Methods Research Report

Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies

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

Methods Research Report

Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies

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 Evidence-based Practice Center
Research Triangle Park, NC

Investigators:
Meera Viswanathan, Ph.D.
Nancy D. Berkman, Ph.D.

AHRQ Publication No. 11-EHC028-EF
September 2011

This report is based on research conducted by the RTI International–University of North Carolina Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-0056-I). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. 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 clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the 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 or actions 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 [for reports related to the Effective Health Care Program], or TAP@ahrq.hhs.gov [for reports related to the Technology Assessment Program, located at http://www.ahrq.gov/clinic/techix.htm], or info@ahrq.gov [for other documents developed by Evidence-based Practice Centers.]

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

Suggested citation: Viswanathan M, Berkman ND. Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies. Methods Research Report. (Prepared by the RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2007-0056-I.) AHRQ Publication No. 11-EHC028-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2011. Available at: 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 e-mail to epc@ahrq.hhs.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

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Carla Bann, Mey-Tzy Kuo, Kathleen Lohr, Loraine Monroe, and Elizabeth Tant.

Technical Expert Panel

Wendy Bruening, Ph.D.
ECRI Institute
Plymouth Meeting, PA

Tim Carey, M.D.
University of North Carolina at Chapel Hill
Chapel Hill, NC

Michael Handrigan, M.D.
Agency for Healthcare Research and Quality
Rockville, MD

Marie Griffin, M.D., M.P.H.
Vanderbilt University
Nashville, TN

Katherine Hartmann, M.D., Ph.D.
Vanderbilt University
Nashville, TN

Susanne Hempel, Ph.D.
RAND
Santa Monica, CA

Eric Johnson, Ph.D., M.P.H.
Kaiser Permanente Northwest
Portland, OR

Robert Kane, M.D.
University of Minnesota
Minneapolis, MN

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

Mark Oremus, Ph.D.
McMaster University
Hamilton, ON

Maria Ospina, M.Sc.
University of Alberta
Edmonton, Alberta

Karen Schoelles, M.D.
ECRI Institute
Plymouth Meeting, PA

Jodi Segal, M.D., M.P.H.
Johns Hopkins University
Baltimore, MD

Tatyana Shamliyan, M.D., M.S.
University of Minnesota
Minneapolis, MN

Jonathan Treadwell, Ph.D.
ECRI Institute
Plymouth Meeting, PA

Paul Shekelle, M.D., Ph.D.
RAND Corporation
Santa Monica, CA

Peer Reviewers

Jesse Berlin, Sc.D.
Johnson & Johnson Pharmaceutical
Research and Development
Raritan, NJ

Olaf M Dekkers, M.D., Ph.D.
Leiden University Medical Center
Leiden, Netherlands

Mark Helfand, M.D., M.P.H., M.S.
Oregon Health & Science University
Portland, OR

Robert Kane, MD..
University of Minnesota
Minneapolis, MN

P. Lina Santaguida, Ph.D.
McMaster University
Hamilton, ON

Jonathan Treadwell, Ph.D.
ECRI Institute
Plymouth Meeting, PA

Cognitive Testing Respondents

Susanne Hempel, Ph.D.
RAND Corporation
Santa Monica, CA

Rebecca Jerome, M.L.I.S., M.P.H.
Vanderbilt University
Nashville, TN

Dan Jonas, MD, M.P.H.
University of North Carolina at Chapel Hill
Chapel Hill, NC

Jennifer Kraschnewski, M.D.
University of North Carolina at Chapel Hill
Chapel Hill, NC

Michael Matheny, M.D., M.S., M.P.H.
Vanderbilt University
Nashville, TN

J Nikki McKoy, M.P.H.
Vanderbilt University
Nashville, TN

Melissa McPheeters, M.P.H., Ph.D.
Vanderbilt University
Nashville, TN

Sandra Micucci, M.Sc.
Vanderbilt University
Nashville, TN

Brett Nishikawa, M.D.
University of North Carolina at Chapel Hill
Chapel Hill, NC

Content Validity Experts

Wendy Bruening, Ph.D.
ECRI Institute
Plymouth Meeting, PA

Marie Griffin, M.D., M.P.H.
Vanderbilt University
Nashville, TN

Susanne Hempel, Ph.D.
RAND
Santa Monica, CA

Eric Johnson, Ph.D., M.P.H.
Kaiser Permanente Northwest
Portland, OR

Mark Oremus, Ph.D.
McMaster University
Hamilton, ON

Jodi Segal, M.D., M.P.H.
Johns Hopkins University
Baltimore, MD

Tatyana Shamliyan, M.D., M.S.
University of Minnesota
Minneapolis, MN

Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies

Structured Abstract

Objective. To create a practical and validated item bank for evaluating the risk of bias and precision of observational studies of interventions or exposures included in systematic evidence reviews.

Study Design and Setting. The item bank was created based on 1,492 questions included in earlier instruments, organized by the quality domains identified by Deeks and colleagues. Items were eliminated and refined through face validity, cognitive, content validity, and interrater reliability testing.

Results. The resulting RTI item bank, consisting of 29 questions for evaluating the risk of bias and precision of observational studies of interventions or exposures: (1) captures all of the domains critical for evaluating this type of research; (2) is comprehensive and can be easily lifted "off the shelf" by different researchers; (3) can be adapted to different topic areas and study types (e.g., cohort, case control, cross-sectional and case series studies); and (4) provides sufficient instruction to apply the tool to varied topics.

Conclusions. One bank of items, with specific instructions for focusing abstractor evaluations, can be created to judge the risk of bias and precision of the variety of observational studies that may be used in systematic and comparative effectiveness reviews.

Contents

Introduction.....1
Approaches To Assessing the Risk of Bias and Precision of Studies.....3
Organization of the Manuscript.....4
Methods.....5
Compiliation of Potential Questions for Item Bank.....5
Face Validity Testing.....5
Cognitive Testing and Experienced Reviewer Review.....5
Content Validity Testing.....6
Interrater Reliability.....6
Post-Test Revisions.....7
Results.....9
Compiliation of Potential Questions for Item Bank.....9
Face Validity Testing.....11
Cognitive Testing and Experienced Reviewer Review.....11
Content Validity Testing.....11
Interrater Reliability.....36
Post-Test Revisions.....36
Discussion.....38
References.....43

Tables
Table 1. Threats to Precision and Validity.....2
Table 2. Studies Included in Interrate Reliability Testing.....8
Table 3. Validity and Reliability Results and Disposition of Questions.....12
Table 4. Item Bank Questions Mapped to Risk of Bias, Precision, and Methods Domain.....38

Figures
Figure 1. Disposition of Questions for Item Bank.....10

Appendixes
Appendix A. AC1 Statistic
Appendix B. Item Bank for Assessment of Risk of Bias and Precision for Observational Studies of Interventions or Exposures

Introduction

In the past decade, the number of publications included in PubMed has increased at an average annual rate of nearly 6 percent. In 1998, PubMed recorded 467,364 citations for the year; by 2008, the number was 816,597. This steady expansion in the volume of published literature increases the complexity and variability of literature that policymakers, clinicians, and patients need to evaluate to make informed health care choices. Systematic reviews that compare interventions play a key role in synthesizing the evidence.[1] The assessment of the design and conduct of individual studies is central to the endeavor of synthesis. Systematic reviews routinely use assessments of the design and conduct of studies for interpreting results and grading the strength of the body of evidence; they may also use these assessments to select studies for the review, meta-analysis, and for interpreting heterogeneous findings.[2]

Although well-designed and well-implemented randomized controlled trials (RCTs) have long been considered the gold standard for evidence, they frequently cannot answer all relevant clinical questions. RCTs may be unethical,[3] limited in their ability to address harms because of limited size or length of followup,[4] or lack of applicability to vulnerable subpopulations.[5] Studies with other designs—such as quasi-experimental studies that mimic RCT design features with the exception of randomization and observational studies that lack randomization, allocation concealment, blinding of participants and interventionists, and in some instances, control groups— may fill these gaps, but the tradeoff is a wider range of sources of bias, including biases in selection, performance, detection of effects, and attrition; these biases have the potential to alter effect sizes unpredictably.[6,7] The evaluation of these studies requires validated tools to assess the likelihood of bias.

Critical appraisal of study methodology and terminology has varied and is evolving: overlapping terms include quality assessment, assessment of internal validity, risk of bias, and evaluation of study limitations, but a central construct is the believability of the findings. We elect to use the term "assessment of risk of bias and precision" as the most representative of our goals. The purpose of the assessment of bias and precision is to evaluate the degree to which the effects reported by the study represent the "true" causal relationship between exposure and outcome, that is, the accuracy of the estimation. Rothman et al. note that the accuracy of an estimate depends upon its validity (the absence of bias or systematic error) and precision (the absence of random error).[8]

A thorough assessment of threats to the validity and precision of an estimate is critical to understanding the believability of a study. Validity can be improved by reducing bias in selection, performance, detection, measurement, attrition, and reporting, and by adequately addressing the role of confounders. Precision can be improved by increased study size and improved study efficiency.

Table 1 presents a taxonomy of threats to validity and precision and draws upon two well-cited sources: the *Cochrane Handbook for Systematic Reviews of Interventions*[7] and *Modern Epidemiology*.[8] This particular taxonomy does not, however, represent the universe of ways in which sources of bias can be classified. Confounding, for instance, is often separately evaluated from selection bias.

Table 1. Threats to precision and validity*

Threats	Definition
Threats to precision (random error)	
Inadequate study size	-
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.
Threats to validity (systematic error)	
Selection bias	Systematic differences in baseline characteristics of the groups that are compared (for multiple-arm studies) or within the group (for single-arm or cross-sectional studies).e.g. from self-selection of treatments, physician-directed selection of treatments, or demographic characteristics, failure to account for intention-to-treat clinical, or social characteristics. Includes confounding from differential selection before exposure and disease as well as selection bias where exposure and/or disease influence the selection of the participants
Performance bias	Systematic differences in the care provided to participants in the comparison groups other than the intervention under investigation (for multiple-arm studies) or within groups (for single-arm and cross-sectional studies), e.g., variation in delivery of the protocol, difference in co-interventions, inadequate blinding of providers and participants (variation unlikely in observational studies)
Attrition bias	Systematic differences among the comparison groups in the loss of participants from the study (for multiple-arm studies) or within groups (for single-arm and cross-sectional studies) and how they were accounted for in the results, e.g., incomplete followup, differential attrition
Detection bias	Systematic differences in outcomes assessment among the comparison groups(for multiple-arm studies) or within groups (for single-arm and cross-sectional studies, e.g., inadequate assessor blinding, differential outcome assessment)
Reporting bias	Systematic differences between reported and unreported findings, e.g., differential reporting of outcomes or harms, potential for bias in reporting through source of funding
Information bias	Systematic differences caused by measurement errors, e.g., recall bias

* From Rothman et al., 2003[8] and Higgins et al., 2006.[7]

Several reviews of critical appraisal tools have found no gold standard.[9-12] Deeks and colleagues undertook an extensive review of quality appraisal tools of nonrandomized studies. Of 213 identified tools, only 6[13-18] meet their criteria of evaluating 6 core elements of internal validity (creation of the intervention groups, comparability of the groups at the analysis stage, allocation to intervention, similarity of groups for key prognostic characteristics by design, identification of prognostic factors, and the use of case-mix adjustment) and were specifically designed for use in systematic reviews.[10] These tools vary on their coverage of important criteria[10] and vary in their focus on reporting or methods description (that is, questions regarding whether or not the authors reported a particular element of the study in a manuscript) versus judgment of risk of bias (that is, questions regarding whether the conduct of the study altered the believability of the results).

Some current tools such as the Newcastle Ottawa[14] are scales that use an implicit weighting. The use of uniform weights may be difficult to justify in all contexts;[7] for a particular topic, a single flaw might substantially increase risk of bias, but the use of uniform weights would prevent that determination. Additionally, these tools may require modification or may not be applicable for specific designs such as cross-sectional or case series.[15] Also, use of terms such as “adequate” and “appropriate” without adequate guidance may result in differences in interpretation within teams working on the same systematic review. In practice, the idiosyncrasies of topics require and often result in each new review developing its own risk of bias rating tool. These tools may not have adequate instruction for reviewers, leading to inconsistent standards within and across reviews.

Our objective was to create a practical and validated item bank for evaluating the risk of bias of observational studies of interventions or exposures that: (1) captures all of the domains critical for evaluating this type of research; (2) is comprehensive and can be easily lifted "off the shelf" by different researchers; (3) can be adapted to different topic areas and study types (e.g., cohort, case control, cross-sectional and case series studies); and (4) provides sufficient instruction to apply the tool to varied topics. Our risk of bias and precision assessment item bank provides a means to assess the extent to which a study's conclusions can be trusted and identify threats to the accuracy of an estimate.

The item bank is applicable to:

- studies of interventions or exposures and is not designed to evaluate diagnostic studies;
- a variety of observational study design types studies, such as cohort and case-control designs, case series, and cross-sectional designs;
- evaluating internal validity only and does not evaluate external validity (applicability).

Our item bank is appropriate for studies that lack random allocation to an intervention and rely on association between changes or differences in exposure or intervention and changes or differences in an outcome of interest.[19] Applicable designs include studies with controls (cohort and case-control) as well as studies without controls that rely on changes or differences in exposure (cross-sectional and case series).[20] We focused on these specific designs as common examples of observational studies. Although we did not test the reliability of our item bank for other designs, we note that the item bank may be used for other designs as well, with some modifications. For instance, quasi-experimental studies will need to add, in addition to the questions selected from our item bank, questions from RCT appraisal tools on allocation concealment and blinding of patients and interventionists.

The intent of the item bank is to serve as a comprehensive source of questions that have been validated and tested. We anticipate that principal investigators will select specific items based on the needs of the review topic. The specific choice of items for a review will depend upon the potential sources of bias in the studies included for that topic.

Approaches To Assessing the Risk of Bias and Precision of Studies

As noted by Deeks et al., approaches to evaluating the quality of observational studies focus on either the evaluation of bias or on "methods description," that is, the evaluation of the "objective characteristics of each study's methods as they are described by the primary researchers" (Deeks et al., p. 23).[10] Study appraisal based on risk-of-bias lists potential sources of bias (Table 1), relies heavily on judgment, and is supported by transparency in recording the judgment. One constraint of this approach is that threats to validity and precision can occur at various points in the study. Assessing these threats without explicit reference to methods used at each stage of research would require a relatively abstract evaluation and could result in poor interrater reliability. The latter approach of "methods description" is easier to implement because methods for each stage of research tend to correspond well with how manuscripts are written. This approach relies less on judgment[10] but may fall short of evaluating believability. One solution is to use both approaches, using the methods description for each stage of research as the primary framework to facilitate ease of review, but clearly specifying how the design and

conduct of the study at that stage addresses threats to validity and precision. This approach allows some degree of reliance on reporting of results as well as judgment on the part of the reviewer. Our proposed item bank began with this approach to identify questions relevant to each of the 12 domains identified by Deeks et al.: (1) background/context, (2) sample definition and selection, (3) interventions/exposure, (4) outcomes, (5) creation of treatment groups, (6) blinding, (7) soundness of information, (8) followup, (9) analysis comparability, (10) analysis outcome, (11) interpretation, and (12) presentation and reporting.[10] Abstractors review a manuscript primarily to identify sources of bias but look to identify this information in the order that information is typically presented in a manuscript.

