Draft Methods Research Report

Number XX

Assessing Confounding, the Risk of Bias and Precision of 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. [redacted]

Prepared by: [redacted]

Investigators:

AHRQ Publication No. xx-EHCxxx
<Month Year>

This draft evidence report/technology assessment is distributed solely for the purpose of pre-release peer review. It has not been otherwise disseminated by AHRQ. It does not represent, and should not be construed to represent, an AHRQ determination or policy.

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

Suggested Citation:

<Authors> Assessing Confounding, the Risk of Bias and Precision of Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank<EPC> Evidence-based Practice Center, Contract No. <Contract Number>. (AHRQ to provide the remainder)

This information is distributed solely for the purposes of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the Agency for Healthcare Research and Quality. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services determination or policy.

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. 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 bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

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.

We welcome comments on this evidence report. 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 e-mail 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
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: <Acknowledgments>.

Peer Reviewers

<Name>
<Place>
<City>, <ST>

<Name>
<Place>
<City>, <ST>

Assessing Confounding, the Risk of Bias and Precision of 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 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 their 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. This methods project proposes 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

Introduction 3
Empirical Assessments of Risk of Bias 4
Need for a Revised Approach to Evaluating Observational Studies 4
Project Objectives 6
Methods 6
Results 7
Expand the Analytic Framework to Consider Sources of Confounding 8
Identify the Source of Bias and Confounding Specific to Each Included Design 9
Assess Individual Studies for Risk of Bias and Potential Confounders 10
Consider Competing Hypotheses for Observational Studies 14
Discussion 15
Recommendations for Future Methods Development 16
References 17

Tables

Table 2. Questions from the item bank Working Group voting participants considered very or somewhat important for evaluating risk of bias and precision by observational study design type 11

Table 3. Questions and instructions in revised RTI Item Bank following voting exercise by category of bias or precision assessed 12

Table 3. Questions and instructions in revised RTI Item Bank following voting exercise by category of bias or precision assessed (continued) 12

Figures

Figure 1. Comprehensive framework assessing the validity of causal links for observational studies in systematic reviews 8

Appendixes

Appendix A. Approaches to Assessing the Risk of Bias in Studies A-1

Appendix B. Item Bank for Assessment of Risk of Bias and Precision for Observational Studies Of Interventions or Exposures B-1

Appendix C. Taxonomy on Study Design C-1

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] This way of thinking about causal inference implies that three conditions need to be met: consistency, exchangeability, and positivity.[2] Consistency is "the condition that causal contrast involves two or more well-defined interventions."[3, p. S10] Exchangeability refers to the condition of independence between the counterfactual outcome and the actual treatment, in other words, the treated and the untreated are exchangeable.[2] The condition of positivity requires that the each participant in the study has a probability of receiving every value of the treatment that is greater than zero, in other words, a positive value.[2]

Randomization offers each participant equal probability of being assigned to all levels of treatment (positivity), distributes unmeasured confounding randomly (exchangeability), and clearly sets up contrasts between intervention states (consistency). 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] They may be limited in their ability to address certain outcomes, such as adverse effects or long term benefits, because of limited size or length of followup.[5] RCTs may provide insufficient information on vulnerable subpopulations.[6] In the absence of sufficient evidence from RCTs, 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 is mirrored by other recent guides that offer cautious support for reliance on observational studies.[8, 9]

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.[10] When reviewers rely on evidence from observational studies, the inferences they draw from observational studies need to account for potential important confounding, as researchers did in a recent asthma study.[a]

In this project, we reviewed methodological considerations when evaluating the validity of studies included in systematic reviews, particularly 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 in

[a] As an example, a recent systematic review examined the association between the use of acetaminophen and risk of asthma.[11] A subsequent prospective cohort study attributed the association to confounding by indication.[12] 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 paracetamol 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 to confounding by indication. The researchers commented that a randomized trial set up for 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.

