Institute for Quality and Efficiency in Health Care
Cologne, Germany

Holger Schünemann, M.D., M.Sc., Ph.D.
McMaster University
Hamilton, Ontario, Canada

Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank

Structured Abstract

Objectives: To develop a framework for the assessment of the risk of bias and confounding against causality from a body of observational evidence, and to refine a tool to aid in identifying risk of bias, confounding, and precision in individual studies.

Methods: In conjunction with a Working Group, we sought to develop an overarching approach to assess the effect of confounding across the body of observational study evidence and within individual studies. We sought feedback from Working Group members on critical sources of bias most common to each observational study design type. We then refined and reduced the set of "core" questions that would most likely be necessary for evaluating risk of bias and confounding concerns for each design and refined the instructions provided to users to improve clarity and usefulness.

Results: We developed a framework that identifies additional steps necessary to evaluate the validity of causal claims in observational studies of benefits and harms from interventions. With the help of the Working Group, we narrowed the list of RTI Item Bank questions for evaluating risk of bias and precision from 29 to 16. Working Group members also provided their opinion of the most important questions for assessing risk of bias for four common observational study design types.

Conclusions: Attributing causality to interventions from such evidence requires prespecification of anticipated sources of confounding prior to the review, followed by appraisal of potential confounders at three levels: outcomes, studies, and the body of evidence. We propose a substantial expansion in the critical appraisal of confounding when systematic reviews include observational studies for evaluation of benefits or harms of interventions. Questions about burden, reliability, and validity remain to be answered. Consensus around specific items necessary for evaluating risk of bias for different types of observational study designs does not yet exist.

Contents

Tables

Figures

Appendixes

Introduction

A primary concern for systematic reviews of the effectiveness of interventions is to evaluate the causal relationship between the intervention and outcomes. Despite the basic nature of causality as the motivator for much of science, definitions of causality tend to be circular ("to cause" is to produce an effect, and "to produce" is to cause).[1] Definitions of causality that reference the counterfactual offer a way out of the circularity.[1] A causal effect is "a counterfactual contrast between the outcomes of a single unit under different treatment possibilities."[1, p. 1914] For example, a pregnant woman will deliver her child using a single mode of delivery. Although her selected mode of delivery may have resulted in respiratory distress syndrome, we can postulate a different outcome should the woman have employed another mode of delivery. This approach to thinking about causal inference implies that studies that seek to establish a causal link need to meet three conditions: (1) a causal contrast involving two or more well-defined interventions,[2] (2) independence between the counterfactual outcome and the intervention, and (3) each participant in the study having a positive probability (greater than zero) of being assigned to each of the evaluated interventions.[3]

Randomization offers a participant an equal probability of being assigned to each treatment alternative, distributes unmeasured confounding randomly, and clearly sets up contrasts between interventions. Because randomization ensures that these three conditions for causal inference are met, randomized clinical trials (RCTs) are generally considered the gold standard for evidence of benefits.

RCTs, however, cannot always be used to answer questions on the causal link between interventions or exposures and outcomes. RCTs may be unethical.[4] A review consisting of RCTs alone may provide insufficient information on adverse effects, long-term benefits,[5] or vulnerable subpopulations.[6] In the absence of sufficient evidence from RCTs, Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) guidance suggests that systematic reviews may include observational studies to help answer questions about the causal link between the intervention or exposure and outcomes.[7] This approach, mirrored by other recent guides developed by the Cochrane Collaboration and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group, offers cautious support for reliance on observational studies.[8-10]

In addition to other bias risks, when including observational study evidence, systematic reviewers have to contend with the possibility of confounding; that is, the potential that extraneous factors, rather than the factors of interest (the intervention or exposure), influenced the results.[1] "Confounding by indication" or "allocation bias" refers to a common cause that influences both treatment and outcome.[11] In the example of mode of delivery, fetal distress increases the likelihood that the mother will undergo cesarean delivery in addition to the likelihood of neonatal respiratory distress. Confounding from the indication of fetal distress makes it difficult to assess the independent effect of the mode of delivery on neonatal respiratory distress. When reviewers rely on evidence from observational studies, the inferences they draw need to account for potential important confounding, as researchers did in a recent asthma study.[a]

[a] As an example, a recent systematic review examined the association between the use of acetaminophen and risk of asthma.[12] A subsequent prospective cohort study attributed the association to confounding by indication.[13] That is, the study showed that individuals who subsequently developed asthma were taking acetaminophen to manage early infections of the lower respiratory tract; however, after adjustment for respiratory infections, or when acetaminophen use was restricted to non-respiratory tract infections, no association was found. Therefore, they concluded that acetaminophen use was not an independent risk factor for development of asthma, and that previous positive findings were due

In this project, we reviewed methodological considerations when evaluating the validity of studies included in systematic reviews; specifically, concerns related to risk of bias and confounding in observational studies. Then, based on earlier work and current project activities, we developed a framework for the assessment of the risk of bias and confounding across a body of observational study evidence and refined an existing tool to aid in identifying risk of bias and confounding in individual studies.

Empirical Assessments of Risk of Bias

The higher theoretical risk of some types of bias in observational studies compared with RCTs (described in detail in Appendix A, based on our overview of the literature) raises a fundamental question when assessing causality in observational studies: do observational studies routinely provide a more inflated estimate of effect than trials?[14] If so, can their results be discounted to arrive at a closer approximation of the true effect? Empirically, however, reviews have found no difference[15,16] or inconsistent differences between types of designs (that is, RCTs sometimes had smaller estimates of effect than observational studies and vice versa).[17,18] This unpredictability in direction and magnitude of effect means that systematic reviews cannot rely on a rule of thumb (based on type of design)[19] to discount evidence from observational studies. Instead, a careful assessment of the risk of bias in each individual observational study that accounts for its unique clinical context is necessary to evaluate the validity of estimates from observational studies.[14]

Another important question when assessing causality in observational studies is whether empirical assessments of risk of bias help to shape our understanding of the validity of evidence. MacLehose et al. were unable to show associations between study quality and relative risk in predictable ways, suggesting the need for improved instrumentation for evaluating risk of bias for different types of observational studies: they note "compromises and ambiguities" arising from the use of the same instrument for all study designs and that "[d]eveloping an instrument to assess and characterize different studies is an urgent priority."[16, p. 45] Concern for better instrumentation is echoed by other researchers. Shrier et al. note that ideally reviews would weigh the results of a study with the potential for bias, but that approach requires that quality scores "be highly correlated with bias; therefore, there must be agreement on which items create which biases, in which direction and of what magnitude."[19, p. 1208] Consensus on the direction and magnitude of bias caused by aspects of study design and performance does not yet exist.

Need for a Revised Approach to Evaluating Observational Studies

In a review of critical appraisal tools for assessing the risk of bias of observational studies, Deeks et al. identified key quality domains (background, sample definition and selection, interventions, outcomes, creation of treatment groups, blinding, soundness of information, followup, analysis of comparability, analysis of outcomes, interpretation, and presentation and reporting) but found no gold standard for how the assessment should best be accomplished.[20] Tools typically focus their assessments on either: (1) capturing manuscript authors' descriptions or reporting of the methods they used in designing or conducting particular elements of the study, or (2) judging the risk of bias based on the study's design or implementation (whether the

---

to confounding by indication. The researchers commented that setting up an RCT with the aim of studying the adverse effect would be infeasible. The example shows how diligent analyses of observational data can find solutions to problems of confounding by indication.

conduct of the study altered the validity of results). A recent review by Mhasker et al. reveals the inadequacies of risk of bias assessment tools, all of which rely in varying degrees on reporting. They compared quality in protocols to reported quality in publications from 429 RCTs. Their results showed that reported quality (from publications) did not adequately reflect the actual high quality of the trials (from protocols); moreover, the associations between poor allocation concealment or blinding and effect size as reported in publications did not persist when examining the association between these two quality domains and effect size based on descriptions in the protocols.[21] Although the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement may improve the reporting of observational studies[22] in the future, risk of bias instruments for observational studies that rely on reporting suffer from the same constraints as risk of bias instruments for trials and study conduct is less likely to be based on published protocols.

Another constraint of some existing instruments is the reliance on a scale to summarize their findings. Empirical tests of the validity and reliability of these scales suggest the need for critical analysis of individual bias components rather than dependence solely on checklists and scales. Juni et al. noted dramatically different results in meta-analyses when different quality rating scales were used.[23] For observational studies in particular, mechanical scoring of items on a checklist that focuses on quality of reporting, ethical issues, background, rationale, and so on, will fail to assess the critical question: whether the outcomes can be attributed to the effects of the intervention.[24]

As noted above, available empirical evidence cautions against merely applying weights to observational studies based on scores on a bias checklist.[24-26] Rather, the risk of bias requires interpretation based on an understanding of the content of the topic, particularly when evaluating studies in a heterogeneous body of observational evidence (e.g., differently defined interventions and outcomes, choice of analytic methods). Existing guidance offers limited assistance with interpretation of risk of bias in observational studies, particularly with regard to confounding. "The Cochrane Handbook of Systematic Reviews,"[9] AHRQ guidance,[7] and the Institute of Medicine (IOM) standards for systematic review[8] all detail reasons to include observational studies for harms and attendant risks and to monitor longer-term outcomes.[27] These documents caution that the risk of bias will always be greater for observational studies than for RCTs. All offer general guidance on sources of bias for observational studies.