Organization of the Manuscript

The remainder of this manuscript describes the methods used in compiling, validating, and testing the reliability of the item bank, followed by results and discussion.

Methods

The project was conducted in two phases. The preliminary period, phase 1, resulted in the compilation of potential questions for the item bank. Phase 2 included face validity testing, cognitive testing, content validity testing, and interrater reliability testing. Our goal was to create an item bank where users choose a relevant subset of the included questions. Nonetheless, during the development process, feedback from our Technical Expert Panel (TEP) and testers indicated that a large bank of items would be a constraint to usability. Therefore, we sought to reduce the number of items in the bank to questions relevant for evaluating risk of bias and precision in observational studies and eliminated potential questions concerning applicability, those that were limited in relevance to RCTs, were not relevant to systematic reviews, overlapped with other questions, or had responses that were uninterpretable in the context of evaluating bias or precision. When possible, we attempted to focus on direct evaluations of bias and precision, and subsumed evaluations of limitations resulting from deficiencies in quality of reporting into these questions. We also indicated when specific questions may be excluded for individual studies or for the body of evidence being evaluated. These deletions and modifications occurred over each of the stages of validity and reliability testing, and as a result, some items did not proceed to next stage of testing; the results section provides additional details on the disposition of specific questions.

Compilation of Potential Questions for Item Bank

During phase 1, we compiled a large number of items which had been used previously in AHRQ-sponsored systematic reviews of the evidence and other instruments to evaluate the quality (risk of bias, precision, and other threats to validity) of individual observational studies.[21-108] Some instruments were identified and forwarded to the team by TEP members.[13,16,109-112] To ensure that items addressed all important domains related to risk of bias and other threats to validity, we sorted items into quality domains identified by Deeks et al.[10] We created a prototype item bank containing questions addressing all relevant domains and corresponding multiple choice response categories. The tool included instructions for use and directions for interpreting individual items for principal investigators and abstractors.

Face Validity Testing

In phase 2, we convened a Technical Expert Panel (TEP) composed of 16 senior staff from across Evidence-based Practice Centers (EPCs) and the Agency for Healthcare Research and Quality (AHRQ). The TEP provided expert input throughout the process of finalizing the item bank including reviewing and advising on the proposed conceptual framework; ensuring that the item bank questions evaluated all the critical domains identified by Deeks et al. listed above; sharing their knowledge of existing instruments that have been used for measuring risk of bias or precision; and evaluating the face validity of an early draft of the item bank containing 60 questions (whether they believed that questions would be interpreted correctly and appeared to measure what they were intended to measure).

Cognitive Testing and Experienced Reviewer Review

A preliminary version of the item bank containing 44 questions was reviewed by 9 potential users who were staff at 6 AHRQ-funded Evidence-based Practice Centers (RAND/RTI

International–University of North Carolina [RTI-UNC], University of Alberta, University of Connecticut, University of Minnesota, and Vanderbilt University). Each reviewer independently participated in cognitive testing of the instrument and responded to questions concerning the readability of particular instructions, questions, and response categories. The cognitive interviewer (at least one of the study principal investigators [PIs] accompanied by a notetaker) quizzed the interviewee to determine whether all portions of the instrument were being read and interpreted in the manner they were intended. In addition, because interviewees were experienced in conducting systematic reviewers, each was asked to also evaluate the instrument in relation to whether it contained sufficient and appropriate questions to obtain information on all critical domains.

Content Validity Testing

Seven TEP members participated in a content validity testing of a 42-item version of the item bank. Content validity raters reviewed each item bank question and determined whether they considered the question to be essential, useful, or not necessary for evaluating study risk of bias and precision. Raters repeated the exercise four times separately in relation to cohort, case control, case series, and cross-sectional studies. Results are summarized through a Content Validity Ratio (CVR) score that describes the extent to which the group of reviewers considered each item to be essential to the operationalization of each of the theoretical domains.[113] The CVR varies from -1.00 to +1.00. A CVR = 0.00 would indicate that half of reviewers considered an item to be essential. For the purposes of this study, we considered a CRV >0 to indicate that an item was essential. All items rated essential were moved forward for interrater reliability; in addition, items for which the majority of the respondents felt that the item was either essential or useful for at least one study design were also moved forward.

Interrater Reliability

We tested the performance of the item bank by conducting interrater reliability testing. Twelve individuals with varying levels of experience in conducting systematic reviews were asked to use the item bank to independently rate 10 studies that had previously been included in a systematic review of the literature (the relevant information that the rater needed to consider in relation to the study was presented in one article). Articles selected were intended to represent a cross-section of topic areas and risk of bias and precision concerns that can arise in observational studies.[114-123] The 10 studies are listed in Table 2. For each study, reviewers were instructed to evaluate all questions included in the item bank in relation to the key questions of the study's original systematic review. The materials provided to reviewers included a copy of the article (study) and summary information from the systematic review including: key questions, key outcomes (benefits and/or harms), any important confounding variables and the conceptual model (analytic framework) included with the review.

Reviewers were asked to complete all questions in a 40-item version of the item bank for each study. The response category for all questions was a multiple-choice format. Because we anticipated that some questions would not be relevant to each study, we included a "not applicable" response category for a number of questions. Reviewers were also encouraged to comment on whether they considered particular questions to be irrelevant for evaluating a study; and to provide descriptive feedback concerning construction of individual questions as well as the instrument overall, including ease of using the instrument to review the study, any study risk

of bias, precision, or other quality-related issues that were not captured by the instrument and the time it took to review each study using the instrument.

The study team calculated summary statistics describing agreement between reviewers including mean percent agreement (and associated standard deviation) and first order agreement coefficients (AC1 statistic) for each question in relation to each study and across studies. The AC1 statistic is a summary measure ranging from -1 (when agreement is zero) to 1 (when agreement is 100 percent), that adjusts results for chance agreement and is considered appropriate when there are multiple raters.[124] Based on previous work by Walter, Eliasziw and Donner (1998), with 10 raters and 10 articles, we calculated that we had at least 80 percent power to detect that our intraclass correlation is significantly different from zero (based on observed intraclass correlations of 0.2). Appendix A presents background information for calculating the AC1. Fleiss' kappa statistics were initially calculated as measures for summarizing the data but are not presented because of the so-called "paradox" of the kappa statistic where one can have high agreement but low kappa scores,[125] raising caution concerning their interpretation.

Because each question included multiple response categories, reliability testing evaluated agreement between raters by comparing the most common response category selected by raters to all other response options for a particular question in relation to each study. Summary statistics by question were calculated across the 12 reviewers and 10 studies. Across all studies, we summarized mean agreement across questions by quality domain and by the two analytic approaches to answering the question (solely determining if specific information is reported in the article or using judgment to evaluate the study's approach to addressing a bias or precision concern). The PIs also reviewed and considered all descriptive comments made by reviewers.

Post-Test Revisions

Based on face validity, cognitive testing, content validity, and inter-rater reliability testing, we either added, deleted, or revised questions (including question syntax, response categories, and instructions). We elected to revise questions when comments elicited during each phase of testing suggested a need for changes to the question, response categories, or instructions.

Table 2. Studies included in interrater reliability testing

Study	Intervention/Exposure Group Methods	Comparison group?	Study Focus	Mean time needed by rater to review study (Range across raters)[1]
Baker (2002) Functional health literacy and the risk of hospital admission among Medicare managed care enrollees	Prospective cohort	No	Health care delivery	47 minutes (25-90)
Coleman FH (1990) Safety and efficacy of combined ritodrine and magnesium sulfate for preterm labor: a method for reduction of complications	Retrospective cohort	Yes, prospective cohort	Medication treatment	51 minutes (20-90)
Crisp AH et al (1992) Long-term mortality in anorexia nervosa. A 20-year followup of the St George's and Aberdeen cohorts	Prospective cohort	No	Disease outcome/harms	52 minutes (29-70)
Daniel (1999) Effectiveness of community-directed diabetes prevention and control in a rural Aboriginal population in British Columbia, Canada	Prospective cohort	Yes, prospective cohort	Community-based intervention	63 minutes (25-90)
De Lieto (2003) Immunohistological detection of insulinlike growth factor type 1 receptor and uterine volume changes in gonadotropin releasing hormone analog-treated uterine leiomyomas	Prospective cohort	Yes, prospective cohort	Medication treatment	43 minutes (20-70)
Fouad et al (1997) A hypertention control program tailored to unskilled minority workers	Prospective cohort	Yes, retrospective cohort	Community, workplace intervention	51 minutes (20-90)
Hendderson (2006) Pregnancy weight gain and risk of neonatal complications	Case control	Yes	Disease outcome/harms	44 minutes (15-60)
Kinney TR (1999) Safety of hydroxyurea in children with sickle cell anemia: HUG-KIDS study, a phase l/ll trial	Prospective cohort	No	Medication harms	39 minutes (25-60)
Schindl (2003) Elective cesarean delivery vs. spontaneous delivery: a comparative experience of birth experience	Prospective cohort	Yes, prospective	Surgery outcomes	49 minutes (38-60)
van Hamm (1997) Maternal consequences of cesarean section. A retrospective study intraoperative and postoperative maternal complications of cesarean section during a 10-year period	Retrospective cohort	Yes, retrospective	Surgery harms	44 minutes (17-90)

[1]Time estimates were not provided by 2 raters; data presented reflects the mean and range for 10 raters.

Results

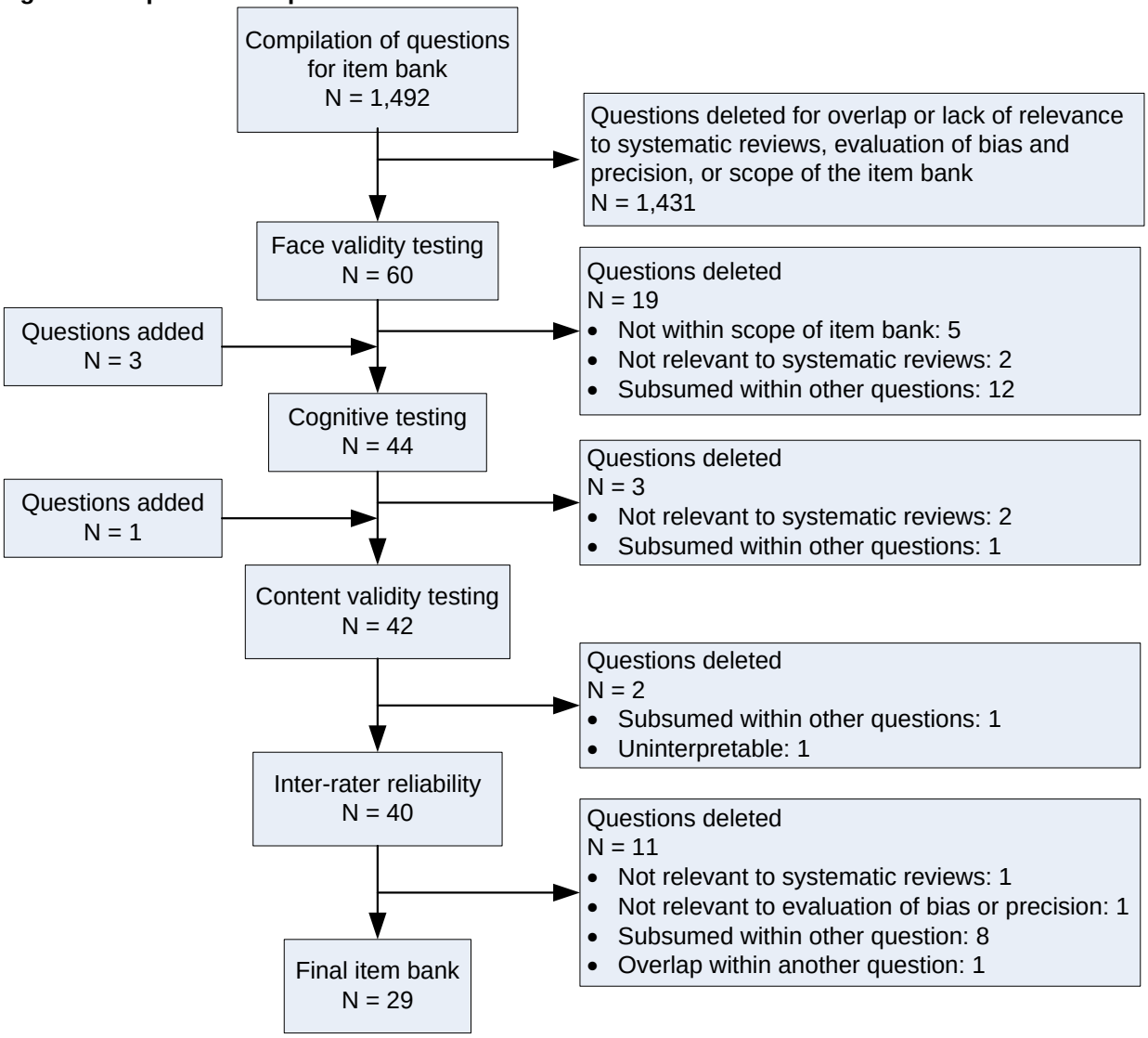
Compilation of Potential Questions for Item Bank

During phase 1 of the project, we reviewed earlier instruments that had been used for evaluating the quality or risk of bias of observational studies and compiled 1,492 items that were available through the published literature and 84[22-24,26-105] of the 90 Agency for Healthcare Research and Quality (AHRQ)-sponsored systematic reviews that had been completed at the start of the project (2007); the remaining 6[25,34,106-109] were either solely focused on RCTs or on evaluation of genetic tests. All items that we collected were categorized into the 12 quality domains (and related subdomains) identified in Deeks et al.[10] We further evaluated the comprehensiveness of the items gathered through that phase by comparing items systematically with those included in other instruments identified through our searches (e.g., Downs and Black[15] and Newcastle Ottawa[14]) or provided to us by our TEP.[13,16,109-112] The intent of casting such a wide net was to ensure a comprehensive set of items.

Many of the 1,492 items were completely or partially redundant. The study team selected 60 items for measuring each of the included domains based on readability and comprehensiveness of instruction (Figure 1). During the development process, we reviewed items and responses and modified wording to ensure that critical domains were represented and to improve readability.