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) raises a fundamental question when assessing causality in observational studies: do observational studies routinely provide a different (more inflated) estimate of effect than trials?[13] If so, can they be discounted to arrive at a closer approximation of the true effect? Empirically, however, reviews have found no difference[14, 15] or inconsistent differences between types of designs (that is, RCTs sometimes had smaller estimates of effect than observational studies, and vice versa).[16, 17] This unpredictability in direction of effect means that systematic reviews cannot rely on a rule of thumb (based on, for instance, type of design)[18] to discount evidence from observational studies. Instead, a careful assessment of risk of bias that accounts for the clinical context and for the individual study performance is necessary to evaluate the validity of estimates from observational studies.[13]

The conclusion that risk of bias assessment need to account for clinical context, gives rise to the next question: can empirical assessments of risk of bias help to shape our understanding of the validity of evidence from observational studies? MacLehose and colleagues evaluated differences in estimates of effect between high and low quality studies using two approaches. They first examined, across interventions, differences between estimates of effect within single manuscripts that compared outcomes from quasi-randomized and observational studies (QEO) and RCTs. Second, they looked at differences in effect sizes across studies of different design types for two different interventions. Their results, for the first strategy, showed no differences in estimates of effect between QEO studies and RCTs of high quality, but greater differences between low quality estimates. Their findings suggest not only that low-quality QEO evidence can be misleading, but also the importance of separating high from low quality. Their evaluation of differences in effect size (the second strategy) for specific interventions failed 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."[15, 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 method 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."[18, p. 1208] A state of consensus on this issue does not yet exist.

Need for a Revised Approach to Evaluating Observational Studies

A review of critical appraisal tools for assessing the risk of bias of observational studies identified key domains but found no gold standard.[19] These tools typically focus their assessments on either capturing descriptions or reporting in the manuscript of the methods authors used in designing or conducting particular elements of the study; or judgments of risk of bias based on these design or implementation decisions (questions regarding whether the conduct of the study altered the validity of results). A recent review by Mhasker and colleagues reveals

the inadequacies of risk of bias assessment tools that rely on reporting. They compared actual quality in protocols to reported quality in publications from 429 RCTs. Their results showed that reported quality did not reflect the actual high quality of the trials; moreover the associations between effect size and reported quality did not persist when examining actual quality.[20]

Another constraint of existing instruments is the reliance on a scale. Empirical tests of the validity and reliability of these scales suggests the need for critical analysis 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.[21] 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.[22]

Our own previous work offers guidance on specific questions for identifying context specific source of biases and confounding for observational studies (Appendix B), [23] but stopped short of offering guidance on a larger framework for its use. A total of 1,492 questions identified in previous quality rating tools were culled and refined through face validity, cognitive, content validity, and inter-rater reliability testing. The resulting RTI Item Bank consists of 29 comprehensive questions which seek to capture the risk of bias and precision domains critical for evaluating this type of research and includes directions for adapting questions to different topic areas and some assistance in selecting questions for different study designs (e.g., cohort, case control, cross-sectional and case series studies).[23] The next step in developing this tool required consensus around which specific questions were essential for each design, and further refinement of the questions.

Available empirical evidence and guidance cautions us against ignoring the risks of bias in observational studies or merely applying weights to observational studies based on scores on a bias checklist.[22, 24, 25] Rather, the risk of bias requires analysis and interpretation, particularly in light of heterogeneity of the evidence. Existing guidance offers limited assistance with analysis and interpretation of observational studies, particularly with regard to confounding. The *Cochrane Handbook of Systematic Reviews*,[9] AHRQ guidance,[7] and the Institute of Medicine standards for systematic review[8] all detail reasons for inclusion of nonrandomized studies for harms, and attendant risks.[26] These documents caution that the risk of bias will always be greater for nonrandomized studies than for randomized studies. All offer general guidance on sources of bias for observational studies.

Both the IOM[8] and AHRQ guidance[27] discuss the role for plausible confounding to increase the strength of evidence in a rating system in which observational studies start out with a low grade, but offer no overarching framework to consider the issue of confounding. The Cochrane guidance notes that issues of confounding cannot be easily addressed within existing instruments on bias and suggests developing new tables to identify prespecified confounders and variables controlled in the analysis for each study. This information is expected to illustrate the extent of heterogeneity in the literature, but the authors do not require that all Cochrane reviews including observational studies develop these tables; they need to demonstrate, however, that they have considered the role of residual confounding in explaining the findings from nonrandomized studies. [9]

The work described above suggests the following needs in evaluating causality in observational studies:(1) consensus around sources of bias for different observational study design;[1, 6, 18] (2) content-specific criteria to rate risk of bias;[1, 6, 18] and (3) use of risk of bias criteria to understand heterogeneity in results rather than merely to weight pooled estimates of

effect;[22, 24, 25] and (4) a comprehensive framework to adequately capture risk of bias and confounding concerns that are specific to a body of observational studies included in a systematic review.