Both the IOM[8] and AHRQ guidance[28] discuss the role for plausible confounding (confounding not controlled for in a study that inflates the observed effect) to increase the strength of evidence in a rating system in which observational studies start out with a lower grade, but offer no framework to consider how to evaluate the risk of confounding. Cochrane guidance notes that issues of confounding cannot be easily addressed within existing instruments evaluating the risk of bias in individual studies in isolation and suggests developing summary tables to identify prespecified confounders and variables controlled for in the analysis for each study. This information is intended to illustrate the extent of heterogeneity in the literature. Although the guidance does not require that all Cochrane reviews that include observational studies develop these tables, reviewers need to demonstrate that they have considered the role of residual confounding (confounding not controlled for in the design or analysis of the study) in explaining the findings from nonrandomized studies.[9]

Earlier conceptual and empirical work described above suggests the following needs for evaluating causality in observational studies: (1) consensus around potential sources of bias for different observational study designs,[1,6,19] (2) content-specific criteria to rate risk of bias,[1,6,19] (3)

use of risk of bias criteria to understand heterogeneity in results across individual studies rather than merely to weight pooled estimates of effect,[24-26] and (4) a framework to adequately capture risk of bias and confounding concerns that are specific to a body of observational studies included in a systematic review.

Earlier Phases of Development

Our own previous work developed a tool consisting of specific questions that reviewers can use for identifying context-specific sources of biases and confounding in observational studies[29] but stopped short of offering guidance on a larger framework for its use. In this earlier project, we developed the tool through first, identifying a large list of questions that had been used previously in EPC systematic reviews or other instruments identified by the research team and a Technical Expert Panel. A resulting list of 1,492 questions was culled to eliminate duplicates and minor wording differences and refined through face, cognitive, and content validity and inter-rater reliability testing. Based on this process, the RTI Item Bank included 29 questions which sought to comprehensively capture the risk of bias and precision domains critical for evaluating observational studies. Precision was included in the bank to allow reviewers to create a tool to evaluate what has been traditionally been referred to as the "quality" of an individual study, inclusion of both risk of bias and precision. The bank also included directions for adapting questions to a specific topic area and assistance in selecting questions relevant for particular study designs (e.g., case series, case-control, cohort, and cross-sectional).[29] The next step in developing this tool required consensus around which specific questions are *essential* for evaluating a study of each design, and further refinement of the questions and related instructions.

Project Objectives

The project's objectives, in response to the needs described above, included creating a framework for evaluating the risk of bias and confounding in individual studies included in a body of observational study evidence, developing consensus around sources of bias for different observational study designs, and making enhancements to the RTI Item Bank. We determined, based on discussions with our expert Working Group, that adequate evaluation of the risk of bias and confounding in observational studies included in a systematic review could not be accomplished solely within the confines of questions included in an item bank evaluating individual studies in isolation. Therefore the project's activities expanded to more adequately accommodate the project's goal. Tasks included: (1) developing a process framework to assess the effect of confounding across the body of observational study evidence as well as within individual studies, (2) identifying critical sources of bias and confounding most common to each observational study design type, (3) refining and reducing the set of "core" questions that would be necessary for evaluating risk of bias and confounding concerns for each design, and (4) refining the instructions provided to users to improve clarity and usefulness so that the RTI Item Bank is a easily accessible and practical tool that can be used across EPCs and other organizations conducting systematic reviews.

Determining, from the questions available within the item bank, the best set of questions to use for studies with specific design features and subject topics requires sufficient epidemiological expertise to classify study designs, experience conducting systematic reviews, and familiarity with risk of bias rating. However, even experienced researchers show poor inter-rater reliability in classifying study designs.[30] To facilitate optimal use of the item bank, further

development was needed to provide guidance to users in appropriately identifying the study designs included in their reviews and the possible bias concerns specific to those designs (sources of bias referred to above). Users could then use the item bank to create an instrument to capture the most likely risk of bias concerns in their subject area (content-specific criteria referred to above). We believe these enhancements will improve the practicality and user-friendliness of instruments created from the item bank and may help promote their inter-rater reliability.

Methods

We convened a Working Group consisting of six systematic review experts, epidemiologists, and trialists and sought their input over the course of three planned conference calls and related email exchanges. When we had less than full attendance for conference calls, we arranged alternate calls for those who could not make the original date. Working Group members responded to call notes through numerous electronic interactions between conference calls.

We asked the Working Group members to comment on the goals and activities of the project. Their input resulted in an expansion of our original objective from further refinement of the RTI Item Bank to also include the development of a process framework to consider the effect of confounding across the body of observational study evidence.

As a precursor to the project, we reviewed the evidence on approaches to assessing the risk of bias in studies to understand how sources of bias might differ between RCTs and observational studies (Appendix A). A previously developed taxonomy of observational studies (Appendix B)[30,31] offers an approach to grouping studies of similar designs, or with similar design features that may relate to bias. This characterization of study design features can be used by systematic reviewers to guide the choice of questions needed for risk of bias assessments of different observational study designs. Studies with different designs, or with different design features, may require (some) different questions for risk of bias assessments. For example, studies identified by the taxonomy as "noncomparative" include case reports and case series that have no comparison group. Therefore, questions to assess risk of bias must be selected with this in mind (e.g., not ask about comparability between groups). Another example of an important study design feature is whether both intervention/exposure and outcome assessment were prospective. When exposure/intervention status were identified retrospectively, there may be concerns about recall bias and misclassification that may be less worrisome in studies that collect and classify exposure information prospectively, particularly if they use standard definitions or criteria.

The study team reviewed the item bank to identify a core set of questions that were needed to evaluate the risk of confounding, bias, and lack of precision of individual studies for a specific, limited set of commonly used observational study designs (case series, case-control, cohort, and cross-sectional).

We revised the bank to consolidate questions addressing the same bias concerns (e.g., use of valid and reliable measures) and separated questions that were likely to be unnecessary because they were limited to study reporting, were redundant, or based on discussions with the Working Group, were considered not relevant for evaluating risk of bias or precision.

Based on the revised item bank, members of the Working Group were asked to rank their choice of specific questions that needed to be included in the item bank to evaluate the risk of bias, confounding, and lack of precision of an observational study for each of the four design types and the subset of questions that would only be relevant for an observational study of a particular design (case series, case-control, cohort, or cross-sectional). Working Group members were provided with the revised version of the bank, which included 16 of the original 29 questions. They were asked to evaluate the importance of each question using a five-point scale that included very important, somewhat important, a little important, not at all important, and not applicable/exclude. They were also asked if they agreed with the study team's recommendation to eliminate each of 10 questions. (One question had been eliminated based on earlier discussions with the Working Group concerning not needing a question establishing whether a study was prospective, retrospective, or mixed [independent of the conduct of the study] and two

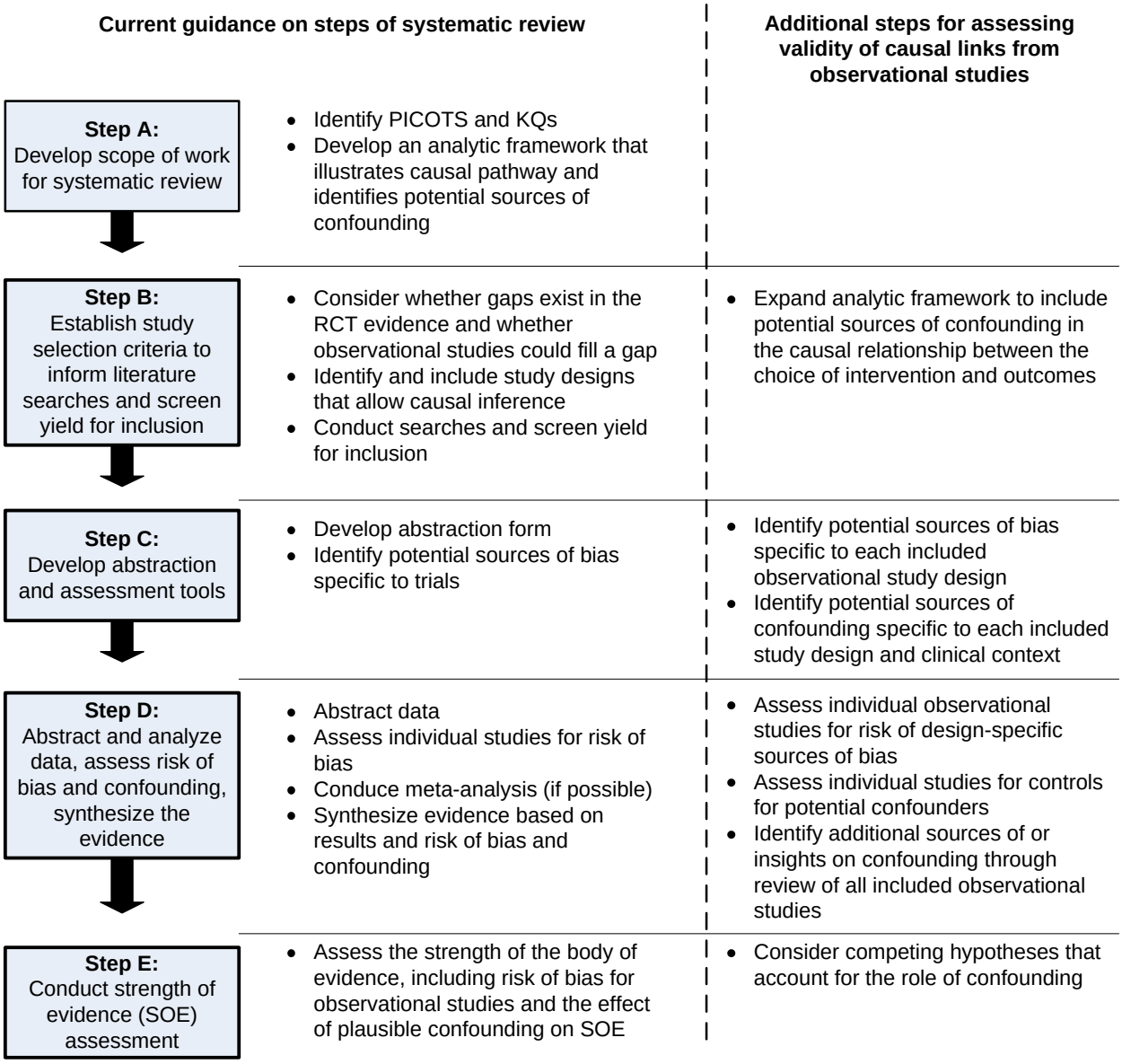
were eliminated based on question consolidations.) We also sought additional comments from Working Group members on each question concerning readability. We had intended to conduct a modified Delphi process with two rounds of voting, but terminated the process with a single round of voting because of a poor response rate (three of six participants voted).