During phase 2 of the project, we crafted explanatory text that can be used by systematic review project directors to individualize the pool of items chosen for a review and assist project directors and abstractors in standardizing item interpretation.

Figure 1. Disposition of questions for item bank



Face Validity Testing

We conducted face validity testing with TEP members. The intent of face validity testing was to ensure that the bank comprehensively included items that could evaluate issues of concerns within each of the methods domains relevant for the evaluation of observational studies and did not include items that the TEP concluded would only be relevant for other study designs. Face validity testing on a 60-item version of the item bank, including instructions for project directors and abstractors, resulted in the elimination of 16 questions. Items eliminated were considered outside scope of the item bank or not relevant to systematic reviews. Two questions were added and several questions were subsumed into other questions (See Figure 1 and Table 3 for details concerning the disposition of specific questions.).[14,15]

Cognitive Testing and Experienced Reviewer Review

On a question-by-question basis, potential users provided feedback on the readability of a 44-item version of the bank including questions, response categories, and instructions. Based on their feedback, we identified particular aspects of individual items that needed revising for greater clarity and direction for the PI and for abstractors separately. Also, because of the substantive expertise of potential users we interviewed, we received comments considering aspects of the item bank design, such as whether particular response categories would provide the distinctions that we were intending and whether additional instructions would be helpful. Reviewers focused on differences between categories such as “no,” “don’t know,” and “not applicable.” Following this review, we eliminated two questions: one because of lack of relevance to systematic reviews and a second because it was subsumed into another question (Figure 1 and Table 3).

Content Validity Testing

An important goal of content validity testing was to identify a core set of questions that experts considered essential in conducting a comprehensive assessment of risk of bias or precision. In the end, a majority of experts considered 24 of a 42-item version of the bank (CRV >0) to be essential across all relevant study designs, 10 items to be either useful or essential across all study designs, and another 6 items to be useful or essential for at least one study design. We eliminated the two questions that the majority of experts considered to be neither essential or nor useful to evaluating risk of bias and precision for any study design type: (1) Is the analysis conducted on an intention-to-treat (ITT) basis? (2) Was the funding for this study derived from a source that does not have a vested interest in its results? The first question was subsumed within another and the second was eliminated from the item bank and not evaluated for interrater reliability (Figure 1 and Table 3).

The source of study funding has been demonstrated to influence the likelihood of publication bias and is therefore an important consideration for evaluating the risk of bias and precision for a body of evidence. Nonetheless, the item was deleted from the evaluation of individual study quality because responses are difficult to interpret in terms of bias. The existence of a source of funding for the study that may have an interest in specific results does not guarantee biased results, nor does the absence of such funding guarantee lack of bias in results.

Although the content validity rating did not easily point to items for exclusion, the results of the content validity rating, in conjunction with interrater reliability scores, helped us determine the need for modifications or deletions.

Table 3. Validity and reliability results and disposition of questions

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
Background/ context	Provision of background information: ***Is the study presented in the context of clinical practice and policy?***	None, study descriptor	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Eliminated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: relevance to systematic reviews This question was deleted based on cognitive testing/expert review because it was determined to be unnecessary for evaluating bias for a systematic review.
	Background/question clearly stated: ***Is the hypothesis/aim/ objective of the study described?***	None, study descriptor	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = -0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = -0.43	**Interrater reliability results:** 94% (83%–100%) 0.88 (0.78–0.98)	Deleted: relevance to systematic reviews This question was deleted because as a study descriptor (reporting of information), content validity testing found the question to generally not be essential for evaluating bias. The question is also unlikely to distinguish risk of bias differences between studies because mean agreement between raters during interrater reliability testing was quite high (94%; almost all responded, "yes").
	Rationale/theoretical framework: ***Was the intervention/ exposure considered within the context of a theoretical framework or causal pathway?***	Selection bias: confounding	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.43 **Cross-sectional:** CVR = -0.71 **Case-control:** CVR = -0.43 **Case series:** CVR = -0.43	**Interrater reliability results:** 52% (42%–83%) 0.10 (-0.01–0.22)	Deleted: relevance to evaluating bias and precision Content validity testing found that the question was not essential across all study designs. The question directly measures reporting and only indirectly suggests risk of bias through failure to account for alternative explanations. It also performed poorly during interrater reliability testing.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
Sample definition and selection	Retrospective/prospective: ***Is the study design prospective, retrospective, or mixed?***	Reporting question related to selection bias; performance bias; detection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.71 **Cross-sectional:** CVR = -1 **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 88% (50%–100%) 0.78 (0.56–1.00)	Retained, considered a second-tier question Content reviewers generally supported retaining this question, particularly in relation to evaluating cohort studies and it performed well in interrater reliability testing. However, the study team acknowledges that increasingly, experts do not necessarily consider one study design to be inherently better than another (see Cochrane Collaboration guidance and Rothman, Greenland & Lash (2008)). Therefore, this is not considered to be a critical question.
	Retrospective/prospective: ***Is the study design appropriate for answering the study's research questions?***	None, overall study quality	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Eliminated	**Cohort:** CVR = NA **Cross-sectional:** CVR = NA **Case-control:** CVR = NA **Case series:** CVR = NA	**Face validity results:** NR **Interrater reliability results:** NR	Deleted: relevance to systematic reviews This question was deleted based on cognitive testing/expert review because it was considered to be relevant for evaluating bias for a systematic review.
	Inclusion/exclusion criteria: ***Are the inclusion/exclusion criteria clearly stated?***	Reporting question, related to selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 1.0 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 1.0 **Case series:** CVR = 0.71	**Interrater reliability results:** 72% (42%–100%) 0.46 (0.28–0.65)	Retained, additional instruction added Interrater reliability was lower in part because there was disagreement between raters concerning whether the correct response was "all" or "partially." We added additional instructions to the PI concerning specification of criteria.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Inclusion/exclusion criteria: ***Are the inclusion/exclusion criteria measured using valid and reliable measures, implemented consistently across all study participants?***	Information bias	**Face validity:** Not evaluated, added after **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = -0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 69% (50%–100%) 0.40 (0.24–0.57)	Retained: additional instruction added during the development process. Interrater agreement varied across studies and so we added direction to the PI to specify important criteria that abstractors need to consider.
	Inclusion/exclusion criteria: ***Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?***	Selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 1.0 **Cross-sectional:** CVR = NA **Case-control:** CVR = 1.0 **Case series:** CVR = NA	**Interrater reliability results:** 61% (33%–92%) 0.30 (0.13–0.46)	Retained Question wording was changed during development process from negative to positive for consistency with majority of other questions. Additional instruction was added to address confusion expressed among interrater reliability raters in relation to single-armed studies.
	Inclusion/exclusion criteria: ***Was the strategy for recruiting participants into the study described?***	Reporting question, related to selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = NA **Case-control:** CVR = NA **Case series:** CVR = NA	**Interrater reliability results:** 68% (42%–92%) 0.37 (0.22–0.53)	Deleted: subsumed in another question Other questions that incorporate this issue as a response category (relevant to reporting) are included in the tool, and so this question considered unnecessary. Also, rater agreement was fairly low, primarily because raters did not agree on the level of detail needed to adequately respond.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Inclusion/exclusion criteria: ***Did the strategy for recruiting participants into the study differ across study groups?***	Performance bias	**Face validity:** Not evaluated **Cognitive testing/ experienced reviewer:** Not evaluated, added after	**Cohort:** CVR = 0.71 **Cross-sectional:** CVR = NA **Case-control:** CVR = Not evaluated **Case series:** CVR = NA	**Interrater reliability results:** 70% (50%–92%) 0.41 (0.27–0.55)	Retained Emphasis changed from negative to positive to be consistent with most other questions; additional detail added to one of the response categories. Clarified in not applicable response category that the question would not apply to studies with one arm.
	Power and sample size: ***Did the authors report conducting a power analysis or some other basis for determining the adequacy of study group sizes for the primary outcome(s) being abstracted?***	Reporting question, related to precision	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = -0.71 **Case-control:** CVR = -0.71 **Case series:** CVR = -0.43	**Interrater reliability results:** 80% (58%-–100%) 0.57 (0.42–0.72)	Deleted: subsumed in another question Content validity ratings supported deleting across study types because it is a measure of precision rather than bias. Because it is a reporting question, it was not considered valuable for evaluating precision although it performed reasonably well in terms of interrater reliability. The tool retains other, more direct questions for evaluating precision.
	Power and sample size: ***Was the sample size sufficiently large to detect a clinically significant difference of 5% or more between groups in at least one primary outcome measure?***	Precision	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = -0.43	**Interrater reliability results:** 55% (33%–75%) 0.14 (-0.02 –0.30)	Retained: additional direction provided to the PI. We have retained this question to address this aspect of precision because it may be important to some reviews. We acknowledge that establishing standards to answer the question can be difficult. Interrater reliability was low for this question, and so we added additional directions for the PI.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Baseline characteristics described: ***Are key characteristics of study participants described?***	Reporting question, related to selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.71 **Case series:** CVR = 0.43	**Interrater reliability results:** 68% (42%–92%) 0.44 (0.28–0.60)	Deleted: subsumed in another question This question was deleted because it is accounted for in responses to questions about controls for differences in baseline characteristics that directly measure bias. Also, raters had difficulty determining which characteristics needed to be described.
Interventions/ Exposure	Clear specification: ***What is the level of detail in describing the intervention or exposure?***	Reporting question, related to performance bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.43 **Case series:** CVR = 0.14	**Interrater reliability results:** 61% (42%–100%) 0.34 (0.19–0.50)	Retained; information added to the PI instructions and response categories Interrater reliability agreement was low, partially because raters differed in their evaluation of the level of detail that they considered sufficient. To address this, additional instruction was added for the PI to specify criteria for abstractors.
	Clear specification: ***What is the level of detail in describing the test reference standard?***	Reporting question related to performance bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: scope This question was deleted based on face validity evaluation by the Technical Expert Panel. It was decided that this item bank would not be used to evaluate diagnostic testing.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Clear specification: ***What is the level of detail in describing the intervention protocol?***	Reporting question related to performance bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel because it was determined to be unnecessary for evaluating bias for a systematic review. The reporting element of this question is incorporated within another question that asks about deviation from protocol.
	Concurrent/concomitant treatment: ***Are concurrent or concomitant treatments described?***	Reporting question, related to performance bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.43 **Case series:** CVR = 0.14	**Interrater reliability results:** 58% (33%–92%) 0.23 (0.07–0.40)	Deleted: subsumed in another question This question was deleted because it was determined by the study team to be unnecessary. The reporting element of this question is incorporated within another question that asks about controlling for concurrent treatment. Also, the question did not perform well because raters had difficulty consistently determining which characteristics needed to be described.
	Description of care for comparison groups: ***Is usual clinical care described?***	Reporting question, related to performance bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = -0.14 **Case-control:** CVR = -0.14 **Case series:** CVR = -0.43	**Interrater reliability results:** 59% (33%–100%) 0.26 (0.07–0.45)	Deleted: overlap This question was deleted because content validity evaluation generally found it to not be essential. Also, agreement between raters was low which may be related to an inability to distinguish usual clinical care from concurrent treatment and to determine the level of detail necessary to report.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
Outcomes	Clear specification: ***Are the potential outcomes, including harms, pre-specified by the researchers?***	Reporting bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 66% (33%–100%) 0.39 (0.23–0.56)	Retained: additional information added to the question and instruction added for the PI Question was edited to address content expert and rater concerns that the question needed more instruction for case-control and that harms need not always be pre-specified. Some raters may have had difficulty answering this question because outcomes of interest were not specified.
	Clear specification: ***Are the primary outcomes described?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel which concluded that it was unnecessary for evaluating bias for a systematic review. The response to this question is included as a response to a question on assessment of outcomes.
	Clear specification: ***Are the secondary outcomes described?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel which concluded that it was unnecessary for evaluating bias for a systematic review. The response to this question is included as a response to a question on assessment of outcomes.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Clear specification: ***Are harms or anticipated adverse events described?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel which concluded that it was unnecessary for evaluating bias for a systematic review. The response to this question is included as a response to a question on assessment of outcomes.
	Objective and/or reliable: ***Do the researchers report the time points for measurement of the primary outcomes?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel and the study team that this question should be addressed as part of the more general question on clear specification of outcomes.
	Objective and/or reliable: ***Do the researchers report the time points for measurement of the secondary outcomes?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel and the study team that this question should be addressed as part of the more general question on clear specification of outcomes.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
Creation of treatment groups	Random allocation: ***If participants are randomized, were appropriate randomization methods used?***	Selection bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: scope This question was deleted based on expert input through face validity evaluation by the Technical Expert Panel that the evaluation of randomization was not necessary for observational studies.
	Clear specification: ***Are the criteria for assignment to study groups described?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel which concluded that it was unnecessary for reporting bias for a systematic review. Consideration of groups is evaluated through the question on uniform application of criteria to comparison groups.
	Allocation ***Is assignment made to study groups randomly?***	Selection bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: scope This question was deleted based on expert input through face validity evaluation by the Technical Expert Panel that the evaluation of randomization was not necessary for observational studies.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Allocation ***Is an explicit case/comparison definition reported?***	Reporting question related to selection bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer results:** NR	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was considered unnecessary because an evaluation of the choice of the case and comparison groups is accomplished through the question clear statement of inclusion/exclusion criteria.
	Allocation ***Is the selection of the comparison group appropriate?***	Selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 1.0 **Cross-sectional:** CVR = NA **Case-control:** CVR = Not evaluated **Case series:** CVR = NA	**Interrater reliability results:** 69% (42%-100%) 0.42 (0.25–0.59)	Retained, additional information added to help clarify the question. Depending on the review, appropriate may take into account feasibility or ethics. Content experts considered this question to be essential for cohort studies.
	Any attempt to balance: ***Any attempt to balance the allocation between the groups?***	Selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = NA **Case-control:** CVR = Not evaluated **Case series:** CVR = NA	**Interrater reliability results:** 63% (42%–92%) 0.32 (0.17–0.46)	Retained: additional clarifying information added to response categories and additional instruction provided to PI We combined "no" and "cannot determine" response categories because there is no appreciable distinction between the categories and raters could not distinguish between the two in a consistent manner. In response categories, we also accounted for designs that address differences between groups through post-hoc approaches such as multivariate analysis.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Clear specification: ***Have researchers reported the possibility of participants having received an unintended intervention?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel which concluded that it was unnecessary for reporting bias for a systematic review. Contamination, as it relates to performance bias is evaluated more directly through a subsequent question concerning unintended exposure.
	Contamination ***Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?***	Performance bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = -0.14 **Case series:** CVR = -0.43	**Interrater reliability results:** 62% (42%–92%) 0.34 (0.21–0.48)	Retained: Response categories edited based on inconsistencies in responses and comments obtained during interrater reliability testing. We added a response category of "partially" and combined "no" and "don't know." The "don't know" response would identify instances of insufficient reporting.
	Contamination ***Did variation from the study protocol compromise the conclusions of the study?***	Performance bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = Not evaluated **Case-control:** CVR = -0.14 **Case series:** CVR = -0.14	**Interrater reliability results:** 51% (42%–67%) 0.21 (0.15–0.27)	Retained: Edited to clarify that the protocol of interest is the study protocol and not an analytic protocol. The question was respecified so that it is written in relation to intervention studies only.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Contamination: ***Is adherence to the protocol reported?***	Reporting question related to performance bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on face validity evaluation by the Technical Expert Panel which concluded that it was unnecessary for reporting bias for a systematic review. The answer to this question is incorporated in a response to the question on whether the execution of the study varied from the protocol.
Blinding	Blind administration ***Were study participants blinded to their group assignment?***	Detection bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: scope This question was deleted based on expert input through face validity evaluation by the Technical Expert Panel that this was not relevant for evaluating bias in observational studies.
	Blind administration: ***Are those administering the intervention blinded to the study assignment or exposure status of participants?***	Detection bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: scope This question was deleted based on expert input through face validity evaluation by the Technical Expert Panel that this was not relevant for evaluating bias in observational studies.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Blind outcome assessment ***Were the outcome assessors blinded to the intervention or exposure status of participants?***	Detection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = Not evaluated **Case-control:** CVR = Not evaluated **Case series:** CVR = 0.43	**Interrater reliability results:** 69% (50%–100%) 0.46 (0.28–0.64)	Retained: Edited to clarify that there may be instances in which the outcome assessment cannot be blinded. Based on rater lack of agreement, we clarified that the response "not applicable" applies to studies in which an assessor cannot be blinded and provided clarifying instructions for the PI.
Soundness of information	Source of information about interventions/ exposure ***Are interventions/ exposures assessed using valid and reliable measures, implemented consistently across all study participants?***	Information bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.71 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.43 **Case series:** CVR = 0.14	**Interrater reliability results:** 82% (42%–100%) 0.65 (0.48–0.82)	Retained: Additional instructions provided for the PI We added instruction for the PI concerning criteria for evaluating what may be valid and reliable measures.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Source of information re outcomes ***Are primary outcomes assessed using valid and reliable measures, implemented consistently across all study participants?***	Information bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.71 **Cross-sectional:** CVR = 0.71 **Case-control:** CVR = 0.71 **Case series:** CVR = 0.14	**Interrater reliability results:** 93% (75%–100%) 0.87 (0.76–0.98)	Retained: Additional instructions provided for the PI We added instruction for the PI concerning criteria for evaluating what may be valid and reliable measures, consistent with the edits for the item concerning interventions/exposures.
Follow-up	Equality of length of follow-up for participants ***In cohort studies, is the length of follow-up different between the groups, or in case-control studies, is the time period between the intervention/exposure and outcome the same for cases and controls?***	Attrition bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.71 **Cross-sectional:** CVR = NA **Case-control:** CVR = 0.71 **Case series:** CVR = NA	**Interrater reliability results:** 57% (42%–92%) 0.19 (0.05–0.33)	Retained: Question and response categories edited We changed the wording of the question from "different" to "same" for all groups, based on comments from raters that the question was confusing. We combined "no" and "don't know," categories, which we thought equivalent.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Length of followup adequate ***Is the length of time following the intervention/exposure sufficient to support the conclusions of the study regarding primary outcomes?***	Attrition bias	**Face validity results:** Asked as one question concerning all outcomes, Retained and changed to 2 questions (primary outcomes and harms) **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = NA **Case-control:** CVR = 0.71 **Case series:** CVR = 0.43	**Interrater reliability results:** 58% (33%–83%) 0.27 (0.16–0.38)	Retained: Edited for clarity; benefits and harms evaluation combined into one question. Based on rater uncertainty, we added additional instruction to PIs to specify outcomes. Benefit and harm outcomes were combined into one question. The question was taken out of the context of what study authors may have concluded, and instead, concerns whether the reviewer considers the follow-up period sufficient to support the results as measured