Project Objectives

As noted above, the project's objectives including creating an overarching framework on the risk of bias and confounding and developing consensus around sources of bias for different observational study design. 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 comparative effectiveness review could not be accomplished solely within the confines of questions included in an item bank evaluating individual studies. Therefore the focus of the project expanded to more adequately accommodate this goal. Tasks included: (1) developing an analytic framework to assess the effect of confounding across the body of observational study evidence as well as within individual studies; (2) integrating a taxonomy to assist reviewers in determining the designs of the observational studies being reviewed; (3) identifying critical sources of bias most common to each observational study design type; (4) refining and reducing the set of "core" questions that would most likely be necessary for evaluating risk of bias and confounding concerns for each design; and (5) refining the instructions provided to users to improve their clarity and usefulness so that the item bank is a readily available and practical tool that can be used across EPCs and others who are conducting systematic reviews.

Determining the best set of questions from those available within the item bank that should be used for specific study designs and subject topics requires a certain level of epidemiological expertise in classifying study designs, experience conducting systematic reviews, and familiarity with risk of bias rating. Even experienced researchers show poor inter-rater reliability in classifying study designs.[28] To facilitate optimal use of the item bank, further development was needed to provide greater guidance to users in appropriately identifying study designs included in their reviews, and the possible bias concerns related specifically to those designs (sources of bias referred to above). Users could then 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. 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 to comment on the goals and activities of the project. As noted above, their input resulted in an expansion of our original objective from further refinement of the RTI Item Bank to also include the development of an overarching framework to consider the effect of confounding across the body of observational study evidence.

We built on a previously developed taxonomy of observational studies (Appendix C)[28, 29] that offered a framework to group studies of similar designs, or with similar design features that may relate to bias. The tool can be used by systematic reviewers to guide the choice of questions

needed for risk of bias assessments of different observational study designs. The premise is that studies of different designs, or with different design features, may require (some) different questions for risk of bias assessments. For example, studies identified by the tool as "non-comparative" include case reports and case series that have no comparison group. Therefore, any questions to assess risk of bias must be selected with this in mind (e.g., not ask about comparability between groups). Another example of important study design features that the taxonomy can assist in clarifying are whether both intervention/exposure and outcome assessment were prospective; to differentiate between prospective and retrospective cohort studies. 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 using standard definitions/criteria.

The study team then 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 limited set of commonly used observational study designs (cohort, case-control, case-series, and cross-sectional designs).

We revised the instrument to consolidate questions addressing the same bias concerns (e.g., use of valid and reliable measures) and separated out 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 confounding, bias, and lack of precision of each included observational study and the subset of questions that would only be relevant for particular observational study designs (case series, case-control, cohort or cross-sectional studies). Working Group members were provided with the revised version of the bank that included 16 of the original 29 questions. They were asked to evaluate the importance of each of the questions 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 ten questions. (One question had been eliminated based on earlier discussions with the working group and two were eliminated based on question consolidations.) We also sought additional comments from Working Group members on each questions. 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).