This document was revised in response to peer review and public comment.

Results

In consultation with our Working Group, we developed a framework to assist in the evaluation of potential confounding in observational studies in systematic reviews (Figure 1). This framework builds on existing routine processes of systematic reviews: developing a scope of work, establishing study selection criteria to inform literature searches and selection of studies for abstraction, developing abstraction and assessment tools, abstracting and analyzing data, assessing the risk of bias and confounding, synthesizing the evidence, and conducting strength of evidence (SOE) assessment. In Figure 1, the central column describes routine tasks required for all systematic reviews in Steps A through E. Methodological guidance is available at length elsewhere on these steps and is not repeated here.[5,28,32-35] The inclusion of observational studies requires additional tasks, highlighted in the right-hand column of the figure and described in detail below. The process described below is simplified to apply to a single outcome. Reviews that evaluate multiple outcomes may need to replicate these steps for each outcome.

Figure 1. Process framework assessing the validity of causal links for observational study evidence in systematic reviews



Expand the Analytic Framework To Consider Sources of Confounding

The Working Group noted the importance of specifying possible sources of confounding at the start of the review (Step A). As shown in Step B, the analytic framework should clearly show potential confounders that could affect the relationship between the choice of intervention and the outcome. In the likely absence of empirical evidence that specific confounders altered estimates of effect, systematic reviewers may need to rely on hypothesized relationships to lay out potential relationships between the confounder, the intervention, and the outcome.[19] Not all

confounders need to be controlled in every study and the use of causal diagrams may help explicate which sets of confounders could alter the effect estimate in an individual study.[36]

The specific sources of confounding will vary based on the clinical context. Some of these sources of confounding can be controlled for in the design or analysis of the study. The potential for uncontrolled confounding in an overall body of evidence may influence what systematic reviewers judge to be admissible study designs. Current EPC guidance on the inclusion of observational studies suggests that reviewers consider two questions: (1) Are there gaps in the evidence from RCTs? (2) Will observational studies provide valid and useful information?[7]

Norris et al. suggest considering the clinical context and natural history of a condition when judging whether observational studies are likely to produce estimates that are too biased to provide valid and useful information.[7] With fluctuating conditions, for instance, patients may improve spontaneously, making it difficult to attribute improvements to treatments through evaluations based on observational studies. Norris et al. note the importance of selection bias and confounding by indication (confounding that results from the patient having a specific indication that influences both the selection of the intervention and the outcome) as particular problems plaguing observational studies but do not offer thresholds for when anticipated concerns from bias would preclude the consideration of these studies. We believe, as do other commentators, that one standard cannot be established across all clinical areas and each review requires an understanding of the threats to validity for that topic.

Identify the Source of Bias and Confounding Specific to Each Included Design

In Step C of the process framework, prior to the review of individual studies, systematic reviewers would consider the observational study designs that could potentially be included in the review, and risks of bias and confounding that are likely to be a concern in the body of evidence.

Based on discussions with the Working Group, we changed the orientation of some of the questions in the item bank so that all of the questions now focus on whether a solution to a bias concern in a study was adequate. This shifts the burden for the reviewer from determining whether the study was "good" (lower risk of bias) to identifying salient problems with the study's approach (higher risk of bias). For example, one question was changed from "Did the study apply inclusion/exclusion criteria uniformly to all comparison groups/arms of the study?" to "Do the inclusion/exclusion criteria vary across the comparison groups of the study?"

Also based on discussions with the Working Group, we eliminated one question from the item bank ("Is the study design prospective, retrospective, or mixed?") prior to conducting the voting exercise. Working Group members found this design classification to be problematic and uninformative because it does not indicate whether an outcome is assessed according to predictable criteria or whether it was specified in a protocol.

In addition to group discussions, three of the six members of the Working Group participated in one round of testing of the instrument and one member provided comments on item bank questions outside the formal exercise. Table 1 presents the results of the testing and identifies questions that were considered very important or somewhat important by at least two of the three participants, with the exception of questions on precision because precision (the absence of random error) is independent of risk of bias (systematic error). Of the 16 questions, 5 were not considered important or very important across study designs by at least two of the three Working Group members. However, because of the small number of participants, we do not recommend

deleting the five questions and instead note Working Group members' concerns and comments on the item bank questions.

One of the Working Group members considered Question 4, concerning whether the study failed to account for variation in execution from the proposed protocol, as unlikely to be reliable because the information was rarely available and also thought that the question did not distinguish between major and minor variations from the protocol. Questions 9 and 10, concerning outcomes missing from the results (benefits and harms), were considered evidence of reporting bias and so outside the scope of the risk of bias assessment. One Working Group member thought that Question 11, concerning whether results were believable after taking study limitations into consideration, expressed weak logic that would result in unexpected findings being discounted. Lastly, several Working Group members expressed concerns about Question 14, which relates to precision. One Working Group member thought that a post hoc power calculation to determine the sufficiency of sample size was no longer needed and that reviewers could instead adopt the GRADE Working Group's guidance on optimal information size. GRADE guidance states that a study (body of evidence) is precise if it includes more than 400 events.[37]

By type of study design, the resulting item bank (see Table 1) includes the following number of questions considered very or somewhat important by the subset of Working Group members, out of the full bank of 16 questions: case series (8), case-control (8), cohort (12), and cross-sectional (10).

Working Group members favored retaining two questions that measured confounding in individual studies: the use of valid and reliable measures of confounding (all study designs) and important confounding variables taken into account in the design and/or analysis (all designs except case series; Table 2).

The four Working Group members who participated in the instrument testing agreed with the study team's decision to exclude additional questions that had been included in an earlier version of the item bank.[29] The excluded questions concerned: the quality of the reporting in the article, such as the description of the inclusion criteria and the intervention because we retained questions that were designed to more directly measure bias; funding source because a clear relationship between source of funding and risk of bias has not been established; applicability, such as the length of followup, because this is not related to internal validity; and redundancies. Questions concerning the validity and/or reliability of various measures were combined (inclusion/exclusion, interventions, outcomes, and confounders), as were questions concerning attrition, which were also reframed as concerns about missing data.

Table 1. Questions from the item bank Working Group voting participants[a] considered very or somewhat important for evaluating risk of bias and precision by observational study design type

Item Bank		Case Series	Case-Control	Cohort	Cross-sectional
Questions to assess the risk of bias					
Q1	Do the inclusion/exclusion criteria vary across the comparison groups of the study?	X[b]	X	X	X
Q2	Does the strategy for recruiting participants into the study differ across groups?	X[b]	X	X	X
Q3	Is the selection of the comparison group inappropriate, after taking into account feasibility and ethical considerations?		X	X	X
Q4	Does the study fail to account for variations in the execution of the study from the proposed protocol?				
Q5	Was the outcome assessor not blinded to the intervention or exposure status of participants?	.		X	X
Q6	Were valid and reliable measures, implemented consistently across all study participants used to assess inclusion/exclusion criteria, intervention/exposure outcomes, participant health benefits and harms, and confounding?	X	X	X	X
Q7	Was the length of followup different across study groups?	X[b]		X	
Q8	In cases of high loss to followup (or differential loss to followup), was the impact assessed (e.g., through sensitivity analysis or other adjustment method)?			X	
Q9	Are any important primary outcomes missing from the results?				
Q10	Are any important harms or adverse events that may be a consequence of the intervention/exposure missing from the results?				
Q11	Are results believable taking study limitations into consideration?				
Questions to assess confounding					
Q6	Were valid and reliable measures, implemented consistently across all study participants used to assess inclusion/exclusion criteria, intervention/exposure outcomes, participant health benefits and harms, and confounding?	X	X	X	X
Q12	Any attempt to balance the allocation between the groups or match groups (e.g., through stratification, matching, propensity scores)?	X	X	X	X
Q13	Were the important confounding variables taken into account in the design and/or analysis (e.g., through matching, stratification, interaction terms, multivariate analysis, or other statistical adjustment such as instrumental variables)?		X	X	X
Questions to assess precision					
Q14	Is the sample size insufficiently large to detect a clinically significant difference of 5% or more between groups in at least one primary outcome measure?				
Q15	Are the statistical methods used to assess the primary benefit outcomes inadequate?	X	X	X	X
Q16	Are the statistical methods used to assess the main harm or adverse event outcomes inadequate?	X	X	X	X

[a]At least two out of three respondents.
[b]Questions can be answered for case series without a uniform set of eligibility criteria, recruitment strategy, or length of followup for all participants.