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Length of followup adequate ***Is the length of time following the intervention/exposure sufficient to support the conclusions of the study regarding harms?***	Attrition bias	**Face validity results:** Asked as one question concerning all outcomes, retained and changed to 2 questions (primary outcomes and harms) **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = NA **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 54% (33%–83%) 0.18 (0.06–0.30)	Deleted: subsumed in another question We deleted this question to create one question concerning sufficient follow-up for both benefits and harms. The PI will need to specify sufficiency of follow-up for specific outcomes if it differs across outcomes.
	Completeness of follow-up ***Are all participants in all study arms accounted for in follow-up?***	Attrition bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question We deleted this question following face validity evaluation by the Technical Expert Panel. We determined that this question would be answered by the question that requires the abstractor to determine the attrition rate.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Completeness of follow-up ***Did attrition from any group exceed [x] percent?***	Attrition bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = NA **Case-control:** CVR = -0.43 **Case series:** CVR = Not evaluated	**Interrater reliability results:** 51% (42%–67%) 0.13 (0.08–0.18)	Retained: Additional instruction provided. While this was not a good performing question based on content validity or interrater reliability testing, due to our concerns about evaluating attrition bias, we are retaining this question. We acknowledge that it can be a difficult question to answer; abstractors were particularly unsure about how to evaluate retrospective studies.
	Completeness of follow-up ***Did attrition differ between groups by more than 20 percentage points?***	Attrition bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = NA **Case-control:** CVR = -0.14 **Case series:** CVR = NR	**Interrater reliability results:** 58% (42%–83%) 0.23 (0.10–0.37)	Retained: Additional instruction provided We provided additional instruction concerning how to evaluate retrospective studies.
Analysis comparability	Analysis of baseline comparability ***Are baseline characteristics similar between groups?***	Selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = NA **Case-control:** CVR = 0.14 **Case series:** CVR = NA	**Interrater reliability results:** 54% (33%–92%) 0.22 (0.08–0.35)	Deleted: subsumed in another question This issue is evaluated through the question that concerns whether baseline differences are addressed (controlled for) through the analysis

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Analysis of baseline comparability ***Does the analysis control for baseline differences between groups?***	Selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = NA **Case-control:** CVR = 0.43 **Case series:** CVR = NA	**Interrater reliability results:** 65% (33%–100%) 0.33 (0.14–0.51)	Retained: Several response categories were modified for greater clarity.
	Identification of prognostic factors (effect modifiers and confounders) ***Does the study identify important confounding variables and effect modifiers?***	Reporting question related to selection bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.71 **Case series:** CVR = -0.14	**Interrater reliability results:** 62% (42%–92%) 0.32 (0.19–0.45)	Deleted: subsumed in another question The answer to this question is part of the response category to a subsequent question on whether confounding variables are accounted for in the design.
	Identification of prognostic factors (effect modifiers and confounders) ***Are confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?***	Information bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = -0.14	**Interrater reliability results:** 55% (33%–92%) 0.20 (0.04–0.37)	Retained: Question and response categories edited to improve clarity This question performed inconsistently across studies and so we added direction to help ensure that adequate specification is provided by the PI.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Case-mix adjustment: ***Were the important confounding and modifying variables taken into account in the design and analysis?***	Selection bias: confounding	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = -0.14 **Case-control:** CVR = 0.43 **Case series:** CVR = -0.71	**Interrater reliability results:** 65% (33%–83%) 0.35 (0.21–0.49)	Retained: Question and response categories edited to improve clarity This question performed inconsistently across studies and so we added direction to help ensure that adequate specification is provided by the PI.
Analysis outcome	Intention-to-treat analysis ***Is the analysis conducted on an intention-to-treat (ITT) basis?***	Attrition bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = NA **Case-control:** CVR = NA **Case series:** CVR = NA	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question The more important question concerning the evaluation of the impact of loss to follow-up is evaluated through the next question.
	Intention-to-treat analysis ***Is the impact of loss to follow-up assessed?***	Attrition bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 60% (42%–83%) 0.26 (0.14–0.39)	Retained: Response categories were edited to improve clarity and address inconsistencies across study raters.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Appropriate analytic methods ***Are findings for all primary outcomes reported?***	Reporting bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.43 **Case series:** CVR = 0.43	**Interrater reliability results:** 79% (50%–100%) 0.63 (0.50–0.76)	Retained: Emphasis of the question changed from reconciliation between what authors said they would and did do to what could be reasonably expected. The emphasis of this question was changed to more directly evaluate whether any important findings were omitted from the publication, in contrast to what the study authors say they intended to evaluate. The PI will need to specify primary outcomes for abstractors.
	Clear specification ***Are the statistical approaches for analyzing the data reported?***	Reporting question related to reporting bias	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on Technical Expert Panel member face validity review because it was determined to be unnecessary for evaluating bias for a systematic review. The evaluation of the precision of estimate based on the study methodology is evaluated through a subsequent question concerning the appropriateness of the methods used in the study.
	Appropriate analytic methods ***Are the statistical methods used to assess the primary outcomes appropriate to the data?***	Precision	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.71 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.71 **Case series:** CVR = 0.14	**Interrater reliability results:** 66% (42%–92%) 0.39 (0.28–0.51)	Retained: Expanded to more comprehensively capture assessment of primary outcomes to include reporting of random variability This question was combined with a subsequent question concerning reporting of the random variability of the outcome.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Appropriate analytic methods ***Have all important harms or adverse events that may be a consequence of the intervention/exposure been reported?***	Reporting bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = -0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 64% (42%–83%) 0.33 (0.22–0.45)	Retained: Emphasis of the question changed from reconciliation between what authors said they would and did do to what could be reasonably expected. The emphasis of this question was changed to more directly evaluate whether any important findings were omitted from the publication, in contrast to what the authors say they intended to evaluate. To reduce the abstractor burden, the PI needs to specify what the abstractors should be looking for.
	Appropriate analytic methods ***Are the statistical methods used to assess the main harm or adverse event outcomes appropriate to the data?***	Precision	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.43 **Case series:** CVR = -0.14	**Interrater reliability results:** 66% (50%–83%) 0.31 (0.21–0.41)	Retained: Expanded to include reporting of random variability to more comprehensively capture assessment of harms: parallels question concerning methods used to report benefits. This question was combined with a subsequent question concerning reporting of the random variability of the outcome.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Appropriate analytic methods ***For cohort studies only, if the outcome has a greater than 10 percent prevalence, is the risk ratio and relative risk calculated directly (not using logistic regression)?***	Precision	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Eliminated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: subsumed in another question This question was deleted based on expert review/cognitive interviews because it was determined that the information is captured through the more general questions concerning appropriate analytic methods.
	Appropriate analytic methods: ***Does the study appropriately report estimates of the random variability in the data for the primary outcomes?***	Reporting question related to precision	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.43 **Cross-sectional:** CVR = 0.43 **Case-control:** CVR = 0.43 **Case series:** CVR = 0.43	**Interrater reliability results:** 66% (42%–92%) 0.38 (0.23–0.53)	Deleted: subsumed in another question Information that would have been obtained through this question can be captured through other analytic questions.
Interpretation	Appropriately based on results ***Are conclusions supported by results with possible biases and limitations taken into consideration?***	Not Applicable, overall study quality	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = 0.14 **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 66% (42%–92%) 0.32 (0.17–0.48)	Retained: edited to capture the abstractor's evaluation of the overall risk of bias and precision of the study and not what was intended by the study's authors. This question provides the abstractor's evaluation of the risk of bias and precision of the study overall. The wording of the question was changed for greater simplicity and understanding.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	Methods Subdomain and Assessment Question	Dimension of Bias or Precision Evaluated	Face Validity and Cognitive Testing/ Experienced Reviewer Results	Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias	Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)	Disposition Rationale
	Interpretation in context ***Are results interpreted appropriately based on study design and statistical analysis?***	Not applicable, overall study quality	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: relevance to systematic reviews Based on expert review, this question was determined to not be relevant for the purposes of conducting a systematic review.
	Interpretation in context ***Are study conclusions presented in the context of prior research?***	Not applicable, overall study quality	**Face validity results:** Eliminated **Cognitive testing/ experienced reviewer:** Not evaluated	**Content validity:** Not evaluated	**Interrater reliability:** Not evaluated	Deleted: relevance to systematic reviews Based on expert review, this question was determined to not be relevant for the purposes of conducting a systematic review.
Presentation and reporting	Completeness, clarity and structure ***Is the source of funding identified?***	Reporting question related to reporting bias	**Face validity results:** Retained **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = 0.14 **Cross-sectional:** CVR = Not evaluated **Case-control:** CVR = 0.14 **Case series:** CVR = 0.14	**Interrater reliability results:** 90% (83%–100%) 0.76 (0.69–0.83)	Retained. Agreement between raters during interrater reliability was high for this question and so question was retained without any changes.

Table 3. Validity and reliability results and disposition of questions (continued)

Methods Domain	**Methods Subdomain and Assessment Question**	**Dimension of Bias or Precision Evaluated**	**Face Validity and Cognitive Testing/ Experienced Reviewer Results**	**Content Validity Results by Study Design Type: Percentage of Experts Who Considered the Question Essential for Evaluating Bias**	**Interrater Reliability Results: Mean Agreement (range) AC1 Statistic (conditional confidence interval)**	**Disposition** **Rationale**
	Completeness, clarity and structure ***Was the funding for this study derived from a source that does not have a vested interest in its results?***	Reporting question related to reporting bias	**Face validity results:** NR **Cognitive testing/ experienced reviewer results:** Retained	**Cohort:** CVR = -0.14 **Cross-sectional:** CVR = -0.14 **Case-control:** CVR = -0.14 **Case series:** CVR = -0.14	**Interrater reliability:** Not evaluated	Deleted: interpretability Added following initial question development but later eliminated. Content validity rating and further discussion with experts recommended that this question be deleted because of the difficulty in identifying on a case-by-case basis organizations that may have a vested interest.

CVR=Content Validity Ratio; NA=not applicable

Interrater Reliability

Interrater reliability testing was conducted on an item bank of 40 questions. Table 2 presents descriptive information on the 10 studies included in the testing, by methodological approach (9 cohort and 1 case-control study), whether the study included a comparator group (6 studies) and type of study (i.e., treatment, harms, and disease outcomes). We also present the average and range of time it took raters to evaluate the risk of bias and precision of a study using the item bank. Overall, it took raters 48 minutes per study, ranging from 17 to 90 minutes.