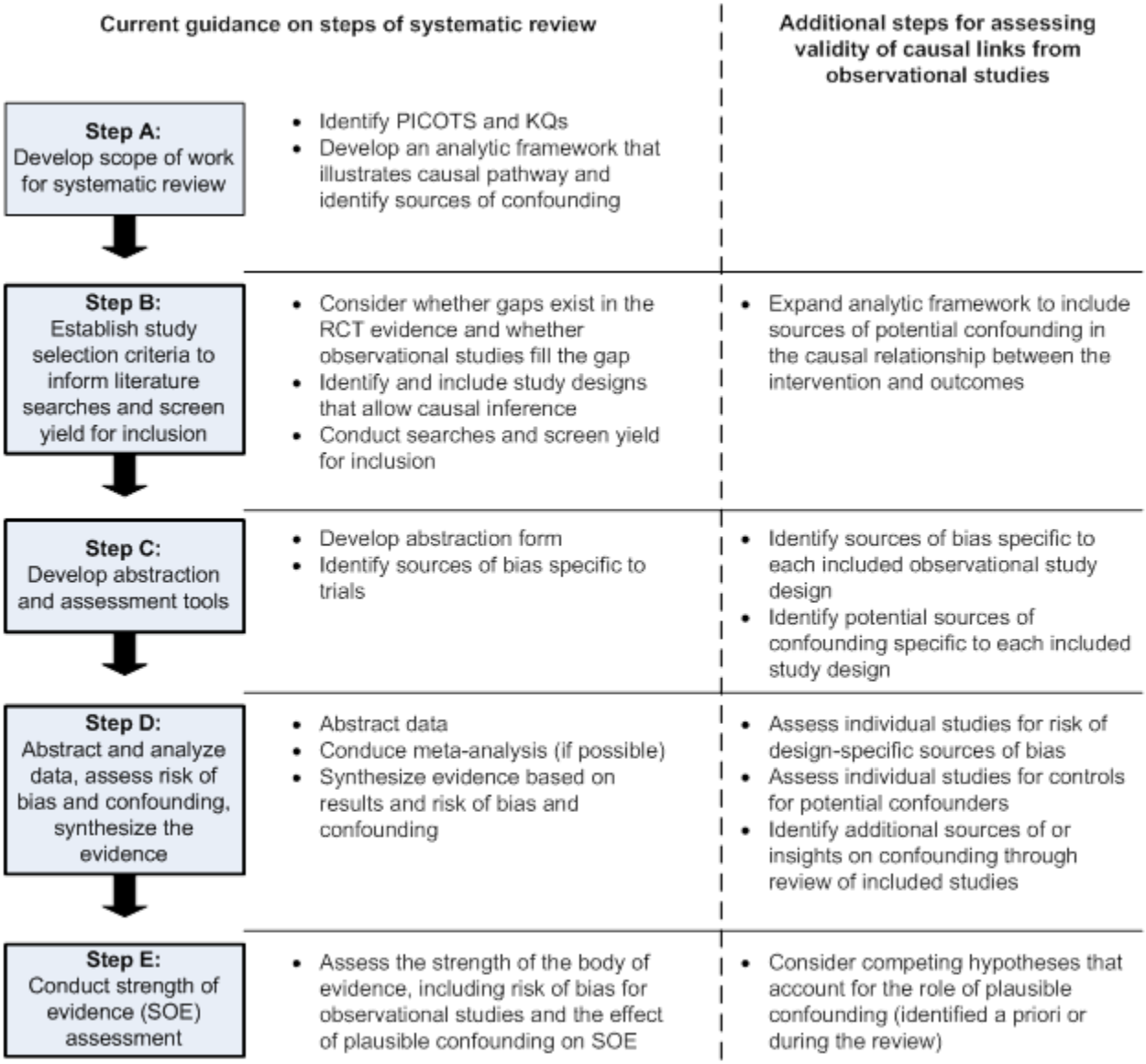
This document will be revised in response to peer review and public comment.

Results

In consultation with the Working Group, we developed a comprehensive framework to address the evaluation of observational studies in systematic reviews (Figure 1). This framework builds on existing routine processes of a systematic review of 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, 27, 30-33] The inclusion of observational studies requires additional tasks highlighted in the right-hand column of the figure and described in detail below.

Figure 1. Comprehensive framework assessing the validity of causal links for observational studies 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. As shown in Step B, the analytic framework should clearly identify potential confounders that would affect the relationship between the intervention and the outcome. The specific sources of confounding will depend upon the clinical context. The potential for uncontrolled confounding in an overall body of evidence may influence what

systematic reviewers judge to be admissible study designs. Current 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 and colleagues suggest considering the clinical context and natural history 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 when evaluating observational studies. Norris et al. note the importance of selection bias and confounding by indication as particular problems plaguing observational studies but do not offer thresholds for when anticipated concerns from bias would preclude the consideration of observational studies. We believe, as do other commentators, that such standards cannot be established across all clinical areas and require an understanding of the threats to validity for each topic.