Table 2. Questions and instructions in revised RTI Item Bank following voting exercise by category of bias or precision assessed

	Questions From Item Bank [Instructions for principal investigator (PI) and/or abstractor]	Type of Bias Assessed
Q1	**Do the inclusion/exclusion criteria vary across the comparison groups of the study?** [*PI: Drop question if not relevant to all included studies. To use this question for studies with one group, the focus of the question on comparison groups and related response categories would need to be changed to individuals.]*	Selection bias
Q2	**Does the strategy for recruiting participants into the study differ across groups?** *[PI: Drop question if not relevant to all included studies. If the recruitment strategy results in pre-intervention differences in prognostic factors that could explain the selection of the intervention and the outcome, confounding can occur. If the strategy results in the selective and differential inclusion of patients (such as prevalent rather than new users), selection bias can occur. To use this question for studies with one group, the focus of the question on comparison groups and related response categories would need to be changed to individuals.]*	Selection bias confounding
Q3	**Is the selection of the comparison group inappropriate?** *[PI: Provide instruction to the abstractor based on the type of study. Interventions with community components are likely to have contamination if all groups are drawn from the same community. Interventions without community components should select groups from the same source (e.g., community or hospital) to reduce baseline differences across groups. For case-control studies, controls should represent the population from which cases arose; that is, controls should have met the case definition if they had the outcome.]*	Selection bias, confounding
Q4	**Does the study fail to account for important variations in the execution of the study from the proposed protocol?** *[PI: Consider intensity, duration, frequency, route, setting, and timing of intervention/exposures. Drop if not relevant for body of literature.]*	Performance bias
Q5	**Was the assessor not blinded to the outcome, exposure, or intervention status of the participants?** *[PI: Clinical assessors may not always be blinded to exposure/intervention as well as outcome status. For studies where patients are selected based on outcome (e.g., case-control), blinding to exposure or intervention status is particularly important. For designs where patients are selected based on exposure status (e.g., cohorts), blinding to outcomes is particularly important. Drop if not relevant to the body of literature.]*	Detection bias
Q6	**Were valid and reliable measures not used or not implemented consistently across all study participants to assess inclusion/exclusion criteria, intervention/exposure outcomes, participant benefits and harms, and potential confounders?** *[PI: Important measures should be identified for abstractors and if there is more than one, they should be listed separately. PI may need to establish a threshold for what would constitute acceptable measures based on study topic. When subjective or objective measures could be collected, the PI will need to consider if subjective measures based on self-report should be considered as being less reliable and valid than objective measures such as clinical reports and lab findings. Some characteristics may require that sources for establishing their validity and/or reliability be described or referenced. If so, provide instruction to abstractors.]*	Detection bias, confounding
Q7	**Was the length of followup different across study groups?** *[Abstractor: When followup was the same for all study participants, the answer is no. If different lengths of followup were adjusted by statistical techniques, (e.g., survival analysis), the answer is no. Studies in which differences in followup were ignored should be answered yes.]*	Attrition bias

Table 2. Questions and instructions in revised RTI Item Bank following voting exercise by category of bias or precision assessed (continued)

	Questions from Item Bank *[Instructions for principal investigator (PI) and/or abstractor]*	**Type of Bias Assessed**
Q8	**In cases of missing data (e.g., overall or differential loss to followup for cohort studies or missing exposure data for case-control studies), was the impact not assessed (e.g., through sensitivity analysis or other adjustment method)?** *[PI: For cohort studies, attrition is measured in relation to the time between baseline (allocation in some instances) and outcome measurement for both retrospective and prospective studies and could include data loss from switching. Attrition rates may vary by outcome and time of measurement. Specify the criterion to meet relevant standards for the topic. Specify measurement period of interest, if repeated measures. For case-control studies, evaluate missing data in relation to exposure status.]*	Attrition bias, detection bias
Q9	**Are any important primary outcomes missing from the results?** *[PI: Identify all primary outcomes that one would expect to be reported in the study, including timing of measurement.]*	Selective outcome reporting
Q10	**Are any important harms or adverse events that may be a consequence of the intervention/exposure missing from the results?** *[PI: Identify all important harms that one would expect be reported in the study, including timing of measurement. Drop if not relevant to body of literature.]*	Selective outcome reporting
Q11	**Did the study fail to balance the allocation between the groups or match groups (e.g., through stratification, matching, propensity scores)?** *[PI: Drop if not relevant to the body of evidence.]*	Confounding
Q12	**Were important confounding variables not taken into account in the design and/or analysis (e.g., through matching, stratification, interaction terms, multivariate analysis, or other statistical adjustment such as instrumental variables)?** *[PI: Provide instruction to abstractors on known confounding variables and inadequate adjustment for confounding for each outcome.]*	Confounding
Q13	**Are results believable taking study limitations into consideration?** *[Abstractor: This question is intended to capture the overall quality of the study. Consider issues that may limit your ability to interpret the results of the study. Review responses to earlier questions for specific criteria.]*	Overall assessment

Assess Individual Studies for Risk of Bias and Potential Confounders

In Step D, the observational studies included in the body of evidence are reviewed to determine their risk of bias. Reviewers must also identify confounders that are controlled in specific study analyses and determine if additional confounders were identified by study authors. In an earlier version of the item bank, we developed detailed instructions to assist reviewers. During this iteration, we made minor edits to the instructions, based on comments from Working Group members, peer reviewers and team members' improvements in readability. Table 2 presents the final included questions, revised detailed instructions for each question, and the type of bias that each question is designed to uncover. Appendix C provides the instrument in detail, along with response categories. As noted above, we have eliminated from the item bank the questions designed to evaluate precision.

Consider Competing Hypotheses for Observational Studies

In Step E, the usual culmination of the systematic review is the assessment of the strength of the body of evidence. The dominant approaches to conducting this step of the systematic review comes from the GRADE Working Group and related systems, such as the AHRQ approach to

grading the strength of evidence.[28] These systems aim to offer a systematic and explicit approach to making judgments about the evidence: systematic reviewers provide an assessment of their confidence about how close the true effect is to the observed effect (strong, moderate, low, very low confidence/insufficient evidence to make a determination)[10] by considering study limitations, indirectness, imprecision, inconsistency, and publication/reporting bias. The body of evidence from trials generally begins with a high confidence and then is marked down for various flaws; nonrandomized/observational study bodies of evidence start with low confidence and then can be marked up if they demonstrate characteristics that would increase the reviewers' confidence in the findings from these designs such as a large magnitude of effect, dose-response relationship, or plausible confounding that would have otherwise weakened the effect estimate.[38,39]

These criteria (limitations of the body of evidence, indirectness, imprecision, inconsistency, and publication/reporting bias) draw from a long tradition of using causal criteria to evaluate the validity of causal links, dating back to the 1964 report of the Advisory Committee to the Surgeon General on the association between smoking and lung cancer,[40] a 1965 keynote address by Bradford Hill,[41] and a 1973 textbook by Susser.[42] Bradford Hill presented nine causal criteria ("viewpoints") to consider when interpreting an association as causal in nature. These included: strength (Is the strength of the association large enough that it would be difficult to explain by other factors?), consistency (Is the association observed by different persons, in different circumstances, and at different times?), specificity (Does a specific exposure lead to a specific outcome?), temporality (Does the exposure precede the outcome?), biological gradient (Does a dose-response curve exist?), plausibility (Is the association biologically plausible?), coherence (Does the observed association fit in with what is known about the natural history of the disease?), experiment (Do experiments with the exposure alter the frequency of the outcome?), and analogy (Do other examples exist that support a similar association?).[41]

Schünemann et al. note that the GRADE approach to assessing the quality of the evidence and strength of recommendations draws both implicitly and explicitly from numerous Bradford Hill criteria for establishing causation. They also postulate that specificity, as defined by Bradford Hill, is not an important criterion for evaluating the effects of interventions[43] (because a single cause can result in multiple effects and vice versa).[b] With regard to how concerns about threats to causality are interpreted for observational study bodies of evidence, both GRADE and the AHRQ EPC approach start observational study evidence at a lower level than trials, thus acknowledging the experimental evidence criterion; they initially downgrade designs that do not address temporality (either through randomization or through concurrent control groups). They do so overall, and make no distinction between study evaluated intended benefit and unintended harm. Strength of evidence assessments require critical consideration of confounders, factors that are more likely to plague observational studies.[c] The primary intent of this approach is to examine whether or not a particular intervention works, or the extent to which it works. Plausible confounding is interpreted as "reverse confounding"; that is, in the absence of confounding, the effect would have been even stronger. This approach is not intended to evaluate alternative reasons for an observed phenomenon: plausible confounding, in this context, serves to raise confidence in the observational data rather than suggest another explanation for the data. Given

[b] Others suggest that specificity does play an important role in causal inference because the construct of an effect having a single cause can be used to distinguish causal relationships from noncausal associations.[44]

[c] Schünemann et al. note that the GRADE system separates judgment about the quality of evidence from the strength of recommendations. Low or moderate quality evidence does not preclude a recommendation for or against an intervention. The implication is that observational studies can still lead to recommendations and action.[43]

the likelihood of multiple observed and unobserved confounders in observational data, the review of evidence on a single hypothesis is unlikely to definitively prove causality. Systematic reviews of observational studies must go beyond grading the evidence to explicitly assess competing hypotheses if they seek to understand causality.

Criticisms of meta-analyses of observational studies point to the need for critical analysis that focuses on study heterogeneity.[26,45] One such criticism relates to the overinterpretation of confidence intervals that exclude the null value. Researchers point out that narrow confidence intervals could arise from spurious precision[26] and represent only one type of uncertainty.[45] Another criticism relates to overreliance of the consistency criterion when evaluating the strength of evidence: MacLure et al. note that the "consistency criterion corroborates the causal hypothesis of interest only indirectly by refuting confounders and biases that differ across studies. It does not refute confounders and biases that recur consistently in many studies."[45, p. 347] Thus a single study that accounts for confounding may be more believable than the pooled estimate from numerous studies that fail to account for confounding.[46]

Our proposed approach, illustrated in Figures 2 and 3 builds on the strength of evidence assessment to evaluate multiple competing hypotheses as a final step of the analysis. An approach such as this one shows that risks of this additional critical appraisal include incompleteness (no review can completely rule out alternative hypotheses) and complexity (multiple pathways to the outcome).[45,47] Figure 2 offers an example of how confounders can be catalogued for each study and Figure 3 illustrates how the synthesis can incorporate competing hypotheses.