Table 3 presents details of results by question including the disposition of each question. Overall, the mean AC1 score per question was relatively low, 0.38 and ranged from 0.10 to 0.88. The mean percent agreement across all items was 66 percent. We found that percent agreement varied by domain, from a high of 90 percent for questions concerning presentation and reporting and 88 percent for those concerning soundness of information to a low of 56 percent for questions concerning followup and 59 percent for those concerning interventions and exposures (results not shown). We further tested if agreement varied significantly between questions that concerned identifying whether specific information was reported in an article and the remaining questions that required more complex judgment on the part of the reviewer. We found that overall, raters agreed 70 percent of the time on their responses to reporting questions and 64 percent on questions requiring judgment ($P = 0.09$).

Poor results from interrater reliability testing resulted in no clear patterns or conclusions. As a result, we did not eliminate any questions based on this stage of testing. Instead, we used the interrater reliability results to identify and revise questions that performed poorly and to add instruction to PIs to help abstractors interpret questions more clearly.

Post-Test Revisions

Based on face validity, cognitive testing, content validity, and interrater reliability testing and study team evaluation, we either deleted questions that were considered unnecessary or revised questions (including question syntax, response categories, and instructions). The original 60-item instrument was reduced to 44 items following face validity testing, 42 items following cognitive testing, and 40 items following content validity testing which were used for interrater reliability testing. Following all testing, the final review of items resulted in a bank of 29 items.

Reasons for deletion of questions include lack of relevance to systematic reviews (one question), lack of relevance to evaluation of bias or precision (one question), and overlap with other questions (one question). When possible we collapsed issues of reporting on a specific source of bias or precision within the response categories for direct evaluation of that source of bias or precision (eight questions). For instance, we deleted a reporting question relating to selection bias ("did the authors report differences in baseline characteristics?") but added a response category within the questions—"did the authors control for differences in baseline characteristics?"—to account for those who reported no differences. (Figure 1 and Table 3 present the disposition of all questions.).

The final version of the RTI Item Bank is presented in Appendix B. The bank contains 29 questions, multiple-choice response categories and extensive instructions for PIs and abstractors to assist them in developing criteria for considering the issue being investigated by the question. The order of the questions in the item bank is according to the study domain structure presented by Deeks and colleagues and is generally intended to allow the reviewer to consider the various risks of bias and precision issues of a study according to the presentation order of a manuscript.

Table 4 maps each question to the risk of bias or precision and methods description and also lists relevant study designs.

Because of the integral nature of reporting to evaluating risk of bias and precision, the item bank evaluates some elements of risk of bias and precision through a cluster of questions. For instance, questions about selection bias in the creation of the sample require questions about whether inclusion/exclusion criteria were reported and measured appropriately before the reviewer can judge whether or not they were applied equally to all arms of the study (Questions 2, 3, and 4). The PI's role in systematic review is to determine the appropriate mix of questions necessary for evaluation the dominant risks of bias and threats to precision in the studies being reviewed.

Discussion

With the increasing use of observational studies in evidence synthesis, systematic reviewers have a greater burden of evaluating risk of bias and precision to identify the effects of these potential concerns on study results. The evaluation of study risk of bias and precision is essentially a subjective exercise, requiring judgments by a reviewer. Nonetheless, this exercise is the only means of evaluating the degree to which a study's results can be believed and is a critical step on the pathway to evaluating the strength of a body of evidence. Our item bank builds on and extends the efforts of previous instruments to (1) create an evaluation tool that is specifically designed to work within the larger context of systematic review methodology and tasks; (2) explicitly focus on believability of the study rather than applicability; (3) comprehensively consider the elements that support believability; and (4) promote transparency and consistency of judgment between pairs of reviewers working on a single review and across reviews, particularly when customization is needed for the specific topic.

Our item bank is intended to be used to interpret the believability of individual studies, but just as importantly, to create building blocks for evaluating the risk of bias and precision for the body of evidence. Systems to grade the strength of a body of evidence such as GRADE and the AHRQ strength of evidence approach use an overall assessment of risk of bias as one key element; other separate elements include applicability and precision. Commonly used instruments such as the Newcastle Ottawa scale[14] and Downs and Black[15] include questions on all three areas: risk of bias, precision, and external validity (applicability) within their ratings for individual studies. These instruments identify questions that evaluate external validity but not questions related to precision. When all these items are included within a rating scale, as in the Newcastle Ottawa scale, the results cannot be used as components for judging the strength of the body of evidence without some manipulation: the external validity and precision elements need to be removed from the overall scores to isolate the risk of bias for a study.

Our item bank focuses explicitly on believability, that is, risk of bias and precision. It excludes applicability entirely. It includes questions on precision to allow for a variety of approaches to analysis. Systematic reviews that rely on meta-analyses may not need to evaluate study-specific elements of precision, particularly sample size and appropriate statistical analysis. Systematic reviews that cannot pool estimates because of heterogeneity of results may choose to include evaluations of precision in addition to risk of bias. The item bank identifies the precision questions clearly so that reviewers can judge whether or not to evaluate those elements.

Table 4. Item bank questions mapped to risk of bias, precision, and methods domain

Methods Domain	Precision	Selection Bias/ Confounding	Performance Bias	Attrition Bias	Detection Bias	Reporting Bias	Information Bias	Overall believability	Total N of items
Background/ context	-	-	-	-	-	-	-	-	0
Sample definition and selection	•Q6 (CH, CC, CS, XS)	•Q1 (CH, CC, CS) •Q2 (CH, CC, CS, XS) •Q4 (CH, CC)	•Q1 (CH, CC, CS) •Q5 (CH, CC)	-	•Q1 (CH, CC, CS)	•Q1 (CH, CC, CS)	•Q3 (CH, CC, CS, XS)	-	6
Interventions/ exposure	-	-	•Q7 (CH, CC, CS, XS)	-	-	-	-	-	1
Outcomes	-	-	-	-	-	•Q8 (CH, CC, CS, XS)	-	-	1
Creation of treatment groups	-	•Q9 (CH, CC) •Q10 (CH, CC)	•Q11 (CH, CC, CS, XS) •Q12 (CH, CC, CS, XS)	-	-	-	-	-	4
Blinding	-	-	-	-	•Q13 (CH, CC, CS, XS)	-	-	-	1
Soundness of information	-	-	-	-	-	-	•Q14 (CH, CC, CS, XS) •Q15 (CH, CC, CS, XS)	-	2
Follow-up	-	-	-	•Q16 (CH, CC) •Q17 (CH, CC, CS) •Q18 (CH, CC, CS) •Q19 (CH, CC)	-	-	-	-	4

Table 4. Item bank questions mapped to risk of bias, precision, and methods domain (continued)

Methods Domain	Precision	Selection Bias/ Confounding	Performance Bias	Attrition Bias	Detection Bias	Reporting Bias	Information Bias	Overall believability	Total N of items
Analysis comparability	-	•Q20 •Q22 (CH, CC, CS, XS)	-	-	-	-	•Q21 (CH, CC, CS, XS)	-	3
Analysis outcome	•Q25 (CH, CC, CS, XS) •Q27 (CH, CC, CS, XS)	-	-	•Q23 (CH, CS)	-	•Q24 (CH, CC, CS, XS) •Q26 (CH, CC, CS, XS)	-	-	5
Interpretation	-	-	-	-	-	-	-	•Q28 (CH, CC, CS, XS)	1
Presentation and reporting	-	-	-	-	-	•Q29 (CH, CC, CS, XS)	-	-	1
Total*	3	7	5	5	2	5	4	1	29

*=number of items sum to greater than total number of items because some items relate to multiple risks of bias; CH=cohort; CC=case-control; CS=case series, XS=cross-sectional; N=number

Our item bank includes nearly all the domains and elements for evaluating observational studies identified by West et al.,[9] in one of three ways: within questions, within response categories to the questions, or within instructions to interpreting the questions. We did not include two elements from the West et al. list of evaluation requirements within our item bank: (1) the study includes clearly focused and appropriate questions and (2) use of concurrent controls. We judged, in consultation with our technical experts and other users, that the former question is not relevant to evaluating bias and precision; our intent is to judge the believability of study results that are relevant to the goals of the systematic review rather than to its own intent. We did not include the latter question because we consider the issue of concurrent controls as a matter of design. The intent of our item bank is not to evaluate the believability of study results based on the type of study design used, rather, we intend for responses to the items in the bank to identify design features (or the lack of features) that could increase the risk of bias.

The wide array of designs and associated risks of biases in observational studies imply that systematic reviewers will need to customize their review form to concentrate on the most critical selection of items for the topic at hand and establish minimum standards for addressing these items. An important contribution of this instrument is that we provide choices through a comprehensive array of items rather than a fixed menu of required elements through an instrument. We provide instruction to teams (with separate instructions to scientific leads and other team members) where customization may be required.

We note the overall low interrater reliability scores we obtained through our testing: one likely reason is that we did not develop instructions for each of the studies on how to customize standards for that particular review. By design, our goal was to document experience using the item bank on a broad range of studies. Our results are also limited by this range: unlike typical systematic review teams that share a common understanding of a single review topic, our raters, themselves with varied backgrounds, were asked to rate studies for 10 different topics.

Although content experts may see value in evaluating multiple nuances related to threats to bias, the detailed consideration of each of those concerns may not always be practical. Our reviewers took an average of approximately 48 hours to complete the 10-study risk of bias and precision review. Instead, a more practical approach is to identify the most critical threats to validity and precision in a body of evidence and then select questions, perhaps no more than 10–15, that can evaluate the concern.

Next steps in evaluating this item bank include the identification of specific biases for which different study designs are most at risk. This exercise will help to identify a core set of questions can be identified that must be used consistently across studies. A second area for research is the assessment of customized questions (that is, the selection of items and standards for specific items) and inter- and intrarater reliability for teams of reviewers working on the same topic. A key consideration in such testing will be the selection of studies with known serious issues in design (such as the selection of prevalent users rather than new users) in order to assess the ability of the instrument to identify potential sources of bias. A third urgent task is to evaluate the empirical basis for each item by measuring the correlation between responses to specific items and effect sizes as means of further culling the item bank.

Our item bank can be used in a variety of systematic reviews as further development of the item bank continues. Although it does not include questions on adequacy of randomization generation, concealment of allocation, and blinding of participants and interventionists because these items are specific to RCTs, many questions in our item bank apply to RCTs as well. From a practical perspective, systematic reviews that include RCTs as well as non-RCT designs may

find it useful to evaluate all study designs with the same questions in order to reduce variability for those items. It will also permit a direct comparison on these items between RCTs, nonrandomized and observational studies. The generation of resulting data may be able to help with a critical question when considering resource allocation for research: when are RCTs essential for answering research gaps?

References

1. Lohr KN. Emerging methods in comparative effectiveness and safety: symposium overview and summary. Med Care. 2007 Oct;45(10 Supl 2):S5-S8.
2. Viswanathan M. Systematic Review: Assessing the Quality of Individual Studies. Rockville, MD: Agency for Healthcare Quality and Review; 2010. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/slides/?pageAction=displaySlides&tk=24.
3. 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):1112-19.
4. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol. 2010;63(5):502-12.
5. Agency for Healthcare Research and Quality. Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews, Version 1.0. Rockville, MD: Agency for Healthcare Research and Quality; 2007. Available at: http://effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf.
6. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. Br Med J. 2001;323:42–46.
7. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.2. London: The Cochrane Collaboration; 2009. Available at: www.cochrane-handbook.org.
8. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott, Williams, & Wilkins; 2008.
9. West SL, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. Rockville, MD: Agency for Healthcare Research and Quality, 2002. AHRQ Publication No. 02-E016.
10. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):iii-x, 1-173.
11. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007 Jun;36(3):666-76.
12. Katrak P, Bialocerkowski AE, Massy-Westropp N, et al. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol. 2004;4:22.
13. Thomas H. Quality assessment tool for quantitative studies. Effective Public Health Practice Project. Toronto: McMaster University.
14. Wells G, Shay B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analysis. Ottawa: University of Ottawa; Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/slides/?pageAction=displaySlides&tk=24.
15. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Commun Health. 1998;52:377–84.
16. Zaza S, Wright-De Aguero LK, Briss PA, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med. 2000 Jan;18(1 Suppl):44-74.
17. Cowley DE. Prostheses for primary total hip replacement. A critical appraisal of the literature. Int J Technol Assess Health Care. 1995 Fall;11(4):770-8.
18. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics. 1989 Nov;84(5):815-27.

19. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000 Apr 19;283(15):2008-12.

20. Peipert JF, Phipps MG. Observational studies. Clin Obstet Gynecol. 1998 Jun;41(2):235-244.

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

22. Lau J, Ioannidis JPA, Balk E, et al. Evaluation of Technologies for Identifying Acute Cardiac Ischemia in Emergency Departments. Evidence Report/Technology Assessment Number 26. (Prepared by The New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality, May 2001. AHRQ Publication No. 01-E006.

23. Sharma M, Clark H, Armour T, et al. Acute Stroke: Evaluation and Treatment. Evidence Report/Technology Assessment No. 127 (Prepared by the University of Ottawa Evidence-based Practice Center under Contract No. 290-02-0021). Rockville, MD: Agency for Healthcare Research and Quality, July 2005. AHRQ Publication No. 05-E023-2.

24. Jadad AR, Boyle M, Cunningham C, et al. Treatment of Attention-Deficit/Hyperactivity Disorder. Evidence Report/Technology Assessment No. 11 (Prepared by McMaster University under Contract No. 290-97-0017). Rockville, MD: Agency for Healthcare Research and Quality, November 1999. AHRQ Publication No. 00-E005.

25. Myers ER, Bastian LA, Havrilesky LJ, et al. Management of Adnexal Mass. Evidence Report/Technology Assessment No. 130 (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-02-0025). Rockville, MD: Agency for Healthcare Research and Quality, February 2006. AHRQ Publication No. 06-E004.

26. Lau J, Balk E, Rothberg M, et al. Management of Clinically Inapparent Adrenal Mass. (Evidence Report/Technology Assessment No. 56 (Prepared by New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality, May 2002. AHRQ Publication No. 02-E014.

27. Chou R, Fu R, Carson S, et al. Empirical Evaluation of the Association Between Methodological Shortcomings and Estimates of Adverse Events. Technical Review No. 13 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024). Rockville, MD: Agency for Healthcare Research and Quality, October 2006. AHRQ Publication No. 07-0003.

28. Hardy M, Coulter I, Venuturupalli S, et al. Ayurvedic Interventions for Diabetes Mellitus: A Systematic Review. Evidence Report/Technology Assessment No. 41 (Prepared by Southern California Evidence-based Practice Center/RAND under Contract No. 290-97-0001). Rockville, MD: Agency for Healthcare Review and Quality, 2001. AHRQ Publication No. 01-E040.

29. Balk E, Chung M, Raman G, et al. B Vitamins and Berries and Age-Related Neurodegenerative Disorders. Evidence Report/Technology Assessment No. 134. (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality, April 2006. AHRQ Publication No. 06-E008.