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

In Step C, systematic reviewers would consider the potential observational study designs, and risks of bias and confounding that are likely to be a concern in the body of evidence being reviewed. Based on discussions with the Working Group, we changed the orientation of some of the questions in the item bank from focusing on whether a solution to a bias concern in a study was adequate to whether it was inadequate (Table 2). This shifts the burden for the reviewer from determining whether the study was "good" to identifying salient problems with the study's approach. For example, one question was changed from, "Did the study apply inclusion/exclusion criteria uniformly to all comparison groups/arms of the study" to "Does 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 prior to conducting the voting exercise, "Is the study design prospective, retrospective, or mixed"? 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.

Three of the six members of the Working Group participated in one round of testing of the instrument and one additional member provided comments on item bank questions but did not participate in the formal exercise. Table 2 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.

Of the 16 questions, five were not considered important or very important across study designs by at least two of the three Working Group members. Particularly because of the small number of participants, we have not generally recommended to delete questions but note Working Group members' concerns. Question 4 concerning whether the study failed to account for variations in the execution of the study from the proposed protocol was considered by one of the Working Group members 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 to be determining 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, 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. One Working Group member thought that a post-hoc power calculation to answer question 14 concerning the evaluation of precision through the sufficiency of sample size was no longer needed and thought that reviewers could instead adopt the GRADE Working Group's guidance on optimal information size as a study (body of evidence) being precise if it includes greater than 400 events.[34]

By study design type, the resulting item bank includes the following number of questions as being very important or somewhat important, out of the full bank of 16 questions: case series (8), case-control (8), cohort (12), and cross-sectional (10) (see Table 2). Working Group members agreed with the study team's decision to exclude additional questions that had been included in an earlier version of the item bank.[23]

As for confounding, 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).

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 which identified confounders are controlled in specific study analyses and to determine if additional confounders were identified by study authors. For each of the questions included in the item bank, we developed or revised detailed instructions to assist reviewers (Table 3). Working Group members made few comments concerning this aspect of the item bank. The study team completed minor edits to improve readability.

Similarly, we developed or revised detailed instructions to assist reviewers when assessing observational studies for potential confounders.

Table 2. 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 follow-up different across study groups?	X[b]		X	
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)?			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] Because case series studies involve one group, differences would be evaluated across individuals.

Table 3. 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. 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 and confounding
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.]*	Selection bias and 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 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.]*	Detection bias
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.]*	Detection bias; Confounding
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.]*	Attrition bias
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).]*	Attrition bias; detection bias

Table 3. 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**
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	**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
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.]*	Confounding
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.]*	Confounding
Q14	**Is the sample size sufficiently large to detect a clinically significant difference between groups in at least one primary outcome measure?** *[PI: Specify a particular percent that would be clinically relevant for each outcome of interest. PI may want to consider optimal information size (OIS) standards established by the GRADE Working Group in Guidelines #6 on imprecision instead of this question.[34] Question relates to precision; reviewers whose evaluation of quality is limited to considerations of systematic error or risk of bias (not random error/precision) need not include this question. Reviewers who include both precision and systematic error in their evaluation of quality but rely on meta-analysis for pooled estimates need not include this question. PIs who choose to include considerations of precision in their assessment may include the question, but should be aware of the need for collaboration between clinical and statistical expertise in determining the threshold for a clinically adequate sample size.]*	Precision
Q15	**Are the statistical methods used to assess the primary benefit outcomes inadequate?** *[Abstractor: Question relates to precision and may not be relevant for systematic reviews that are able to pool data. The statistical techniques used must be appropriate to the data for each outcome and take into account issues such as controlling for dose-response, small sample size, clustering, rare outcomes, and multiple comparisons. In normally distributed data the standard error, standard deviation, or confidence intervals should be reported. In non-normally distributed data, inter-quartile range should be reported. For cohort studies, if the outcome has a greater than 10 percent prevalence, consider if the risk ratio and relative risk need to be calculated]*	Precision
Q16	**Are the statistical methods used to assess the main harm or adverse event outcomes inadequate?** *[Abstractor: Question relates to precision and may not be relevant for systematic reviews that are able to pool data. The statistical techniques used must be appropriate to the data for each outcome and take into account issues such as controlling for dose-response, small sample size, clustering, rare outcomes, and multiple comparisons. In normally distributed data, the standard error, standard deviation, or confidence intervals should be reported. In non-normally distributed data, inter-quartile range should be reported.]*	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 evidence. The dominant approaches to conducting this step of the systematic review comes from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group and related systems, such as the AHRQ approach to grading the strength of evidence.[27] These systems aim to offer a systematic and explicit approach to making judgments about the evidence: systematic reviewers provide an assessment in their confidence about how close the true effect is to the observed effect (strong, moderate, low, very low confidence)[35] by considering limitations of the body of evidence, indirectness, imprecision, inconsistency, and publication bias. Trials generally begin with a high confidence and then are marked down for various flaws; observational studies start with low confidence and then can be marked up if they demonstrate a large magnitude of effect, dose-response relationship, or plausible confounding that would have otherwise weakened the effect estimate.[36]

These criteria draw from a long tradition of using causal criteria to evaluate the validity of causal links, dating back to the report of Advisory Committee to the Surgeon General on the association between smoking and lung cancer,[37] Bradford Hill,[38] and Susser.[39] Bradford Hill presented nine "viewpoints" to consider when interpreting an association as causal in nature. These included strength (is the strength of the association is 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?).[38]

Shunemann and colleagues note that the GRADE approach to assessing the quality of the evidence and strength of recommendations draws implicitly or explicitly from numerous Bradford Hill criteria for causation. They also postulate that specificity in particular is not an important criterion for evaluating the effects of interventions[41] (because a single cause can result in multiple effects and vice-versa).[b] With regard to how causality is interpreted for observational studies, GRADE starts observational studies at a lower grade in comparison with trials, thus addressing the experimental evidence criterion. It does so overall, and makes no distinction between study evaluated intended benefit and unintended harm. GRADE's initial starting point for observational studies results in downgrading designs that do not address temporality (either through randomization or through concurrent control groups). GRADE also requires 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, in the GRADE system, 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 the likelihood of

[b] Others suggest that specificity does play an important role in causal inference because the construct of an effect having asingle cause can be used to distinguish causal relationships from noncausal associations.[40]

[c] Schunemann 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.[41]

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 the grading the evidence if they seek to understand causality.

Criticisms of meta-analyses of observational studies also point to the need for critical analysis that focuses on heterogeneity.[25, 42] 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[25] and represent only one type of uncertainty.[42] Another criticism relates to overreliance of the consistency criterion when evaluating the strength of evidence: MacLure and colleagues 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."[42, 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.[43]

Our proposed approach, illustrated in Figure 1, does not reject the current process for strength of evidence rating: it build on that assessment to evaluate multiple competing hypotheses as a final step of the analysis. Examples of such approaches show that risks of this additional critical appraisal include incompleteness (no review can completely rule out alternative hypotheses) and complexity.[42, 44]

Discussion

This project proposes 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 that 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. These 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 RTI risk of bias tool: 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?

This project is a pilot effort at providing 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. 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, Robins JM. Casual inference. Harvard; 2012.
3. Hernan 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.
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. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep; 15(5):615-25. PMID: 15308962.
11. 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.
12. Lowe AJ, Carlin JB, Bennett CM, et al. Paracetamol use in early life and asthma: prospective birth cohort study. British Medical Journal. 2010 Sep 15; 341:c4616. PMID: ISI:000282093800015.
13. Vandenbroucke JP. Why do the results of randomised and observational studies differ? BMJ. 2011; 343:d7020. PMID: 22065658.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

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

26. 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 Rockville, MD: March 2012. www.effectivehealthcare.ahrq.gov/. AHRQ Publication No. 12-EHC047-EF.

27. 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. Journal of Clinical Epidemiology. 2010 May; 63(5):513-23. PMID: ISI:000276534800009.

28. 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.

29. 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 [Internet]. Rockville, MD: Agency for Healthcare Research and Quality; Dec 2010.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr; 64(4):401-6. PMID: 21208779.

36. 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.

37. 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.

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

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

40. 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.

41. Schunemann 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.

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

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

44. Cornfield J, Haenszel W, Hammond EC, et al. Smoking and Lung Cancer - Recent Evidence and a Discussion of Some Questions. Journal of the National Cancer Institute. 1959; 22(1):173-203. PMID: ISI:A1959WD28600009.