Figure 2. Example of systematic review listing potential confounders when evaluating the effect of cesarean delivery on maternal request on maternal and neonatal outcomes

Table 16. Inclusion of possible confounders

	Nulliparous Only	Includes Preterm	Includes Previa	Includes Repeat Cesarean Delivery	Includes Multiple Gestations
Hannah et al., 2000[20]	No	No	No	Yes	No
Hannah et al., 2002[18]	No	No	No	Yes	No
Hannah et al., 2004[83]	No	No	No	Yes	No
Leiberman et al., 1995[84]	Yes	No	Yes	No	No
Badawi et al., 1998[37]	No	No	Yes	Yes	Unspecified
Bergholt et al., 2003[38]	No	Yes	Yes	Yes	Yes
Burrows et al., 2004[39]	No	No	Unspecified	Yes	No
Dessole et al., 2004[40]	No	Yes	Yes	Yes	Yes
Farrell et al., 2001[41]	Yes	Probably	Unspecified	No	Unspecified
Farrell et al., 2001[42]	Yes	Probably	Unspecified	No	Unspecified

Source: Viswanathan M, Visco AG, Hartmann K, Wechter, ME, Gartlehner G, Wu JM, Palmieri R, Funk MJ, Lux, LJ, Swinson T, Lohr KN. Cesarean Delivery on Maternal Request. Evidence Report/Technology Assessment No. 133. (Prepared by the RTI International-University of North Carolina Evidence-Based Practice Center under Contract No. 290-02-0016.) AHRQ Publication No. 06-E009. Rockville, MD: Agency for Healthcare Research and Quality. March 2006. Page 61.

Figure 3. Example of use of information on confounders in synthesis

[The report graded the evidence on the effect of cesarean delivery on maternal request on neonatal mortality as weak and then noted that only one] "study used a large administrative data set that offered a sample size sufficient to examine rare outcomes, but its retrospective classification of mode of delivery limited its usefulness. For instance, the classification of the cesarean deliveries was limited to either "labored" or "unlabored." The unlabored cesarean deliveries likely included emergency cesareans and those performed for serious maternal and neonatal indications such as placenta previa, severe preeclampsia, breech presentation, fetal distress, and major fetal anomalies. Such maternal and neonatal disorders could seriously affect neonatal mortality and seriously confound the underlying association between neonatal mortality and mode of delivery."

Source: Viswanathan M, Visco AG, Hartmann K, Wechter, ME, Gartlehner G, Wu JM, Palmieri R, Funk MJ, Lux, LJ, Swinson T, Lohr KN. Cesarean Delivery on Maternal Request. Evidence Report/Technology Assessment No. 133. (Prepared by the RTI International-University of North Carolina Evidence-Based Practice Center under Contract No. 290-02-0016.) AHRQ Publication No. 06-E009. Rockville, MD: Agency for Healthcare Research and Quality. March 2006. Page 120-121.

Discussion

Our tool was initially based on earlier instruments, expert panel input, cognitive testing, and face validity testing. This project builds on our earlier work by revising the questions, providing further instructions to users, and proposing a framework for considering confounding across a body of observational study evidence.

Specifically, we made further modifications and enhancements to the RTI Item Bank, including: eliminating an item asking reviewers to designate a study as prospective or retrospective because it was considered uninformative, eliminating items evaluating precision because it is generally no longer coupled with risk of bias assessment, modifying item bank questions so that they all focus on identifying study implementation characteristics that would increase the risk of bias, and enhancing the instructions provided to reviewers who use the tool.

In addition, we propose a substantial expansion in the critical appraisal of confounding when systematic reviews include observational studies for evaluation of benefits of interventions or their harms. Attributing causality to interventions from such evidence requires prespecification of anticipated sources of confounding prior to the review, followed by appraisal of potential confounders at three levels: outcomes, studies, and the body of evidence. During the review, additional insights may emerge. These tasks require a substantial commitment of senior researchers on each systematic review team who are deeply knowledgeable about the topic to conduct tasks that may have sometimes been previously assigned to junior team members. This expanded focus on confounding serves two purposes: first, it helps to explain results when heterogeneous bodies of evidence turn up conflicting results; and second, it helps to undergird the validity of the results of the systematic review, regardless of the level of heterogeneity. The latter may be particularly important in instances when a body of evidence may be relatively homogeneous in showing an effect but individual studies all fail to account for common confounders.

Establishing consensus requires collaboration across a wide range of expertise. This methods project benefited from a diverse set of Working Group members with varying experience, expertise, and perspectives. The Working Group members were particularly engaged in discussions on how to think about the issue of confounding over the course of a systematic review. Our project suffered, however, from the format we employed to engage Working Group members in revising questions for the RTI Item Bank: the length, frequency, and format of communication (telephone and email) did not permit a sustained exchange of ideas. We believe that the intensity of this effort requires one or more in-person meetings to establish consensus on needed items for specific research designs.

Recommendations for Future Methods Development

From a practical perspective, this proposed approach to evaluating confounding raises questions about burden, reliability, and validity. Future efforts should include tests of reliability, including comparing the tool to other approaches, in addition to assessment of the time and effort involved in assessing sources of confounding. Most importantly, the added value of the expanded effort to evaluate confounding needs to be established. Does this effort produce more actionable evidence or a shift in conclusions from the review?

This project is a pilot effort to provide an approach for identifying confounding relevant to various observational study designs. The project also produced user enhancements to the RTI Item Bank by adding assistance in identifying study designs and streamlining the number of

required questions necessary to evaluate risk of bias. While we were able to identify a set of questions for the most common observational study designs, we were not able to establish consensus on required items for each type of design, as we had originally intended. Whether our suggested approach applies to other quasi-experimental designs such as controlled clinical trials or pre-post studies of public health interventions also requires empirical assessment. More work is required to establish consensus on type of study designs and specific sources of bias for each design.

References

1. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22:189-212. PMID: 11274518.
2. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes (Lond). 2008 Aug;32 Suppl 3:S8-14. PMID: 18695657.
3. Hernán MA, Robins JM. Casual inference. Harvard; 2012.
4. Norris SL, Atkins D. Challenges in using nonrandomized studies in systematic reviews of treatment interventions. Ann Intern Med. 2005 Jun 21;142(12 Pt 2):1112-9. PMID: 15968036.
5. 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.
6. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed., Philadelphia, PA: Lippincott, Williams, & Wilkins; 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. Eden J, Levit L, Berg A, et al., eds. Finding What Works in Health Care: Standards for Systematic Reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2011.
9. Reeves BC, Deeks JJ, Higgins JPT, et al. Chapter 13: Including non-randomized studies. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration; 2011.
10. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. PMID: 21208779.
11. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25. PMID: 15308962.
12. Etminan M, Sadatsafavi M, Jafari S, et al. Acetaminophen use and the risk of asthma in children and adults: a systematic review and metaanalysis. Chest. 2009 Nov;136(5):1316-23. PMID: 19696122.
13. Lowe AJ, Carlin JB, Bennett CM, et al. Paracetamol use in early life and asthma: prospective birth cohort study. Br Med J. 2010 Sep 15;341:c4616. PMID: 20843914.
14. Vandenbroucke JP. Why do the results of randomised and observational studies differ? BMJ. 2011;343:d7020. PMID: 22065658.
15. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000 Jun 22;342(25):1887-92. PMID: 10861325.
16. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess. 2000;4(34):1-154. PMID: 11134917.
17. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001 Aug 15;286(7):821-30. PMID: 11497536.
18. Odgaard-Jensen J, Vist GE, Timmer A, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev. 2011(4):MR000012. PMID: 21491415.
19. Shrier I, Boivin JF, 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 Nov 15;166(10):1203-9. PMID: 17712019.

20. 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.

21. Mhaskar R, Djulbegovic B, Magazin A, et al. Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. J Clin Epidemiol. 2012 Jun;65(6):602-9. PMID: 22424985.

22. 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. J Clin Epidemiol. 2008 Apr;61(4):344-9. PMID: 18313558.

23. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ. 2001 Jul 7;323(7303):42-6. PMID: 11440947.

24. Colliver JA, Kucera K, Verhulst SJ. Meta-analysis of quasi-experimental research: are systematic narrative reviews indicated? Med Educ. 2008 Sep;42(9):858-65. PMID: 18715482.

25. Blair A, Burg J, Foran J, et al. Guidelines for application of meta-analysis in environmental epidemiology. ISLI Risk Science Institute. Regul Toxicol Pharmacol. 1995 Oct;22(2):189-97. PMID: 8577954.

26. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. BMJ. 1998 Jan 10;316(7125):140-4. PMID: 9462324.

27. Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 12-EHC047-EF. Rockville, MD: March 2012. www.effectivehealthcare.ahrq.gov/ehc/products/322/998/MethodsGuideforCERs_Viswanathan_IndividualStudies.pdf

28. 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. PMID: 19595577.

29. 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. PMID: 21959223.

30. Hartling L, Bond K, Santaguida PL, et al. Testing a tool for the classification of study designs in systematic reviews of interventions and exposures showed moderate reliability and low accuracy. J Clin Epidemiol. 2011 Aug;64(8):861-71. PMID: 21531537.

31. 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 (Prepared by University of Alberta Evidence-based Practice Center under Contract No. 290-002-0023). AHRQ Publication No. 11-EHC007-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2010.

32. Helfand M. AHRQ series editorial: public involvement improves methods development in comparative effectiveness reviews. J Clin Epidemiol. 2010 May;63(5):471-3. PMID: 20346860.

33. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):484-90. PMID: 19716268.

34. 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 May;63(5):481-3. PMID: 18834715.

35. 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.

36. Shrier I. Structural Approach to Bias in Meta-analyses. Research Synthesis Methods. 2011;2(4):223-37.

37. 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.

38. 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.

39. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011 Apr;64(4):383-94. PMID: 21195583.

40. United States. Surgeon General's Advisory Committee on Smoking and Health. Smoking and health report of the advisory committee to the Surgeon General of the Public Health Service. Public Health Service publication no 1103. Washington, D.C.: U.S. Dept. of Health, Education, and Welfare, Public Health Service; 1964. p. xvii, 387 p. ill. 24 cm.

41. Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med. 1965 May;58:295-300. PMID: 14283879.

42. Susser M, Andjelkovich D. Causal thinking in the health sciences; concepts and strategies of epidemiology. New York: Oxford University Press; 1973.