30. Catlett C, Perl T, Jenckes MW, et al. Training of Clinicians for Public Health Events Relevant to Bioterrorism Preparedness. (Evidence Report/Technology Assessment No. 51. (Prepared by Johns Hopkins Evidence-based Practice Center under Contract No. 290-97-006). Rockville, MD: Agency for Healthcare Research and Quality, January 2002. AHRQ Pub. No. 02-E011.

31. Bravata DM, McDonald K, Owens DK, et al. Bioterrorism Preparedness and Response: Use of Information Technologies and Decision Support Systems. (Evidence Report/Technology Assessment No. 59 (Prepared by University of California San Francisco B Stanford Evidence-based Practice Center under Contract No.290-97-0013). Rockville, MD: Agency for Healthcare Research and Quality, June 2002. AHRQ Publication No. 02-E028.

32. Appel LJ, Robinson KA, Guallar E, et al. Utility of Blood Pressure Monitoring Outside of the Clinic Setting. Evidence Report/Technology Assessment No. 63 (Prepared by the Johns Hopkins Evidence-based Practice Center under Contract No 290-97-006). Rockville, MD: Agency for Healthcare Research and Quality, November 2002. AHRQ Publication No. 03-E004.

33. Balion C, Santaguida P, Hill S, et al. Testing for BNP and NT-proBNP in the Diagnosis and Prognosis of Heart Failure. Evidence Report/Technology Assessment No. 142. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-02-0020). Rockville, MD: Agency for Healthcare Research and Quality, September 2006. AHRQ Publication No. 06-E014.

34. Viswanathan M, King VJ, Bordley C, et al. Management of Bronchiolitis in Infants and Children. Evidence Report/Technology Assessment No. 69 (Prepared by RTI International-University of North Carolina at Chapel Hill Evidence-based Practice Center under Contract No. 290-97-0011). Rockville, MD: U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality, January 2003. AHRQ Publication No. 03-E014.

35. Ford JG, Howerton MW, Bolen S, et al. Knowledge and Access to Information on Recruitment of Underrepresented Populations to Cancer Clinical Trials. Evidence Report/Technology Assessment No. 122 (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 290-02-0018.) Rockville, MD: Agency for Healthcare Research and Quality, June 2005. AHRQ Publication No. 05-E019-2.

36. Ellis P, Robinson P, Ciliska D, et al. Diffusion and Dissemination of Evidence-based Cancer Control Interventions. Evidence Report/Technology Assessment Number 79. (Prepared by McMaster University under Contract No. 290-97-0017.) Rockville, MD: Agency for Healthcare Research and Quality, May 2003. AHRQ Publication No. 03-E033

37. Whelan TJ, O'Brien MA, Villasis-Keever M, et al. Impact of Cancer-Related Decision Aids. Evidence Report/Technology Assessment Number 46. (Prepared by McMaster University under Contract No. 290-97-0017.) Rockville, MD: Agency for Healthcare Research and Quality, July 2002. AHRQ Publication No. 02-E004.

38. Ammerman A, Lindquist C, Hersey J, et al. Efficacy of interventions to modify dietary behavior related to cancer risk. Evidence Report/Technology Assessment No. 25 (Contract No. 290-97-0011 to the Research Triangle Institute-University of North Carolina at Chapel Hill Evidence-based Practice Center), AHRQ Rockville, MD: Agency for Healthcare Research and Quality, February 2001. Publication No. 01-E029.

39. McAlister F, Ezekowitz J, Wiebe N, et al. Cardiac Resynchronization Therapy for Congestive Heart Failure. Evidence Report/Technology Assessment No. 106. (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023.) Rockville MD: Agency for Healthcare Research and Quality, November 2004. AHRQ Publication No. 05-E001-2.

40. Oliver D. Schein OD, David S. Friedman DS, Lee A. Fleisher LA, et al. Anesthesia Management During Cataract Surgery. Evidence Report/Technology Assessment No. 16. (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 290-097-0006.) Rockville, MD: Agency for Healthcare Research and Quality, December 2001. AHRQ Publication No. 01-E017.

41. Jampel H, Lubomski L, Friedman D. Treatment of Coexisting Cataract and Glaucoma. Evidence Report/Technology Assessment Number 38. (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. 290-97-0006.) Rockville, MD: Agency for Healthcare Research and Quality, June 2003. AHRQ Publication No. 03-E041.

42. Viswanathan M, Ammerman A, Eng E, et al. Community-Based Participatory Research: Assessing the Evidence. Evidence Report/Technology Assessment No. 99 (Prepared by RTI University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016). Rockville, MD: Agency for Healthcare Research and Quality, July 2004. AHRQ Publication 04-E022-2. .

43. Rostom A, Dubé C, Cranney A, et al. Celiac Disease. Evidence Report/Technology Assessment No. 104. (Prepared by the University of Ottawa Evidence-based Practice Center, under Contract No. 290-02-0021.) Rockville, MD: Agency for Healthcare Research and Quality, September 2004. AHRQ Publication No. 04-E029-2.

44. Grady D, Chaput L, Kristof M. Results of Systematic Review of Research on Diagnosis and Treatment of Coronary Heart Disease in Women. Evidence Report/Technology Assessment No. 80. (Prepared by the University of California, San Francisco-Stanford Evidence-based Practice Center under Contract No 290-97-0013.) Rockville, MD: Agency for Healthcare Research and Quality, May 2003. AHRQ Publication No. 03-0035.

45. Marinopoulos SS, Dorman T, Ratanawongsa N, et al. Effectiveness of Continuing Medical Education. Evidence Report/Technology Assessment No. 149 (Prepared by the Johns Hopkins Evidence-based Practice Center, under Contract No. 290-02-0018.) Rockville, MD: Agency for Healthcare Research and Quality, January 2007. AHRQ Publication No. 07-E006.

46. McCrory DC, Brown C, Gray RN, et al. Management of Acute Exacerbations of COPD. Evidence Report/Technology Assessment No. 19 (Contract 290-97-0014 to the Duke University Evidence-based Practice Center). Rockville, MD: Agency for Healthcare Research and Quality, March 2001. AHRQ Publication No. 01-E003.

47. Flamm CR, Aronson N, Bohn R, et al. Use of Epoetin for Anemia in Chronic Renal Failure. Evidence Report/Technology Assessment No. 29 (Prepared by the Blue Cross and Blue Shield Association Technology Evaluation Center under Contract No. 290-97-0015). Rockville, MD: Agency for Healthcare Research and Quality, August 2001. AHRQ Publication No. 01-E016.

48. Viswanathan M, Visco AG, Hartmann K, et al. 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.) Rockville, MD: Agency for Healthcare Research and Quality, March 2006. AHRQ Publication No. 06-E009.

49. Bonito AJ, Palton LL, Shugars DA, et al. Management of Dental Patients Who are HIV-positive. Evidence Report/Technology Assessment No. 37 (Contract 290-97-0011 to the Research Triangle Institute-University of North Carolina at Chapel Hill Evidence-based Practice Center). Rockville, MD: Agency for Healthcare Research and Quality, March 2002. AHRQ Publication No. 01-E042.

50. Bader JD, Bonito AJ, Shugars DA. Cardiovascular Effects of Epinephrine on Hypertensive Dental Patients. Evidence Report/Technology Assessment Number 48. (Prepared by Research Triangle Institute under Contract No. 290-97-0011.) Rockville, MD: Agency for Healthcare Research and Quality, July 2002. AHRQ Publication No. 02-E006.

51. Golden S, Boulware LE, Berkenblit G, et al. Use of Glycated Hemoglobin and Microalbuminuria in the Monitoring of Diabetes Mellitus. Evidence Report/Technology Assessment No. 84 (Prepared by Johns Hopkins Evidence-based Practice Center under Contract No. 290-97-0006). Rockville, MD: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, October 2003. AHRQ Publication No. 04-E001.

52. Ross SD, Levine C, Ganz N, et al. Systematic Review of the Current Literature Related to Disability and Chronic Fatigue Syndrome. Evidence Report/Technology Assessment No. 66 (Prepared by MetaWorks Inc. Evidence-based Practice Center under Contract No 290-97-0016). Rockville, MD: Agency for Healthcare Research and Quality, December 2002. AHRQ Publication No. 03-E007.

53. Segal JB, Eng J, Jenckes MW, et al. Diagnosis and Treatment of Deep Venous Thrombosis and Pulmonary Embolism. Evidence Report/Technology Assessment Number 68. (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. 290-97-0007.) Rockville, MD: Agency for Healthcare Research and Quality, March 2003. AHRQ Publication No. 03-E016.

54. Berkman ND, Bulik CM, Brownley KA, et al. Management of Eating Disorders. Evidence Report/Technology Assessment No. 135. (Prepared by the RTI International-University of North Carolina Evidence-Based Practice Center under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality, April 2006. AHRQ Publication No. 06-E010.

55. Lorenz K, Lynn J, Morton SC, et al. End-of-Life Care and Outcomes. Evidence Report/Technology Assessment No. 110. (Prepared by the Southern California Evidence-based Practice Center, under Contract No. 290-02-0003.) Rockville, MD: Agency for Healthcare Research and Quality, December 2004. AHRQ Publication No. 05-E004-2.

56. Aronson N, Flamm CR, Mark D, et al. Endoscopic Retrograde Cholangiopancreatography. Evidence Report/Technology Assessment Number 50. (Prepared by Blue Cross and Blue Shield Association under Contract No. 290-97-001-5.) Rockville, MD: Agency for Healthcare Research and Quality, June 2002. AHRQ Publication No. 02-E017

57. Ross SD, Estok R, Chopra S, et al. Management of Newly Diagnosed Patients with Epilepsy: A Systematic Review of the Literature. Evidence Report/Technology Assessment No. 39 (Contract 290-97-0016 to MetaWorks, Inc.) Rockville, MD: Agency for Healthcare Research and Quality, September 2001. AHRQ Publication No. 01-E038.

58. Chapell R, Reston J, Snyder D. Management of Treatment-Resistant Epilepsy. Evidence Report/Technology Assessment No. 77. (Prepared by the ECRI Evidence-based Practice Center under Contract No 290-97-0020.) Rockville, MD: Agency for Healthcare Research and Quality, May 2003. AHRQ Publication No. 03-0028.

59. Viswanathan M, Hartmann K, Palmieri R, et al. The Use of Episiotomy in Obstetrical Care: A Systematic Review. Evidence Report/ Technology Assessment No. 112. (Prepared by the RTI-UNC Evidence-based Practice Center, under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality, May 2005. AHRQ Publication No. 05-E009-2.

60. Perrin EC, Cole CH, Frank DA, et al. Criteria for Determining Disability in Infants and Children: Failure to Thrive. Evidence Report/Technology Assessment No. 72 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality, March 2003. AHRQ Publication No. 03-E026.

61. McDonagh MS, Carson S, Ash JS, et al. Hyperbaric Oxygen Therapy for Brain Injury, Cerebral Palsy, and Stroke. Evidence Report/Technology Assessment No. 85 (Prepared by the Oregon Health & Science University Evidence-based Practice Center under Contract No 290-97-0018). Rockville, MD: Agency for Healthcare Research and Quality, September 2003. AHRQ Publication No. 04-E003.

62. Berkman ND, DeWalt DA, Pignone MP, et al. Literacy and health outcomes. Evidence Report/Technology Assessment No. 87. (Prepared by RTI International-University of North Carolina under Contract No. 290-02-0016). Rockville, MD: Agency for Healthcare Research and Quality, 2004. AHRQ Publication No. 04-E007-2.

63. Oremus M, Hanson M, Whitlock R, et al. The Uses of Heparin To Treat Burn Injury. Evidence Report/Technology Assessment No. 148. (Prepared by the McMaster University Evidence-based Practice Center, under Contract No. 290-02-0020). Rockville, MD: Agency for Healthcare Research and Quality, December 2006. AHRQ Publication No. 07-E004.

64. Gebo KA, Jenckes MJ, Chander G, et al. Management of Chronic Hepatitis C. Evidence Report/Technology Assessment No. 60 (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No 290-97-0006). Rockville, MD: Agency for Healthcare Research and Quality, July 2002. AHRQ Publication No. 02-E030.

65. Santaguida PL, Balion C, Hunt D, et al. Diagnosis, Prognosis, and Treatment of Impaired Glucose Tolerance and Impaired Fasting Glucose. Evidence Report/Technology Assessment No. 128. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-02-0020). Rockville, MD: Agency for Healthcare Research and Quality, September 2005. AHRQ Pub. No 05-E026-2.

66. Buscemi N, Vandermeer B, Friesen C, et al. Manifestations and Management of Chronic Insomnia in Adults. Evidence Report/Technology Assessment No. 125. (Prepared by the University of Alberta Evidence-based Practice Center, under Contract No. C400000021.) Rockville, MD: Agency for Healthcare Research and Quality, June 2005. AHRQ Publication No. 05-E021-2.

67. Cole C, Binney G, Casey P, et al. Criteria for Determining Disability in Infants and Children: Low Birth Weight. Evidence Report/Technology Assessment No. 70 (Prepared by Tufts New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality, December 2002. AHRQ Publication No. 03-E010.

68. Meenan RT, Saha S, Chou R, et al. Effectiveness and Cost-Effectiveness of Echocardiography and Carotid Imaging in the Management of Stroke. Evidence Report/Technology Assessment Number 49. (Prepared by Oregon Health & Science University Evidence-based Practice Center under Contract No. 290-97-0018.) Rockville, MD: Agency for Healthcare Research and Quality, July 2002. AHRQ Publication No. 02-EO22.

69. Cook D, Meade M, Guyatt G, et al. Criteria for Weaning From Mechanical Ventilation. Evidence Report/Technology Assessment No. 23 (Prepared by McMaster University under Contract No. 290-97-0017). Rockville MD: Agency for Healthcare Research and Quality, November 2000. AHRQ Publication No. 01-E010.

70. Buscemi N, Vandermeer B, Pandya R, et al. Melatonin for Treatment of Sleep Disorders. Evidence Report/Technology Assessment No. 108. (Prepared by the University of Alberta Evidence-based Practice Center, under Contract No. 290-02-0023.) Rockville, MD: Agency for Healthcare Research and Quality, November 2004. AHRQ Publication No. 05-E002-2.

71. Nelson HD, Haney E, Humphrey L, et al. Management of Menopause-Related Symptoms. Evidence Report/Technology Assessment No. 120. (Prepared by the Oregon Evidence-based Practice Center, under Contract No. 290-02-0024.) Rockville, MD: Agency for Healthcare Research and Quality, March 2005. AHRQ Publication No. 05-E016-2.

72. Beach MC, Cooper LA, Robinson KA, et al. Strategies for Improving Minority Healthcare Quality. Evidence Report/Technology Assessment No. 90. (Prepared by the Johns Hopkins University Evidence-based Practice Center, Baltimore, MD.) Rockville, MD: Agency for Healthcare Research and Quality, January 2004. AHRQ Publication No. 04-E008-02.

73. McCrory DC, Pompeii LA, Skeen MB, et al. Criteria to Determine Disability Related to Multiple Sclerosis. Evidence Report/Technology Assessment No. 100. (Prepared by the Duke Evidence-based Practice Center, Durham, NC, under Contract No. 290-02-0025.) Rockville, MD: Agency for Healthcare Research and Quality, May 2004. AHRQ Publication No. 04-E019-2.

74. Ip S, Glicken S, Kulig J, et al. Management of Neonatal Hyperbilirubinemia. Evidence Report/Technology Assessment No. 65 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019). Rockville, MD: U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality, January 2003. AHRQ Publication No. 03-E011.

75. Shekelle PG, Morton SC, Maglione MA, et al. Pharmacological and Surgical Treatment of Obesity. Evidence Report/Technology Assessment No. 103. (Prepared by the Southern California-RAND Evidence-Based Practice Center, Santa Monica, CA, under contract Number 290-02-0003.) Rockville, MD: Agency for Healthcare Research and Quality, July 2004. AHRQ Publication No. 04-E028-2.

76. Hodge W, Barnes D, Schachter H, et al. Effects of Omega-3 Fatty Acids on Eye Health. Evidence Report/Technology Assessment No. 117 (Prepared by University of Ottawa Evidence-based Practice Center under Contract No. 290-02-0021.) Rockville, MD: Agency for Healthcare Research and Quality, July 2005. AHRQ Publication No. 05-E008-2.

77. MacLean CH, Issa AM, Newberry SJ, et al. Effects of Omega-3 Fatty Acids on Cognitive Function with Aging, Dementia, and Neurological Diseases. Evidence Report/Technology Assessment No. 114 (Prepared by the Southern California Evidence-based Practice Center, under Contract No. 290-02-0003.) Rockville, MD: Agency for Healthcare Research and Quality, February 2005. AHRQ Publication No. 05-E011-2.

78. MacLean CH, Mojica WA, Morton SC, et al. Effects of Omega-3 Fatty Acids on Lipids and Glycemic Control in Type II Diabetes and the Metabolic Syndrome and on Inflammatory Bowel Disease, Rheumatoid Arthritis, Renal Disease, Systemic Lupus Erythematosus, and Osteoporosis. Evidence Report/Technology Assessment. No. 89 (Prepared by Southern California/RAND Evidence-based Practice Center, under Contract No. 290-02-0003). Rockville, MD: Agency for Healthcare Research and Quality, March 2004. AHRQ Publication No. 04-E012-2.

79. Lewin GA, Schachter HM, Yuen D, et al. Effects of Omega-3 Fatty Acids on Child and Maternal Health. Evidence Report/Technology Assessment No. 118. (Prepared by the University of Ottawa Evidence-based Practice Center, under Contract No. 290-02-0021.) Rockville, MD: Agency for Healthcare Research and Quality, August 2005. AHRQ Publication No. 05-E025-2.

80. Schachter HM, Kourad K, Merali Z, et al. Effects of Omega-3 Fatty Acids on Mental Health. Evidence Report/Technology Assessment No. 116. (Prepared by the University of Ottawa Evidence-based Practice Center, Under Contract No. 290-02-0021.) Rockville, MD: Agency for Healthcare Research and Quality, July 2005. AHRQ Publication No. 05-E022-2.

81. Bonis PA, Chung M, Tatsioni A, et al. Effects of Omega-3 Fatty Acids on Organ Transplantation. Evidence Report/Technology Assessment No. 115 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality, February 2005. AHRQ Publication No. 05-E012-2.

82. Schachter H, Reisman J, Tran K, et al. Health Effects of Omega-3 Fatty Acids on Asthma. Evidence Report/Technology Assessment No. 91 (Prepared by University of Ottawa Evidence-based Practice Center under Contract No. 290-02-0021). Rockville, MD: Agency for Healthcare Research and Quality, March 2004. AHRQ Publication No. 04-E013-2.

83. Wang C, Chung M, Lichtenstein A, et al. Effects of Omega-3 Fatty Acids on Cardiovascular Disease. Evidence Report/Technology Assessment No. 94 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center, under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality, March 2004. AHRQ Publication No. 04-E009-2.

84. Marcy M, Takata G, Chan LS, et al. Management of Acute Otitis Media. Evidence Report/Technology Assessment No. 15 (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-97-0001). Rockville, MD: Agency for Healthcare Research and Quality, May 2001. AHRQ Publication No. 01-E010.

85. Shekelle P, Takata G, Chan LS, et al. Diagnosis, Natural History, and Late Effects of Otitis Media with Effusion. Evidence Report/Technology Assessment No. 55 (Prepared by Southern California Evidence-based Practice Center under Contract No 290-97-0001, Task Order No. 4). Rockville, MD: Agency for Healthcare Research and Quality, May 2003. AHRQ Publication No. 03-E023.

86. Hickam DH, Severance S, Feldstein A, et al. The Effect of Health Care Working Conditions on Patient Safety. Evidence Report/Technology Assessment Number 74. (Prepared by Oregon Health & Science University under Contract No. 290-97-0018.) Rockville, MD: Agency for Healthcare Research and Quality, April 2003. AHRQ Publication No. 03-E024.

87. Gaynes BN, Gavin N, Meltzer-Brody S, et al. Perinatal Depression: Prevalence, Screening Accuracy, and Screening Outcomes. Evidence Report/Technology Assessment No. 119. (Prepared by the RTI-University of North Carolina Evidence-based Practice Center, under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality, February 2005. AHRQ Publication No. 05-E006-2.

88. Holtzman J, Schmitz K, Babes G, et al. Effectiveness of Behavioral Interventions to Modify Physical Activity Behaviors in General Populations and Cancer Patients and Survivors. Evidence Report/Technology Assessment No. 102 (Prepared by the Minnesota Evidence-based Practice Center, under Contract No. 290-02-0009.) Rockville, MD: Agency for Healthcare Research and Quality, June 2004. AHRQ Publication No. 04-E027-2.

89. Myers ER, Blumrick R, Christian AL, et al. Management of prolonged pregnancy. Evidence Report/Technology Assessment No. 53 (Prepared by Duke Evidence-based Practice Center, Durham, NC, under Contract No. 290-97-0014). Rockville, MD: Agency for Healthcare Research and Quality, May 2002. AHRQ Publication No. 02-E018.

90. Bush DE, Ziegelstein RC, Patel UV, et al. Post-Myocardial Infarction Depression. Evidence Report/Technology Assessment No. 123. (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 290-02-0018.) Rockville, MD: Agency for Healthcare Research and Quality, May 2005. AHRQ Publication No. 05-E018-2.

91. Long A, McFadden C, DeVine D, et al. Management of Allergic and Nonallergic Rhinitis. Evidence Report/Technology Assessment No. 54 (Prepared by New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019). Rockville, MD: Agency for Healthcare Research and Quality, May 2002. AHRQ Pub. No. 02-E024. .

92. Balk E, Chung M, Chew P, et al. Effects of Soy on Health Outcomes. Evidence Report/Technology Assessment No. 126. (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022.) Rockville, MD: Agency for Healthcare Research and Quality, August 2005. AHRQ Publication No. 05-E024-2.

93. McCrory DC, Samsa GP, Hamilton BB, et al. Treatment of Pulmonary Disease Following Cervical Spinal Cord Injury. . Evidence Report/Technology Assessment Number 27. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-97-0014.) Rockville, MD: Agency for Healthcare Research and Quality, September 2001. AHRQ Publication No. 01-E014.

94. DeForge D, Blackmer J, Moher D, et al. Sexuality and Reproductive Health Following Spinal Cord Injury. Evidence Report/Technology Assessment No. 109 (Prepared by the University of Ottawa Evidence-based Practice Center under Contract No. 290-02-0021). Rockville, MD: Agency for Healthcare Research and Quality, November 2004. AHRQ Publication No. 05-E003-2.

95. Ranney L, Melvin C, Lux LJ, et al. Tobacco use: prevention, cessation, and control. Evidence Report/Technology Assessment No. 140. (Prepared by the RTI-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality, 2006. AHRQ Pub. No. 06-E015.

96. Kiddoo D, Klassen TP, Lang ME, et al. The Effectiveness of Different Methods of Toilet Training for Bowel and Bladder Control. Evidence Report/Technology Assessment No. 147. (Prepared by the University of Alberta Evidence-based Practice Center, under contract number 290-02-0023). Rockville, MD: Agency for Healthcare Research and Quality, December 2006. AHRQ Publication No. 07-E003.

97. Kane RL, Saleh KJ, Wilt TJ, et al. Total Knee Replacement. Evidence Report/Technology Assessment No. 86 (Prepared by the Minnesota Evidence-based Practice Center, Minneapolis, MN). Rockville, MD: Agency for Healthcare Research and Quality, December 2003. AHRQ Publication No. 04-E006-2.

98. Viswanathan M, Hartmann K, McKoy N, et al. Management of Uterine Fibroids: An Update of the Evidence. Evidence Report/Technology Assessment No. 154 (Prepared by RTI International-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016.) Rockville, MD: Agency for Healthcare Research and Quality, July 2007. AHRQ Publication No. 07-E011.

99. Guise J-M, McDonagh MS, Hashima J, et al. Vaginal Birth After Cesarean (VBAC). Evidence Report/Technology Assessment No. 71 (Prepared by the Oregon Health & Science University Evidence-based Practice Center under Contract No 290-97-0018). Rockville, MD: Agency for Healthcare Research and Quality, March 2003. AHRQ Publication No. 03-E018.

100. Velmahos GC, Kern J, Chan L, et al. Prevention of Venous Thromboembolism After Injury. Evidence Report/ Technology Assessment No. 22. (Prepared by Southern California Evidence-based Practice Center/RAND under Contract No. 290-97-0001.) Rockville, MD: Agency for Healthcare Research and Quality, November 2000. AHRQ Publication No. 01-E004.

101. Chan LS, Kipke MD, Schneir A, et al. Preventing Violence and Related Health-Risking Social Behaviors In Adolescents. Evidence Report/Technology Assessment No. 107 (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-02-2003.) Rockville, MD: Agency for Healthcare Research and Quality, October 2004. AHRQ Publication No. 04-E032-2.

102. Beach J, Rowe BH, Blitz S, et al. Diagnosis and Management of Work-Related Asthma. Evidence Report/Technology Assessment No. 129. (Prepared by the University of Alberta Evidence-based Practice Center, under Contract No. 290-02-0023.) Rockville, MD: Agency for Healthcare Research and Quality, November 2005. AHRQ Publication No. 06-E003-2.

103. Coulter ID, Hardy ML, Favreau JT, et al. Mind-Body Interventions for Gastrointestinal Conditions. Evidence Report/Technology Assessment No. 40 (Prepared by Southern California Evidence-based Practice Center/RAND under Contract No. 290-97-0001). Rockville, MD: Agency for Healthcare Research and Quality, July 2001. AHRQ Publication No. 01-E030.

104. Hersh WR, Hickam DH, Severance SM, et al. Telemedicine for the Medicare Population: Update. Evidence Report/Technology Assessment No. 131 (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024.) Rockville, MD: Agency for Healthcare Research and Quality, February 2006. AHRQ Publication No. 06-E007.

105. Levine C, Armstrong K, Chopra S, et al. Diagnosis and management of specific breast abnormalities. Evidence Report/Technology Assessment No. 33 (Prepared by MetaWorks, Inc., Boston, MA under Contract No. 290-97-0016). Rockville, MD: Agency for Healthcare Research and Quality, September 2001. AHRQ Publication No. 01-E046.

106. Matchar DB, Thakur ME, Grossman I, et al. Testing for Cytochrome P450 Polymorphisms in Adults With Non-Psychotic Depression Treated With Selective Serotonin Reuptake Inhibitors (SSRIs). Evidence Report/Technology Assessment No. 146. (Prepared by the Duke Evidence-based Practice Centerunder Contract No. 290-02-0025.) Rockville, MD: Agency forHealthcare Research and Quality, November 2006. AHRQ Publication No. 07-E002.

107. Seidenfeld J, Samson DJ, Bonnell CJ, et al. Management of Small Cell Lung Cancer. Evidence Report/Technology Assessment No. 143. (Prepared by Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-based Practice Center under Contract No. 290-02-0026.) Rockville, MD: Agency for Healthcare Research and Quality, July 2006. AHRQ Publication No. 06-E016.

108. Myers ER, Havrilesky LJ, Kulasingam SL, et al. Genomic Tests for Ovarian Cancer Detection and Management. Evidence Report/Technology Assessment No. 145. (Prepared by the Duke University Evidence-based Practice Center under Contract No. 290-02-0025.) Rockville, MD: Agency for Healthcare Research and Quality, October 2006. AHRQ Publication No. 07-E001.

109. ECRI Institute Evidence-based Practice Center. Quality item checklist for SINGLE-GROUP studies. 2008.

110. University Of Alberta Evidence-Based Practice Centre. Quality Assessment Tool for Observational Analytical Studies.

111. National Institute for Health and Clinical Excellence. The guidelines manual. London: National Institute for Health and Clinical Excellence; 2007. Available at: www.nice.org.uk.

112. Khan KS, ter Riet G, Glanville J, et al., eds. Undertaking Systematic Reviews of Research on Effectiveness,. 2nd ed. York, U.K.: NHS Centre for Reviews and Dissemination, 2001.

113. Johnston P, Wilkinson K. Enhancing validity of critical tasks selected for college and university program portfolios. National Forum of Teacher Education Journal. 2009;19(3):1-6.

114. Crisp AH, Callender JS, Halek C, et al. Long-term motrality in anorexia nervosa: A 20-year follow-up of the St. George's and Aberdeen cohorts. Br J Psychiatry. 1992;161:104-7.

115. Daniel M, Green LW, Marion SA, et al. Effectiveness of community-directed diabetes prevention and control in a rural Aboriginal population in British Columbia, Canada. Soc Sci Med. 1999;48:815-832.

116. Hedderson MM, Weiss NS, Sacks DA, et al. Pregnancy weight gain and risk of neonatal complications. Obstet Gynecol. 2006;108:1153-61.

117. Di Lieto A, Iannotti F, De Falco M, et al. Immunohistochemical detection of insulin-like growth factor type I receptor and uterine volume changes in gonadotropin-releasing hormone analog-treated uterine leiomyomas. Am J Obstet Gynecol. 2003;188:702-6.

118. Coleman FH. Safety and efficacy of combined ritodrine and magnesium sulfate for preterm labor: a method for reduction of complications. American Journal of Perinatology. 1990;7(4):366-9.