43. Schünemann H, Hill S, Guyatt G, et al. The GRADE approach and Bradford Hill's criteria for causation. J Epidemiol Community Health. 2011 May;65(5):392-5. PMID: 20947872.

44. Weiss NS. Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology. 2002 Jan;13(1):6-8. PMID: 11805580.

45. Maclure M. Demonstration of deductive meta-analysis: ethanol intake and risk of myocardial infarction. Epidemiol Rev. 1993;15(2):328-51. PMID: 8174661.

46. Vandenbroucke JP. Statins and infections: sloppy causal thinking in meta-analyses of observational research. BMJ. 2012 1 February.

47. Cornfield J, Haenszel W, Hammond EC, et al. Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst. 1959 Jan;22(1):173-203. PMID: 13621204.

Appendix A. Approaches to Assessing the Risk of Bias in Studies

Approaches to critical appraisal of study methodology and related terminology has varied and is evolving. Overlapping terms include quality, internal validity, risk of bias, or study limitations, but a central goal is an assessment of the validity of the findings. We use the phrase "assessment of risk of bias" as the most representative of the goal of evaluating the degree to which the effects reported by a study represent the "true" causal relationship between exposure and outcome.

A valid estimate requires the absence of bias or systematic error in selection, performance, detection, measurement, attrition, and reporting and adequacy in addressing potential confounding. The interpretation of the effect of an estimate also requires the evaluation of precision (the absence of random error through adequate study size and study efficiency).[1] For studies that do not lend themselves to quantitative pooling of estimates, reviewers will likely be making assessments of individual study risk of bias and precision at the same time, supporting the broader notion of evaluating study quality. Thorough assessment of these threats to the validity of an estimate is critical to understanding the believability and interpretation of a study.

Table 1 builds on and revises, drawing on numerous sources,[1-3] the Cochrane Collaboration taxonomy[4] of threats to validity and precision to expand the discussion of confounding and selection bias.

The inclusion of observational studies considerably expands the challenges in establishing causal inference in systematic reviews. Observational studies cannot, by design, offer establish causality through features such as randomization and concealment of allocation. They are therefore at greater risk than RCTs for confounding by indication and selection bias.

In contrast, threats to validity and precision from performance bias, detection bias, inadequate sample size, and lack of study efficiency do not differ markedly in theory between RCTs and observational studies (although some features such as blinding of assessors that protect against detection bias are more likely in experimental designs than in observational studies). For both designs, these risks of bias can threaten the validity of results. Performance bias and detection of effects have the potential to alter effect sizes unpredictably and need to be evaluated as well in observational studies.[4,5] These sources of bias can invalidate the results of observational studies, as can problems with confounding and selection bias. In relation to risks from confounding, Greenland and Morgenstern note that "[b]iases of comparable or even greater magnitude can arise from measurement errors, selection (sampling) biases, and systematically missing data, as well as from model-specification errors. Even when confounding and other systematic errors are absent, individual causal effects will remain unidentified by statistical observations."[6, p. 208]

Table 1. Threats to validity and precision[a]

Threats	Definition
Threats to validity (systematic error)	
Confounding by indication or allocation bias	Systematic differences between baseline characteristics of the groups that arise when patient prognostic characteristics, such as disease severity or comorbidity, influence both treatment source and outcomes. Confounders are the common cause for intervention and exposure; they occur before exposure. Confounding by indication can occur from self-selection of treatments or physician-directed selection of treatments.
Selection bias	Selection bias occurs when studies are conditioned on (that is, they differentially select for) common effects of the exposure and the outcome. Selection bias occurs after exposure and arises when the association between exposure and outcome is different for those who participate compared with those who do not participate in a study (i.e., all those who are theoretically eligible). Includes inappropriate selection of controls in a case-control study, differential loss to follow-up for groups being compared (attrition bias), incidence-prevalence bias, nonresponse bias, and in- or exclusion of specific groups for study.
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.
Detection bias	Systematic differences in outcomes assessment among groups being compared, including 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.
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).
Threats to precision (random error)	
Inadequate study size	Not powered to test study hypothesis for an individual study. (Multiple underpowered small studies can result in a precise meta-analysis estimate)
Lack of study efficiency	Absence of needed stratification in design. When confounding and effect modifiers do not exist, an equal apportionment ratio between exposed and unexposed is the most efficient design. Comparisons within strata may be required to account for known confounders and effect modifiers. Matching on stratification variables allows for an efficient design.

[a] From Hernan et al., 2004,[2] Rothman et al., 2003,[1] Higgins et al., 2006,[4] and Viswanathan et al., 2012.[3]

Confounding

Although confounding is more of a concern for observational studies than RCTs, RCTs are not entirely free of these concerns. Confounding that is random (by chance) is expected to be equally distributed between arms by randomization, but RCTs may not always successfully randomize such potential confounders. Confounding by chance (that is confounding that is unknown, unmeasured, or poorly measured, but expected to be equally distributed) should occur with the same probability in RCTs and observational studies, because, it is, by definition, occurring by chance.[7] Confounding by indication or contraindication, on the other hand, occurs when both treatment and outcome are influenced by a third factor, namely prognosis. Confounding by indication can occur when the expectation of prognosis influences the patient or provider's selection of treatment as well as the potential outcome. Trials, by virtue of concealed randomization, avoid this source of confounding, by "breaking the link between prognosis and prescription."[8, p. e67] Observational studies of benefits that cannot address source of confounding by design (through, for instance, restriction to patients with the same prognosis) must account for it in analysis to the extent possible. Vandenbroucke notes that confounding by contraindication is not always a concern for observational studies of harms: harms are often unanticipated outcomes, so an expectation of prognosis for harms is unlikely to have influenced the selection of treatment.[8, p. e67] The extent to which confounding by indication may occur in observational

studies of harms lies on a gradient, from the completely unanticipated (as with ACE inhibitors and angioneurotic edema) to the potentially likely (as with cardiac arrhythmias induced by anti-arrhythmic drugs).

Confounding by indication may also occur when the person allocating treatment is influenced by factors other than prognosis. Shrier et al. offer the example of the appearance of fatigue causing the physician to select one diabetes treatment over another.[7] When confounding by indication can be controlled, for instance, by the inclusion of physician-rated appearance of patient fatigue as a covariate in modeling, the effect of the potential confounder can be accounted for. Each analysis need only account for those confounders that are expected to have an effect on the outcome and that have not already been accounted for by the inclusion of other closely correlated variables. Thus, the assessment of the potential for bias from confounding requires context-specific understanding of the relationship between treatment and outcomes and study-specific evaluation.[1,6,7]

Selection Bias

Selection bias refers to the selection of the subset of those eligible, when that selection is conditioned upon variables that are the common effect of causes of the exposures and outcomes.[1,2] The specific risks of selection bias vary by type of observational study design. Differential loss to follow-up, for instance, is a concern for cohort studies but not for case-series or cross-sectional studies. On the other hand, a choice of control group that is intrinsically either more or less exposed than the source population of the cases, is a specific form of selection bias in case-control studies.

The concerns regarding attrition bias, however, are similar for RCTs and observational studies in theory: both types of studies may be weakened by high rates of overall or differential attrition. In practice, observational studies may have higher risks of attrition bias: they may have longer time horizons and fewer resources to follow up with participants

Lastly, RCTs often suffer less from inadvertent selection bias caused by the analysis. Usually, in RCTs, investigators will not adjust for other consequences of the treatment, nor will they adjust for consequences of the outcome. However, in large data-base analysis, that is sometimes done unwittingly, as with restriction of the analysis on folic acid supplementation on congenital malformation to live births.[2]

Performance Bias

The risk of performance bias may differ in practice between RCTs and observational studies. In a RCT the intervention is typically standardized so that study participants within groups are exposed (for the most part) to similar interventions or control conditions. Further, co-interventions can be standardized and/or monitored and fidelity to the intervention can be assessed and reported. For observational studies, this standardization may not be possible; therefore, researchers may not have control over how the intervention was administered or the level of exposure. As a result, observational studies may not be able to clearly define intervention states. Hernan and Taubman note that when the principle of consistency (a causal contrast between two or more well defined interventions) is not met (as with studies exploring whether obesity leads to increased mortality), other requirements of caual inference such as exchangeability and positivity cannot be met.[9] This will also vary by specific observational study design features, particularly whether the intervention/exposure occurred prospectively vs. retrospectively with respect to the conduct of the study. Blinding of the providers and

participants may also be of variable concern across different designs. For instance, in a retrospective study, the intervention or exposure would likely have been administered outside the context of a research study; therefore, blinding of the intervention/exposure may not be applicable. In a prospective study, by contrast, blinding of providers and participants is critical to limiting bias that arises as a result of knowing the study hypothesis and what the study participants are receiving. Blinding, or other measures that make assessments objective, is possible in assessing the exposure status of individuals in case-control studies.

Detection Bias

As with performance bias, detection bias may be more problematic in observational studies because outcome assessment may not be standardized as it is more typically with RCTs. For example, in RCTs the same outcome assessment tools are used often with protocols for their implementation and assessment of results. As above, this source of bias can also vary across observational studies, particularly for prospective vs. retrospective designs. In retrospective designs where the measurement of outcomes has occurred prior to the start of the study, the researchers have no control over how those assessments were made, including choice of measurement tools, whether tools were valid and reliable (or a process to ensure their validity/reliability was employed), and how results were interpreted. Blinding of outcome assessors serves to limit detection bias, but not all designs can employ a blinded approach to assessing outcomes. In some studies, researchers serve as outcome assessors and can be blinded; in other studies, participants provide self-reported outcome data and cannot be blinded. As above, blinding of outcome assessors may be of variable concern across different observational study designs.

Information bias is related to detection bias and arises from how measurement and assessment of exposure and outcomes are made. In theory, all designs run the risk of bias from the use of poorly validated measures.

Reporting Bias

Reporting bias is a concern across all research regardless of design: authors are more likely to publish studies and selectively report outcomes that show statistical significance. However, with the recent emphasis on standardized reporting of RCTs and prospective trial registration, reporting bias may become less problematic with RCTs or at least more easily detected as a problem in specific RCTs. While no empirical evidence exists that reporting bias is a greater concern for observational studies, fewer standards exist for observational studies, making reporting bias a concern of equal or greater magnitude than for RCTs.

References

1. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed., Philadelphia, PA: Lippincott, Williams, & Wilkins; 2008.

2. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25. PMID: 15308962.

3. 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. PMID: 21959223.

4. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.2. London: The Cochrane Collaboration; 2009. www.cochrane-handbook.org. Accessed June 9, 2011.

5. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001 Jul 7;323(7303):42-6. PMID: 11440947.

6. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22:189-212. PMID: 11274518.

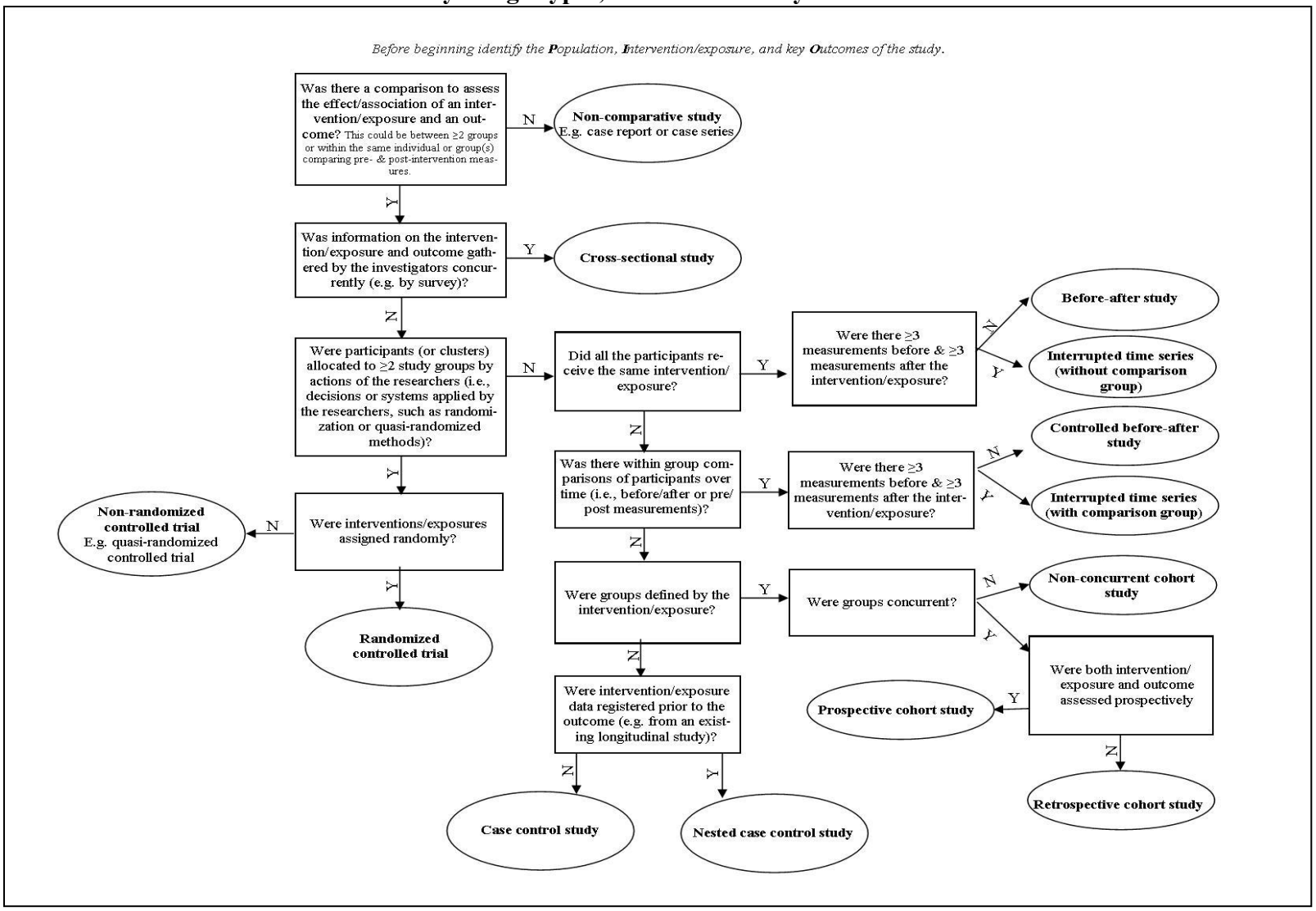
7. Shrier I, Boivin JF, Steele RJ, et al. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. Am J Epidemiol. 2007 Nov 15;166(10):1203-9. PMID: 17712019.

8. Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. PLoS Med. 2008 Mar 11;5(3):e67. PMID: 18336067.

9. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes (Lond). 2008 Aug;32 Suppl 3:S8-14. PMID: 18695657.

Appendix B. Taxonomy on Study Design

Users: To determine observational study design types, use the taxonomy and definitions below



DESIGN ALGORITHM FOR STUDIES OF INTERVENTIONS AND EXPOSURES

When using the algorithm, it is recommended that you do not rely on the design labels assigned by the authors of the report, but rather work through the questions in the algorithm based on the methods presented in the report and the definitions provided below.

Study Design Key

Below is a list of definitions that correlate with study designs assigned by the accompanying taxonomy. At the end of this list are some additional concepts that may be useful during study design classification.

Non-randomized trial

A study in which individuals or groups of individuals (e.g., community, classroom) are assigned to the intervention or control by a method that is not random (e.g., date of birth, date of admission, judgement of the investigator). Individuals or groups are followed prospectively to assess differences in the outcome(s) of interest. The unit of analysis is the individual or the group, as appropriate.

Randomized trial

A study designed to test the efficacy of an intervention on an individual, a group of individuals, or clusters (e.g., classrooms, communities). Individuals or clusters are randomly allocated to receive an intervention or control/comparison (e.g., placebo or another intervention) and are followed prospectively to assess differences in outcomes. The unit of analysis is the individual, group of individuals, or the cluster, as appropriate. Variations in treatment assignment and measurement produce different types of studies including factorial, cross-over, parallel, stepped wedge and Solomon four-group.

Prospective cohort study

A study in which individuals in the group without the outcome(s) of interest (e.g., disease) are classified according to exposure status at baseline (exposed or unexposed) and then are followed over time to determine if the development of the outcome of interest is different in the exposed and unexposed groups.

Retrospective cohort study

A study in which a group of individuals is identified on the basis of common features that were determined in the past. The group is usually assembled using available data sources (e.g., administrative data). Individuals are classified according to exposure status (exposed or unexposed) at the time the group existed and are followed up to a prespecified endpoint to determine if the development of the outcome of interest is different in the exposed or unexposed groups.

Interrupted time series with comparison group

A study in which multiple observations over time are “interrupted” by an intervention or exposure and in which two series are examined (one is a comparison group). There must be at least 3 observations before and at least 3 observations after the intervention or exposure for each group. The investigator(s) does not assign or have control over the intervention/exposure, which

may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation, educational program, service delivery model) but does have control over the timing of the measurement and the variables being measured.

Controlled before-after study
A study in which the outcome(s) of interest is measured both before and after the intervention or exposure in two or more groups of individuals. In this study design the study group receives the intervention or exposure and the comparison group(s) does not. This type of study includes interventions that may be in the control of the investigator (e.g., a surgical procedure) as well as interventions that may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation). In all cases the investigator(s) has control over the timing of the measurement and the variables being measured.

Non-concurrent cohort study
A study in which 2 or more groups of individuals are identified on the basis of common features at different time points. Individuals in each group are classified according to exposure status (exposed or unexposed) at the time the groups existed or were created. They are followed to determine if the development of the outcome of interest is different in the exposed or unexposed groups.

Nested case control study
A study where exposed and control subjects are drawn from the population of a prospective cohort study. Baseline data are obtained at the time the population is identified; the population is then followed over a period of time. The study is then carried out using persons in whom the disease or outcome has developed and a sample of those who have not developed the outcome of interest (controls).

Case control study
A study in which participants are selected based on the known outcome(s) of interest (e.g., disease, injury). Exposure status is then collected based on the participants' past experiences. Exposure status is compared between the two (or more) groups: those who have the outcome of interest and those who do not have the outcome of interest (controls). This is a retrospective study that collects data on events that have already occurred.

Interrupted time series (without a comparison group)
A study in which multiple observations over time are "interrupted" by an intervention or exposure. There must be at least 3 observations before the intervention and at least 3 observations after the intervention; otherwise, the study is considered a before-after study. The investigator(s) does not assign or have control over the intervention/exposure, which may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation, educational program, service delivery model) but does have control over the timing of the measurement and the variables being measured.

Before-after study
A study of an intervention or exposure in which the investigator(s) compares the outcome(s) of interest both before and after the intervention in the same group of individuals. This includes

interventions that may be in the control of the investigator (e.g., a surgical procedure) as well as interventions that may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation). In all cases the investigator(s) has control over the timing of the measurement and the variables being measured.

Cross-sectional study
A study in which both the exposure and the outcome status in a target population are assessed concurrently, that is, at the same point in time or during a brief period of time. The temporal sequence of cause and effect cannot necessarily be determined. They are most commonly used to assess prevalence. A common method for data collection is a survey.

Non-comparative study (case series)
Examples of this design include:

- A study that presents a description of a single patient or participant. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of the individual.
- A study that describes the experience of a group of patients with a similar diagnosis and/or treatment. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of a condition.
- A study in which data are collected at a series of points in time on the same population to observe trends in the outcome(s) of interest.

Additional Concepts

Cluster
The term 'cluster' refers to a unit of allocation or analysis in a clinical trial. Examples of clusters include hospitals, schools, neighborhoods, or entire communities.

Cluster randomized controlled trial
Synonym: *community trial; group randomized trial*
A randomized controlled trial in which the units of randomization and analysis are groups of people or communities (e.g., classroom, hospital, town). Typically, several communities receive the intervention and several different communities serve as controls.

Cohort
The term 'cohort' refers to a group of individuals (or other organizational units) who have a common feature when they are assembled (e.g., birth year, place of employment, medical condition, place or time period of medical treatment) and are followed over time. They can be followed prospectively or examined retrospectively.

Experimental study
A type of study in which investigators have direct control over the timing, course, and assignment of the intervention. Experimental studies investigate an intervention to determine its effect on the outcome(s) of interest. In an experimental study a population is selected to receive a specific intervention the effects of which are measured by comparing the outcomes in the

experimental group with the outcomes of a control group that has received another intervention or placebo. Examples include randomized controlled trial, cluster randomized controlled trial, nonrandomized trial, n-of-one trial. See also **observational study**.

Observational study
A study in which the investigator(s) does not control the exposure/ intervention status of study participants (i.e., the assignment of the intervention or exposure of interest is not under the control of the investigator(s)). The simplest form of observational study is the case report or case series, which describes the clinical course of individuals with a particular condition or diagnosis. Observational studies include descriptive and analytic studies. See also **experimental study**.

Quasi-experimental study
A type of study in which the investigator(s) evaluates the effect of an intervention but does not have full control over the timing, course, or allocation of the intervention. They are often used when it is not possible to conduct a true experimental study.

FROM: Hartling L, Bond K, Harvey K, Santaguida PL, Viswananthan M, Dryden DM. Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures. Agency for Healthcare Research and Quality; December 2010. Methods Research Report. AHRQ Publication No. 11-EHC-007. Available at http://effectivehealthcare.ahrq.gov/.

Appendix C. Item Bank for Assessing Risk of Bias and Confounding for Observational Studies of Interventions or Exposures

This item bank is intended to evaluate the quality of studies examining the outcomes of interventions, treatments, or exposures. Eligible study designs include observational studies (cohort studies, case-control, case-series, and cross-sectional studies). It is not intended to rate the quality of studies concerning the accuracy of diagnostic tests. Abstractors can use the empty text box included with each question to document an explanation of their rating for later review. This may be particularly helpful in relation to a "cannot determine" response choice.

Study Definitions

Case series

Description: A study that describes a group of patients with a similar diagnosis and/or treatment. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of a condition through a collection of individual case reports.

Design features:

1. There is no comparison between groups to assess the effect/association of an intervention/exposure and an outcome.
2. There is no comparison with the same group over time.

Cross-sectional study

A study in which both the exposure and the outcome status in a target population are assessed concurrently that is, at the same point in time or during a brief period of time. The temporal sequence of cause and effect cannot necessarily be determined. They are most commonly used to assess prevalence. A common method for data collection is a survey.

Case control study

A study in which participants are selected based on the known outcome(s) of interest (e.g., disease, injury). Exposure status is then collected based on the participants' past experiences. Exposure status is compared between the two (or more) groups: those who have the outcome of interest and those who do not have the outcome of interest (controls). This is a retrospective study that collects data on events that have already occurred.

Cohort studies

A study in which individuals in the group without the outcome(s) of interest (e.g., disease) are classified according to exposure status (exposed or unexposed) and then are followed over time

to determine if the development of the outcome of interest is different in the exposed and unexposed groups.

Q1: Do the inclusion/exclusion criteria vary across the comparison groups of the study? *[PI: Drop question if not relevant to all included studies. To use this question for studies with one group, the focus of the question on comparison groups and related response categories would need to be changed to individuals.]*

PI:

Yes, varies ..... ☐

Partially: some, but not all criteria, applied to all groups or not clearly stated if some criteria are applied to all groups ..... ☐

No, does not vary ..... ☐

Cannot determine: article does not specify ..... ☐

Not applicable: study has only one group and so does not include comparison groups ..... ☐

Explanation for rating:

Q2: Does the strategy for recruiting participants into the study differ across groups? *[PIs: Drop question if not relevant to all included studies. To use this question for studies with one group, the focus of the question on comparison groups and related response categories would need to be changed to individuals.]*

PI:

Yes, differs ..... ☐

No, does not differ ..... ☐

Cannot determine ..... ☐

Not applicable: one study group ..... ☐

Explanation for rating:

Q3: Is the selection of the comparison group inappropriate, after taking into account feasibility and ethical considerations? *[PI: Provide instruction to the abstractor based on the type of study. Interventions with community components are likely to have contamination if all groups are drawn from the same community. Interventions without community components should select groups from the same source (e.g., community or hospital) to reduce baseline differences across groups. For case-control studies, controls should represent the population from which cases arose; that is, controls should have met the case definition if they had the outcome.]*

PI:

Yes, inappropriate.....	☐	***Explanation for rating:***
No, not inappropriate	☐	
Cannot determine or no description of the derivation of the comparison group.....	☐	
Not applicable: study does not include a comparison group (case series, one study group)	☐	

Q4: Does the study fail to account for important variations in the execution of the study from the proposed protocol? *[PI: Consider intensity, duration, frequency, route, setting, and timing of intervention/exposures. Drop if not relevant for body of literature.]*

PI:

Yes, fails to account	☐	***Explanation for rating:***
Partially, fails to account	☐	
No, does not fail to account	☐	
Cannot determine.....	☐	
Not applicable: not an intervention study or no variations	☐	

Q5: Was the outcome assessor not blinded to the intervention or exposure status of participants? *[PI: There may be circumstances where clinical evaluators cannot be blinded to exposure status. Drop if not relevant to the body of literature.]*

PI:

Yes, not blinded.....	☐	**Explanation for rating:**
No, blinded.....	☐	
Not applicable: assessor cannot be blinded	☐	

Q6: Were valid and reliable measures, implemented consistently across all study participants used to assess inclusion/exclusion criteria, intervention/exposure outcomes, participant health benefits and harms, and confounding? *[PI: Important measures should be identified for abstractors and if there is more than one, they should be listed separately. PI may need to establish a threshold for what would constitute acceptable measures based on study topic. When subjective or objective measures could be collected, subjective measures based on self-report may be considered as being less reliable and valid than objective measures such as clinical reports and lab findings. Some characteristics may require that sources for establishing their validity and/or reliability be described or referenced. If so, provide instruction to abstractors.]*

PI:

Yes, valid and reliable measure used.....	☐	**Explanation for rating:**
No, valid and reliable measure not used	☐	
Cannot determine or measurement approach not reported	☐	

Q7: Was the length of follow-up different across study groups? *[Abstractor: When follow-up was the same for all study participants, the answer is no. If different lengths of follow-up were adjusted by statistical techniques, (e.g., survival analysis), the answer is no. Studies in which differences in follow-up were ignored should be answered yes.]*

PI:

Yes, different or cannot determine ..... ☐

No, not different or remedied through analysis... ☐

Not applicable: cross-sectional or only one group followed over time..... ☐

Explanation for rating:

Q8: In cases of high loss to follow-up (or differential loss to follow-up), was the impact assessed (e.g., through sensitivity analysis or other adjustment method)? [PI: Attrition is measured in relation to the time between baseline (allocation in some instances) and outcome measurement for both retrospective and prospective studies and could include data loss from switching. Attrition rates may vary by outcome and time of measurement. Specify the criterion to meet relevant standards for the topic. Specify measurement period of interest, if repeated measures. Cochrane standard for attrition is 20 percent for shorter term (<1 year) and 30 percent for longer term (≥ 1 year).]

PI:

Yes, impact assessed ..... ☐

No, impact not assessed..... ☐

Cannot determine..... ☐

Not applicable: no loss to follow-up or loss to follow-up was not considered to be high, cross-sectional study, or case-control study selected on outcome..... ☐

Explanation for rating:

Q9: Are any important primary outcomes missing from the results? *[PI: Identify all primary outcomes that one would expect to be reported in the study, including timing of measurement.]*

PI:

Yes, important outcome(s) missing	☐	**Explanation for rating:**
No important outcome(s) missing.....	☐	
Cannot determine.....	☐	

Q10: Are any important harms or adverse events that may be a consequence of the intervention/exposure missing from the results? *[PI: Identify all important harms that one would expect be reported in the study, including timing of measurement. Drop if not relevant to body of literature.]*

PI:

Yes, important outcomes missing.....	☐	**Explanation for rating:**
No important outcomes missing.....	☐	
Assessment of harms not applicable to this study.....	☐	

Q11: Are results believable taking study limitations into consideration? *[Abstractor: This question is intended to capture the overall quality of the study. Consider issues that may limit your ability to interpret the results of the study. Review responses to earlier questions for specific criteria.]*

PI:

Yes, believable ..... ☐

No, not believable ..... ☐

Explanation for rating:

Questions to Assess Confounding (Q6, Q12-13)

Q12: Any attempt to balance the allocation between the groups or match groups (e.g., through stratification, matching, propensity scores). *[PI: Drop if not relevant to the body of evidence.]*

PI:

Yes or study accounts for imbalance between groups through a post hoc approach such as multivariate analysis..... ☐

No or cannot determine ..... ☐

Not applicable: study does not include a comparison group (case series or one study group) ..... ☐

Explanation for rating:

Q13: Were important confounding variables not taken into account in the design and/or analysis (e.g., through matching, stratification, interaction terms, multivariate analysis, or other statistical adjustment such as instrumental variables)? [PI: Provide instruction to abstractors on known confounding variables and inadequate adjustment for confounding for each outcome.]

PI:

Yes, not accounted for or not identified.....	☐	**Explanation for rating:**
Partially: some variables taken into account or adjustment achieved to some extent.....	☐	
No: taken into account…	☐	
Cannot determine.....	☐	

Modified from: 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. PMID: 21959223.