119. Schindl M, Birner P, Reingrabner M, et al. Elective cesarean section vs. spontaneous delivery: a comparative study of birth experience. Acta Obstet Gynecol Scand. 2003;82:834-40.

120. Fouad MN, Kiefe CI, Bartolucci AA, et al. A hypertension control program tailored to unsilled and minority workers. Ethnicity Dis. 1997;7:191-9.

121. Baker DW, Gazmararian JA, Williams MV, et al. Functional health literacy and the risk of hospital admission among Medicare managed care enrollees. Am J Public Health. 2002;92(8):1278-83.

122. Kinney TR, Helms RW, O'Branski EE, et al. Safety of hydroxyurea in children with sickle cell anemia: results of the HUG-KIDS Study, a phase I/II trial. Blood. 1999;94:1550-4.

123. Van Ham MAPC, van Dongen PWJ, Mulder J. Maternal consequences of caesarean section. A retrospective study of intra-operative and postoperative maternal complications of caesarean section during a 10-year period. European Journal of Obstetrics and Gynecology. 1997;74:1-6.

124. Blood E, Spratt KF. Disagreement on agreement: Two alternative agreement coefficients. Paper 186-2007. SAS Global Forum. 2007:1-12.

125. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9.

Appendix A. AC1 Statistic

AC1 was originally introduced by Gwet in 2001 (Gwet, 2001). The interpretation of AC1 is similar to generalized kappa (Fleiss, 1971), which is used to assess interrater reliability of when there are multiple raters. Gwet (2002) demonstrated that AC1 can overcome the limitations that kappa is sensitive to trait prevalence and rater's classification probabilities (i.e., marginal probabilities), whereas AC1 provides more robust measure of interrater reliability. The section below shows the formula used to compute AC1. The first formula also shows that AC1 differs from generalized kappa in the way that how the chance correction was computed (i.e. how the $p_{e\gamma}$ is computed). In addition, the computation is unweighted, thus the ordering of the response category is not taken into account. Our computation of AC1 is conducted using the macro code provided by Blood et al (2007) (http://mcrc.hitchcock.org/SASMacros/Agreement/AC1AC2.TXT).

$$AC1 = \frac{p_a - p_{e\gamma}}{1 - p_{e\gamma}},$$

where p_a is the overall agreement probability including by chance or not by chance, and $P_{e\gamma}$ is the chance-agreement probability. Their computation formulas are as follows:

$$p_a = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\frac{r_{iq}(r_{iq}-1)}{r(r-1)}\right\}$$

$$p_{e\gamma} = \frac{1}{Q-1}\sum_{q=1}^{Q}\pi_q(1-\pi_q)$$

$$\pi_q = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{iq}}{r},$$

where i is the number of studies rated, q is the number of categories in the rating scale,
r_{iq} is the number of raters who classified the i^{th} studies into the q^{th} category,
r is the total number of raters, and
π_q is the probability that a rater classifies an study into categories q and computed as follows

References

Blood E, Spratt KF. Disagreement on agreement: Two alternative agreement coefficients. Paper 186-2007. SAS Global Forum. 2007:1-12.

Fleiss JL. Measuring nominal scale agreement among many raters, Psychological Bulletin 1971;76:378–2.

Gwet K. Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters. Gaithersburg, MD: STATAXIS Publishing Company; 2001.

Gwet K. Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity. Statistical Methods for Inter-Rater Reliability Assessment 2002;2:1–9.

.

Appendix B. Item Bank for Assessment of Risk of Bias and Precision 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). Some questions may be applicable to quasi-experimental designs. 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.

Sample Definition and Selection

Retrospective/Prospective

1. **Is the study design prospective, retrospective, or mixed?** *[Abstractor: Prospective design requires that the outcome has not occurred at the time the study is initiated and information is collected over time to assess relationships with the outcome (and includes nested case-control studies). Mixed design includes case-control or cohort studies in which one group is studied prospectively and the other retrospectively. A retrospective design analyzes data from past records. The question is not applicable to cross-sectional studies.]*

Prospective	☐	**Explanation for rating:**
Mixed	☐	
Retrospective	☐	
Cannot determine/not applicable	☐	

Inclusion/Exclusion Criteria

2. **Are critical inclusion/exclusion criteria clearly stated (does not require the reader to infer)?** *[Principal Investigator (PI): Provide direction to abstractors by listing individual criteria of a priori significance and minimal requirements for criteria to be considered "clearly stated." Include this question to identify specific inclusion/exclusion criteria that should be consistently recorded across studies] [Abstractor: Use "Partially" if only some criteria are stated or if some criteria are not clearly stated (corresponding to directions provided by the PI). Note that studies may describe inclusion criteria alone (i.e., include x), exclusion criteria (i.e., do not include x), or a combination of inclusion and exclusion criteria.]*

PI:

Yes ..... ☐

Partially: some, but not all, criteria stated or some criteria not clearly stated ..... ☐

No ..... ☐

Explanation for rating:

3. **Are the inclusion/exclusion criteria measured using valid and reliable measures?** *[PI: Separately specify each criterion that abstractors should consider based on its relevance to study bias. It is unlikely that all criteria will need to be evaluated in relation to this question. Provide direction to abstractors on valid and reliable measurement of each criterion that is to be considered. For example, prior exposure or disease status is a frequent inclusion/exclusion criterion, particularly in inception cohorts. Subjective measures based on self-report tend to have lower reliability and validity than objective measures such as clinical reports and lab findings. Replicate question to evaluate each individual inclusion/exclusion criterion.]*

PI:

Yes ..... ☐

No..... ☐

Cannot determine; measurement approach not reported ..... ☐

Explanation for rating:

4. ***Did the study apply inclusion/exclusion criteria uniformly to all comparison groups/arms of the study?*** *[PI: Drop question if not relevant to entire body of evidence (e.g., all case-series, single-arm studies).]*

PI:

Yes ..... ☐

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

No ..... ☐

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

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

Explanation for rating:

5. **Was the strategy for recruiting participants into the study the same across study groups/arms of the study?** *[PIs: This question is likely to be more relevant for prospective or mixed designs than retrospective designs. Drop question if not relevant to entire body of evidence (e.g., all studies generally have only one arm).]*

PI:

Yes ..... ☐

No ..... ☐

Cannot determine ..... ☐

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

Explanation for rating:

6. ***Was the sample size sufficiently large to detect a clinically significant difference of 5% or more between groups in at least one primary outcome measure?*** *[PI: Specify a different percent, if clinically relevant for each outcome of interest. 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.]*

PI:

Yes ..... ☐

No ..... ☐

Explanation for rating:

Interventions/Exposure

Clear Specification

7. ***What is the level of detail in describing the intervention or exposure?*** *[PI: Specify which details need to be stated (e.g., intensity, duration, frequency, route, setting, and timing of intervention/exposure). For case-control studies, consider whether the condition, timing, frequency, and setting of symptoms are provided in the case definition. PI needs to establish criteria for high, medium, or low response.]*

PI:

High: very clear, all PI-required details provided ☐

Medium: somewhat clear, majority of PI-required details provided ..... ☐

Low: unclear, many PI-required details missing..... ☐

Explanation for rating:

Outcomes

Clear Specifications

8. **Are the important outcomes prespecified by the researchers? Do not consider harms in answering this question unless they should have been pre-specified.** *[PI: This question can be asked for all outcomes together or replicated for each event. Each adverse event of interest should be specified for abstractors. Relevant source information includes all study data, including what*

may have been established in relation to an initial randomized controlled trial. Drop question if not relevant (e.g., primary outcome for case-control studies).]

PI:

Yes ..... ☐

Partially ..... ☐

No ..... ☐

Not applicable ..... ☐

Explanation for rating:

Creation of Treatment Groups

Allocation

9. **Is the selection of the comparison group appropriate, 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 ..... ☐

No ..... ☐

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 arm) ... ☐

Explanation for rating:

Any Attempt To Balance

10. **Any attempt to balance the allocation between the groups (e.g., through stratification, matching, propensity scores).** *[PI: This is most likely to be used in case-control study designs. 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 arm) ☐

Explanation for rating:

Contamination

11. **Did researchers isolate the impact from a concurrent intervention or an unintended exposure that might bias results, e.g., through multivariate analysis, stratification, or subgroup analysis?** *[PI: specify interventions or exposures for abstractors.]*

PI:

Yes ..... ☐

Partially..... ☐

No or do not know: concurrent intervention or unintended exposure is not described)..... ☐

Not applicable: no concurrent interventions or unintended exposures likely..... ☐

Explanation for rating:

12. **Did execution of the study vary from the intervention protocol proposed by the investigators and therefore compromise the conclusions of the study?** *[PI: Consider intensity, duration, frequency, route, setting, and timing of intervention/exposures. Drop if not relevant for body of literature.]*

PI:

Yes ..... ☐

Partially ..... ☐

No ..... ☐

Cannot determine ..... ☐

Not applicable: not an intervention study ..... ☐

Explanation for rating:

Blinding

Blind Outcomes Assessment

13. **Were the outcome assessors 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 ..... ☐

No ..... ☐

Not applicable: assessor cannot be blinded ..... ☐

Explanation for rating:

Soundness of Information

Source of Information Re Interventions/Exposure

14. **Are interventions/exposures assessed using valid and reliable measures, implemented consistently across all study participants?** *[PI: Important measures may 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. Replicate question when needed.]*

PI:

Yes ..... ☐

No ..... ☐

Cannot determine or measurement approach not reported ..... ☐

Explanation for rating:

Source of Information for Outcomes

15. **Are outcomes assessed using valid and reliable measures, implemented consistently across all study participants?** *[PI: Primary outcomes should be identified for abstractors and if there is more than one, they may be listed separately. Also, identify any relevant secondary outcomes and harms. Subjective measures based on self-report tend to have lower reliability and validity than objective measures such as clinical reports and lab findings. Note for case-control studies: consider whether the ascertainment of cases was independent of exposure.]*

PI:

Yes ..... ☐

No ..... ☐

Cannot determine or measurement approach not reported ..... ☐

Explanation for rating:

Follow-Up

Equality of Length of Follow-Up for Participants

16. **Is the length of follow-up the same for all groups?** *[For case-control studies, are cases and controls matched on length of followup? Abstractor: When follow-up was the same for all study participants, the answer is yes. If different lengths of follow-up were adjusted by statistical techniques, (e.g., survival analysis), the answer is yes. Studies in which differences in follow-up were ignored should be answered no.]*

Yes ..... ☐

No or cannot determine..... ☐

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

Explanation for rating:

Length of Followup Adequate

17. **Is the length of time following the intervention/exposure sufficient to support the evaluation of primary outcomes and harms?** *[PI: Primary outcomes (including harms) should be identified for abstractors. Important measures may be listed separately. Abstractors should be provided with specific criteria for sufficient length of follow-up based on prior research or theory. Drop if entire body of evidence is cross-sectional or if minimal length of follow-up period is specified through inclusion criteria.]*

PI:

Yes	☐	**Explanation for rating:**
Partially: some primary outcomes are followed for a sufficient length of time	☐	
No	☐	
Cannot determine	☐	
Not applicable: cross-sectional	☐	

Completeness of Follow-Up

18. **Did attrition from any group exceed [x] percent?** *[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 crossover. 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). Drop if entire body of evidence is cross-sectional]*

PI:

Yes	☐	**Explanation for rating:**
No	☐	
Cannot determine: includes retrospective designs not stating number eligible at baseline	☐	
Not applicable: cross-sectional	☐	

19. **Did attrition differ between groups by more than 20 percent?** *[PI: If appropriate, modify difference criterion to meet relevant standards for the topic. Attrition rates may vary by outcome and time of measurement. Drop if entire body of evidence is cross-sectional or case series.]*

PI:

Yes ..... ☐	**Explanation for rating:**
No ..... ☐	
Cannot determine: includes retrospective designs not stating number eligible at baseline.... ☐	
Not applicable: cross-sectional or only one group followed—case series, one-arm study..... ☐	

Analysis Comparability

Assessment of Baseline Comparability

20. **Does the analysis control for baseline differences between groups?** *[PI: Drop if entire body of evidence is case series or case control. Define adequate control. List critical baseline differences that need to be controlled.]*

Yes ..... ☐	**Explanation for rating:**
No ..... ☐	
Insufficient reporting to be able to determine..... ☐	
Not applicable: only one group, no comparison group (case series), or case-control study, no difference in measured baseline characteristics... ☐	

Identification of Prognostic Factors (Effect Modifiers and Confounders)

21. **Are confounding and/or effect modifying variables assessed using valid and reliable measures across all study participants?** *[PI: 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 ..... ☐

No ..... ☐

Cannot determine or source for measures not reported ..... ☐

Not applicable: no confounders or effect modifiers included in the study ..... ☐

Explanation for rating:

Case-Mix Adjustment

22. **Were the important confounding and effect modifying variables taken into account in the design and/or analysis (e.g., through matching, stratification, interaction terms, multivariate analysis, or other statistical adjustment)?** *[PI: Provide instruction to abstractors on adequate adjustment for confounding and testing for effect modification.]*

PI:

Yes ..... ☐

Partially: some variables taken into account or adjustment achieved to some extent ..... ☐

No: not accounted for or not identified ..... ☐

Cannot determine ..... ☐

Explanation for rating:

Analysis Outcome

Intention-to-Treat Analysis

23. **In cases of high loss to follow-up (or differential loss to follow-up), is the impact assessed (e.g., through sensitivity analysis or other adjustment method)?**

Yes ..... ☐

No ..... ☐

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:

Appropriate Analytic Methods

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

PI:

Yes ..... ☐

No ..... ☐

Cannot determine ..... ☐

Explanation for rating:

25. **Are the statistical methods used to assess the primary benefit outcomes appropriate to the data?** *[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 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]*

Yes ..... ☐

Partially ..... ☐

No ..... ☐

Cannot determine ..... ☐

Explanation for rating:

26. **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, including timing of measurement, that one would expect be reported in the study. Drop if not relevant to body of literature.]*

PI:

Yes	☐	**Explanation for rating:**
Partially	☐	
No	☐	
Assessment of harms not applicable to this study	☐	

27. ***Are the statistical methods used to assess the main harm or adverse event outcomes appropriate to the data?*** *[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 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.]*

Yes	☐	**Explanation for rating:**
Partially	☐	
No	☐	
Not applicable: harms not reported	☐	

Interpretation

Appropriately Based on Results

28. **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.]*

Yes	☐	**Explanation for rating:**
Partially	☐	
No	☐	

Presentation and Reporting

Completeness, Clarity, and Structure

29. **Is the source of funding identified?** *[PI: The relevance of this question will depend upon the topic. This question may be modified to identify particular sources of funding (e.g., industry, government, university, or foundation funding).]*

PI:

Yes ..... ☐

No ..... ☐

Explanation for rating: