# Inclusion of Nonrandomized Studies of Interventions in Systematic Reviews of Intervention Effectiveness: An Update

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ
Agency for Healthcare
Research and Quality

This work was funded by the Agency for Healthcare Research and Quality through the following contracts: Scientific Resource Center (290-2017-00003-C), Brown University Evidence-based Practice Center (EPC) (75Q80120D00001), ECRI Institute–Penn Medicine EPC (75Q80120D00002), Pacific Northwest EPC (75Q80120D00006), Johns Hopkins University EPC (75Q80120D00003), Mayo Clinic EPC (75Q80120D00005), Minnesota EPC (75Q80120D00008), RTI–University of North Carolina EPC (75Q80120D00007), and University of Southern California EPC (75Q80120D00009). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

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

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

AHRQ or U.S. Department of Health and Human Services endorsement of any derivative products that may be developed from this report, such as clinical practice guidelines, other quality enhancement tools, or reimbursement or coverage policies, may not be stated or implied. A representative from AHRQ served as a Contracting Officer's Representative and reviewed the contract deliverables for adherence to contract requirements and quality.

AHRQ appreciates appropriate acknowledgment and citation of its work. Suggested language for acknowledgment: This work was based on a Methods Guide for Comparative Effectiveness Reviews paper, Inclusion of Nonrandomized Studies of Interventions in Systematic Reviews of Intervention Effectiveness: An Update, by the Evidence-based Practice Center Program at the Agency for Healthcare Research and Quality (AHRQ).

**Suggested citation:** Saldanha IJ, Skelly AC, Vander Ley K, Wang Z, Berliner E, Bass EB, Devine B, Hammarlund N, Adam GP, Duan-Porter D, Bañez LL, Jain A, Norris SL, Wilt TJ,

Leas B, Siddique SM, Fiordalisi CV, Patino-Sutton C, Viswanathan M. Inclusion of Nonrandomized Studies of Interventions in Systematic Reviews of Intervention Effectiveness: An Update. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Scientific Resource Center under Contract No. 290-2017-00003-C.) AHRQ Publication No. 22-EHC033. Rockville, MD: Agency for Healthcare Research and Quality; September 2022. DOI: https://doi.org/10.23970/AHRQEPCMETHODSGUIDENRSI. Posted final reports are located on the Effective Health Care Program search page.

**Prepared for:**
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

**Contract No. 290-2017-00003-C**

**Prepared by:**
Scientific Resource Center
Portland, OR

**Investigators:**
Ian J. Saldanha, M.B.B.S, M.P.H., Ph.D.
Andrea C. Skelly, Ph.D., M.P.H.
Kelly Vander Ley, Ph.D.
Zhen Wang, Ph.D.
Elise Berliner, Ph.D.
Eric B. Bass, M.D., M.P.H.
Beth Devine, Pharm.D., Ph.D., M.B.A.
Noah Hammarlund, Ph.D.
Gaelen P. Adam, M.L.I.S., M.P.H.
Denise Duan-Porter, M.D., Ph.D.
Lionel L. Bañez, M.D.
Anjali Jain, M.D.
Susan L. Norris, M.D., M.P.H.
Timothy J. Wilt, M.D., M.P.H.
Brian Leas, M.S.
Shazia M. Siddique, M.D., M.S.H.P.
Celia V. Fiordalisi, M.S.
Cecilia Patino-Sutton, M.D., Ph.D.
Meera Viswanathan, Ph.D.

# 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 healthcare in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new healthcare 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.

Strong methodological approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a Methods Guide for Comparative Effectiveness Reviews. This Guide presents issues key to the development of systematic reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The Methods Guide for Comparative Effectiveness Reviews is a living document and will be updated as further empiric evidence develops and our understanding of better methods improves.

If you have comments on this Methods Guide paper, they may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov.

Robert Otto Valdez, Ph.D., M.H.S.A.
Director
Agency for Healthcare Research and Quality

Arlene S. Bierman, M.D., M.S.
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Craig A. Umscheid, M.D., M.S.
Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Lionel L. Bañez, M.D.
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

# Acknowledgments

The authors gratefully acknowledge the following individuals for their comments on previous versions of this report: Ethan Balk, M.D., M.P.H., from the Brown Evidence-based Practice Center (EPC), Providence, RI; Mary Butler, Ph.D., M.B.A., from the Minnesota EPC, Minneapolis, MN; Jennifer S. Lin, M.D., from the Kaiser Permanente EPC, Portland, OR; and EPC Director Craig A. Umscheid, M.D., M.S., from AHRQ. The authors also gratefully acknowledge Edi Kuhn, B.S., B.A., at the Scientific Resource Center, Portland, OR, for administrative assistance and assistance with managing references. Finally, the authors gratefully acknowledge Edwin Reid, M.S., M.A.T., M.F.A. at the Scientific Resource Center, Portland, OR, for editorial assistance with this report.

# Key Considerations

- This document updates guidance on including nonrandomized studies of interventions (NRSIs) in systematic reviews of interventions. NRSIs are observational or experimental studies of the effectiveness and/or harms of interventions, in which participants are not randomized to intervention groups.
- NRSIs are a valuable source of information about the effectiveness and harms of interventions, not just a supplemental source to fill gaps in the evidence from randomized controlled trials (RCTs).
- In deciding whether to include NRSIs, consider the balance between improved utility to end-users and threats to validity based on the following questions:
    - What are the decisional dilemmas and Key Questions being addressed, and how will the end-users of the systematic review use the evidence to inform decision making?
        - Are the decisional dilemmas centered on efficacy, effectiveness, or harms? To what extent are RCTs and NRSIs likely to address these dilemmas? Are NRSIs likely to fill gaps in the RCT evidence base?
        - To what extent do available RCTs and NRSIs address the populations, interventions, comparators, and outcomes of the Key Questions?
        - Has the topic evolved in a way that increases or decreases the value of NRSIs?
    - Is it logical and likely for RCTs to have addressed the Key Questions adequately?
    - How serious is the risk of bias in NRSIs that address the Key Questions likely to be?
    - To what extent are NRSIs and RCTs likely to complement each other?
- When NRSIs are included:
    - Ensure that the systematic review team members are familiar with topic-specific data source considerations and advanced analytic methods for NRSIs.
    - Follow guidance specific to NRSIs in developing the protocol and conducting the systematic review.
    - Report decisions, methods, and results transparently.
    - Discuss strengths, limitations, and caveats of including, or not including, NRSIs.

# Inclusion of Nonrandomized Studies of Interventions in Systematic Reviews of Intervention Effectiveness: An Update

## Structured Abstract

**Introduction:** Nonrandomized studies of interventions (NRSIs) are observational or experimental studies of the effectiveness and/or harms of interventions, in which participants are not randomized to intervention groups. There is increasingly widespread recognition that advancements in the design and analysis of NRSIs allow NRSI evidence to have a much more prominent role in decision making, and not just as ancillary evidence to randomized controlled trials (RCTs).

**Objective:** To guide decisions about inclusion of NRSIs for addressing the effects of interventions in systematic reviews (SRs), this chapter updates the 2010 guidance on inclusion of NRSIs in Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) SRs. The chapter focuses on considerations for decisions to include or exclude NRSIs in SRs.

**Methods:** In November 2020, AHRQ convened a 20-member workgroup that comprised 13 members representing 8 of 9 AHRQ-appointed EPCs, 3 AHRQ representatives, 1 independent consultant with expertise in SRs, and 3 representatives of the AHRQ-appointed Scientific Resource Center. The workgroup received input from the full EPC Program regarding the process and specific issues through discussions at a virtual meeting and two online surveys regarding challenges with NRSI inclusion in SRs. One survey focused on current practices by EPCs regarding NRSI inclusion in ongoing and recently completed SRs. The other survey focused on the appropriateness, completeness, and usefulness of existing EPC Program methods guidance. The workgroup considered the virtual meeting and survey input when identifying aspects of the guidance that needed updating. The workgroup used an informal method for generating consensus about guidance. Disagreements were resolved through discussion.

**Results:** We outline considerations for the inclusion of NRSIs in SRs of intervention effectiveness. We describe the strengths and limitations of RCTs, study design features and types of NRSIs, and key considerations for making decisions about inclusion of NRSIs (during the stages of topic scoping and refinement, SR team formation, protocol development, SR conduct, and SR reporting). We discuss how NRSIs may be applicable for the decisional dilemma being addressed in the SR, threats to the internal validity of NRSIs, as well as various data sources and advanced analytic methods that may be used in NRSIs. Finally, we outline an approach to incorporating NRSIs within an SR and key considerations for reporting.

**Conclusion:** The main change from the previous guidance is the overall approach to decisions about inclusion of NRSIs in EPC SRs. Instead of recommending NRSI inclusion only if RCTs are insufficient to address the Key Question, this updated guidance handles NRSI evidence as a valuable source of information and lays out important considerations for decisions about the inclusion of NRSIs in SRs of intervention effectiveness. Different topics may require different decisions regarding NRSI inclusion. This guidance is intended to improve the utility of the final

product to end-users. Inclusion of NRSIs will increase the scope, time, and resources needed to complete SRs, and NRSIs pose potential threats to validity, such as selection bias, confounding, and misclassification of interventions. Careful consideration must be given to both concerns.

## Contents

# 1. Introduction

## 1.1 Rationale for Updated Guidance

Nonrandomized studies of interventions (NRSIs) are observational or experimental studies of the effectiveness and/or harms of interventions, in which participants are *not* randomized to intervention groups. This guidance document on the inclusion of NRSIs in Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) comparative effectiveness systematic reviews (SRs) updates guidance developed in 2010 by AHRQ EPC investigators on the inclusion of observational studies in comparative effectiveness SRs.[1, 2]

The 2010 guidance noted the need for routine consideration of the appropriateness of including observational studies and explicit documentation and justification for the decision. The guidance offered a framework for systematic reviewers ("reviewers") to consider when making the decision. Reviewers were guided to consider two primary questions: (1) Are there gaps in the evidence from randomized controlled trials (RCTs)? and (2) Will observational studies provide valid and useful information?[1, 2]

Of note, we are using the term NRSIs (instead of observational studies) because this guidance and the older 2010 guidance are specifically relevant for SRs of *interventions*. The need for careful consideration of the value and consequences of including NRSIs continues to be recommended by various methodologists.[3, 4] Nevertheless, there is now increasingly widespread recognition that, in many instances, advancements in the design and analysis of NRSIs allow NRSI evidence to have a much more prominent role in decision making,[5, 6] and not just as ancillary evidence to RCTs, as was suggested in the 2010 guidance.

In addition to methodological advancements in the field, the AHRQ EPC Program also periodically updates its methods to ensure that the needs of its end-users are being met. The EPC Program has served the needs of various guideline developers and, increasingly, Learning Health Systems (LHSs). In 2012, the Institute of Medicine defined an LHS as a system "in which science and informatics, patient-clinician partnerships, incentives, and culture are aligned to promote and enable continuous and real-time improvement in both the effectiveness and efficiency of care."[7] Over the years, many healthcare systems have embraced continuous learning, and large amounts of data are produced every day in patient-clinician encounters. NRSIs are one method of harnessing these nonrandomized data from non-research or so-called "real-world" settings.

The incorporation of evidence from NRSIs in comparative effectiveness SRs has the potential to inform decision making in LHS and healthcare more broadly. A recent AHRQ stakeholder engagement project with leaders of LHS around the United States identified some challenges with applying the findings of SRs produced by the AHRQ EPC Program to healthcare decision making.[8] LHS leaders noted that findings of SRs, especially those including few NRSIs or without consideration of local health system data, may often not be as useful for LHS decision making. A chief concern raised was the perception that evidence summarized does not adequately apply to the LHS's typical patient population.[8, 9] Including and considering the findings of NRSIs appropriately in EPC evidence syntheses has the potential to increase the applicability of EPC Program SRs to healthcare decision making by LHS and by healthcare decision makers more broadly.

## 1.2 Considerations When Deciding Whether To Include NRSIs

Although low-quality evidence on interventions may often be better than no evidence at all for important clinical and policy decision making, whether, and to what extent, to consider NRSIs in SRs of interventions requires carefully assessing multiple considerations. These considerations relate to the clinical topic, scope, and decisional dilemmas that the SR aims to address; the potential applicability of research studies to clinical practice; the uncertainty around the internal validity of NRSIs; the advantages and disadvantages of novel data sources and analytic methods; and the methods to be used when NRSIs are included (e.g., searching for NRSIs, assessing their risk of bias, and synthesizing, grading, and reporting evidence that includes NRSIs). This section introduces these considerations and refers to later sections in this guidance document that provide further elaboration.

### 1.2.1 Nature of the Clinical Decisional Dilemmas

Archie Cochrane, after whom the Cochrane Collaboration is named, is credited for articulating the distinction between the questions of “can it work?” (i.e., the extent to which an intervention does more good than harm under the most *ideal* circumstances; a question of efficacy) and “will it work?” (i.e., whether an intervention does more good than harm under the *usual* circumstances of routine practice; a question of effectiveness).[10] Although RCTs usually exist along a spectrum ranging from those highly focused on efficacy to those highly focused on effectiveness,[11] it is useful to consider whether the decisional dilemma being addressed in the SR pertains more to the efficacy or effectiveness of the interventions of interest.[12] Clinical dilemmas related to interventions generally pertain to their effectiveness. Most, if not all, SRs conducted as part of AHRQ EPC Program address complex multicomponent interventions and implementations, where the “will it work” (i.e., effectiveness) question is most important. For many of these SRs, no or only a few RCTs may be available, and where they exist, are likely to be narrow in scope, all of which increases the relevance of NRSIs.

### 1.2.2 Potential Applicability of Study Results to Clinical Practice

For intervention effectiveness, although RCTs, when conducted well, may be the study design that is methodologically strongest (i.e., least susceptible to bias), their findings may be less applicable than NRSIs to clinical practice. This may be true for several reasons. RCTs often have narrow participant eligibility criteria that exclude some subsets of the population, such as individuals with more severe disease and/or multiple comorbidities. RCTs often tightly control and monitor delivery of the intervention and comparator. Moreover, because of their generally smaller sample size and shorter duration than NRSIs, RCTs more often focus on short-term, intermediate (or surrogate), and/or composite outcomes. By contrast, findings of NRSIs, which often include broader patient populations, flexible intervention and comparator implementations, and longer-term outcomes, may be more representative of clinical practice.

### 1.2.3 Uncertainty Around the Internal Validity of NRSIs

An important consideration regarding the inclusion of NRSIs relates to their validity, particularly the extent to which NRSIs successfully address potential threats to validity, such as selection bias, confounding, and misclassification of interventions. NRSI inclusion in SRs presents a predicament when the evidence base includes older or smaller, poorly conducted NRSIs that may not have adequately accounted for these potential threats to validity. Section 7 expands on these issues.

### 1.2.4 Advantages and Disadvantages of Novel Data Sources and Analytic Methods

The recent decade has witnessed considerable advancements in methods for analyzing NRSIs, such as trial replication methods, propensity scores, and instrumental variable analyses. These methods greatly enhance the potential for causal inferences based on data from NRSIs, but most rely on important assumptions. Use of such methods remains somewhat uncommon. Moreover, interpretation of their results requires a level of familiarity with complex analytic methods that may require SR teams to acquire knowledge of how to assess and analyze the studies. Sections 8 and 9 expand on these issues.

### 1.2.5 Methods To Use When NRSIs Are Included

Inclusion of NRSIs (in addition to RCTs) increases the methodological complexity and resources needed to complete an SR. Planned methods for consideration of both types of evidence should be specified in the SR protocol. Searches should be tailored to ensure that NRSIs are also identified. If NRSIs are included in addition to RCTs, the number of studies to be screened and included in the SR would be expected to increase, often to a great extent, likely increasing the resource needs of the SR team. Tools to assess the risk of bias in NRSIs are different from those used to assess risk of bias in RCTs. The full use of some tools, such as the Risk of Bias in Nonrandomized Studies of Interventions (ROBINS-I),[13] requires advanced expertise in study design methodology and can be time-intensive, especially if the SR includes many NRSIs. The analysis, synthesis, and reporting of NRSIs also require special attention. Section 10 expands on these issues.

## 1.3 Scope of This Guidance

This guidance focuses on considerations involved in the decision to include or exclude NRSIs in SRs of interventions. The guidance targets both the benefits and harms of interventions that are intended to work at any level, such as the individual patient, the health system, or the broader population.

Substantial heterogeneity exists in how NRSIs are classified and described. In this guidance, we define an NRSI as an observational or experimental study of the effectiveness and/or harms of interventions, in which participants are *not* randomized to intervention groups. This definition is consistent with those used by others.[13] NRSIs may include experimental (i.e., investigator-assigned interventions) or non-experimental (i.e., observational) studies.

The sources used by NRSIs to obtain data are broad and include primary data collection in a research context as well as data collected for other purposes, such as electronic health records, patient registries, administrative data, claims data, and others. This guidance document addresses NRSIs regardless of the source or type of data collected and analyzed; this represents an expansion beyond the scope of AHRQ's 2010 report,[1] which was confined to observational studies.

# 2. Workgroup Methods

## 2.1 Composition of the Workgroup

In November 2020, AHRQ convened a workgroup to update the existing (2010) guidance and develop this document. The 20-member workgroup comprised 13 members representing 8 of 9 current AHRQ-appointed EPCs, 3 AHRQ representatives, 1 independent consultant with expertise in SRs, and 3 representatives of the AHRQ-appointed Scientific Resource Center.

## 2.2 Obtaining Input From the Broader EPC Program

The workgroup met virtually approximately twice a month for 10 months. The workgroup also received input from the broader EPC Program regarding the process and specific issues in the following ways:

- All nine current EPCs provided input through discussions at a virtual meeting (August 2021).
- Two online surveys examined challenges with including NRSIs in SRs. One survey focused on current practices by EPCs regarding NRSI inclusion in 19 ongoing and recently completed SRs. The other survey, completed by all EPCs following completion of each EPC SR in the last 3 years, focused on the appropriateness, completeness, and usefulness of existing AHRQ EPC Program methods guidance.

The workgroup considered virtual meeting and survey input when identifying specific aspects of the guidance that needed updating.

## 2.3 Generating Consensus

The workgroup used an informal method for generating consensus on the content of this report. Disagreements were resolved through discussion.

## 2.4 Writing This Report

Individual workgroup members drafted sections of this report. Two workgroup leads (IJS and MV) reviewed, revised, and compiled the sections into a draft report. All workgroup members commented on the complete draft before the draft was circulated to all EPC Directors and AHRQ officers. All workgroup members, led by IJS and MV, reviewed and addressed the comments and approved this final version of the report.

# 3. Strengths and Limitations of RCTs

The ideal study design for examining interventions has traditionally been high-quality RCTs. The randomization of study participants to treatment and comparator groups, when allocation is concealed, minimizes selection bias.[14] Prospective assignment of participants to study arms prior to outcomes occurring helps avoid selection bias that may arise in retrospective NRSIs if the selection of participants into the study is conditional on an outcome related to the intervention being evaluated (e.g., if only survivors are included). When properly executed, the randomization process helps ensure that the study groups are comparable with respect to known and unknown baseline prognostic factors (i.e., confounders).[15] RCTs particularly larger, better conducted studies, generally have registered and/or published protocols and may be required to report deviations from protocols. Such requirements may reduce the potential for bias arising from deviations from intended interventions and selective outcome reporting[13] and may help promote consistency of treatments and measurements that may be challenging in NRSIs, particularly retrospective NRSIs.

Despite the above advantages, RCTs bear some important limitations. *First*, RCTs may be underpowered to detect differences between comparators in harms. RCTs may be of limited value in the assessment of harms of interventions because RCTs are frequently small and/or of too short duration for uncommon harms or longer-term harms to be detected. Given the relatively small sample sizes and short duration of most RCTs, they are frequently underpowered to detect differences on several measures of effectiveness prioritized for a given SR. *Second*, in the situation of rare diseases, RCTs may have to draw from a very small population of interest, which may make enrollment very challenging. *Third*, it may be unethical to perform an RCT due to the absence of clinical equipoise or for other reasons. *Fourth*, it may be infeasible or highly resource-intensive to conduct an RCT, for example to examine very long-term outcomes, which may be more important to patients. Depending on the topic, RCTs may focus on interim (or surrogate) outcomes instead of clinically important or patient-important outcomes. *Finally*, results of some RCTs may not be broadly applicable due to their narrow eligibility criteria for participants, tightly controlled implementation of interventions and comparators, smaller sample size, shorter duration, and focus on short-term, surrogate, and/or composite outcomes. However, applicability may be less of a concern for some RCT designs (e.g., large simple RCTs, pragmatic RCTs).

# 4. Types of NRSIs and Study Design Features

## 4.1 Types of NRSIs

Unfortunately, there is no consensus on NRSI terminology and study categorization; different researchers may refer to the same design using different language.[16, 17] For example, a single-group study in which all participants received the same intervention has been called a before-after study, a pre-post study, a case series, or a cohort study. As another example, a comparative study in which investigators nonrandomly assigned participants to two or more intervention groups has been called a controlled clinical trial, a cohort study, or more generally, a nonrandomized comparative study. While the following text and table articulate the design elements of commonly named NRSI designs, we concur with other methodologists that, when conducting SR tasks, systematic reviewers should use study *methods* rather than study design names to differentiate among NRSI types.[4]

When considering study methods to help differentiate among NRSI types, we suggest the following five domains, although we recognize that this may not be a complete list:

1. Presence of a comparison group receiving a different intervention or not receiving an intervention (controlled vs. uncontrolled/single group)
2. Experimental nature (experimental [i.e., investigator assigns group] vs. non-experimental)
3. Type of control group (historic control vs. concurrent control vs. none)
4. Presence of follow-up over time (yes [i.e., longitudinal] vs. no [i.e., cross-sectional])
5. Temporality, in the case of longitudinal studies (prospective vs. retrospective)

Single-group studies generally cannot contribute information to SRs focused on comparing the effectiveness of interventions. However, some EPC Program comparative effectiveness SRs, such as a 2021 SR on breast reconstruction after mastectomy,[18] have included single-group studies for (noncomparative) quantification of the risks of harms of interventions. We have therefore included single-group studies in this section, but much of the rest of this guidance document (e.g., sections on threats to validity of NRSIs, data sources, and advanced analytic methods) focus on comparative NRSIs.

In the context of considering whether to include NRSIs, two of the most important NRSI characteristics to consider are whether the study has a control group, which could include historical controls ("controlled"), and whether the researchers conducted an experiment (i.e., the study was not purely "observational"). Accordingly, the types of NRSIs (and related design names) include:

**Controlled and experimental NRSIs**

Controlled clinical trial (also known as nonrandomized controlled trial)

**Controlled and non-experimental (i.e., observational) NRSIs**

Prospective cohort study
Retrospective cohort study
Case-control study

**Controlled and either experimental or non-experimental (i.e., observational) NRSIs**

Before-after study (also known as pre-post study)
Interrupted time series

**Uncontrolled and non-experimental (i.e., observational) NRSIs**

Case series (also known as uncontrolled single-arm study)
Case study (also known as case report)

Table 1 provides a summary of common NRSI study designs.

**Table 1. Common types of nonrandomized studies of interventions (NRSIs) based on study design features**

| Type | Design | Brief Description | Control | Control Type | Followup | Temporality | Strengths | Weaknesses | Readings |
|---|---|---|---|---|---|---|---|---|---|
| **Controlled and experimental NRSIs** | **Controlled clinical trial**<br>*(Also known as nonrandomized controlled trial)* | A trial in which participants or clusters of participants are allocated to the intervention or the comparator in a nonrandom fashion (e.g., based on disease severity, clinical history, time of admission). Like an RCT, often follows a well-defined study protocol and implements the intervention and comparator in a well-defined, closely monitored environment. | Yes | Concurrent | Longitudinal | Prospective | • Strict eligibility criteria<br>• Strict followup schedule<br>• Prospective design<br>• Possibility to blind/ mask participants to intervention<br>• Can measure incidence/ risk | • Prone to selection bias and confounding, which can limit causal inference<br>• May have poor generalizability (different from typical clinical practice environment)<br>• Often not suitable for rare outcomes due to typically short followup and small sample sizes<br>• Expensive and time consuming | Dávila-Fajardo 2017[19] |

| Type | Design | Brief Description | Control | Control Type | Followup | Temporality | Strengths | Weaknesses | Readings |
|---|---|---|---|---|---|---|---|---|---|
| **Controlled and non-experimental (i.e., observational) NRSIs** | **Prospective cohort study*** | A study that prospectively recruits and follows participants over time and defines study groups based on exposure (intervention) status. The allocation of study participants to the intervention and the comparison is non-experimental (i.e., it is based on the intervention received in routine practice, without influence of research personnel). | Yes | Concurrent | Longitudinal | Prospective | • Participants reflect routine practice<br>• Temporal association between intervention and outcome can be demonstrated<br>• Can measure incidence/risk | • Prone to selection bias and confounding, which can limit causal inference<br>• Cannot blind/ mask participants to intervention<br>• Often not suitable for rare outcomes<br>• More expensive and time-consuming than retrospective studies | Guyatt 2015,[20] Giovannucci 1993[21] |

| Type | Design | Brief Description | Control | Control Type | Followup | Temporality | Strengths | Weaknesses | Readings |
|---|---|---|---|---|---|---|---|---|---|
|  | **Retrospective cohort study*** | A study that retrospectively identifies study participants based on their exposure (intervention) status. Information on the intervention, comparator, and outcomes of interest are all obtained from historical data. | Yes | Concurrent | Longitudinal | Retrospective | • Participants reflect routine clinical practice<br>• Less expensive and time-consuming than prospective studies because followup has already completed<br>• Temporal association between intervention and outcome can be demonstrated<br>• Can measure incidence/ risk | • Prone to selection bias, confounding, and misclassification of interventions, which can limit causal inference<br>• Cannot blind/mask participants to intervention<br>• Data are generally not specific to research aim<br>• Not suitable for unanticipated outcomes | Guyatt 2015,[20]<br>Go 2017[22] |
|  | **Case-control study** | A study that compares participants with a specific outcome/disease (cases) with participants without the outcome/disease and evaluates the association between previous exposure (intervention) and the outcome. | Yes | Concurrent | Neither (unless nested in a prospective or retrospective cohort study or RCT) | Retrospective (unless nested in a prospective cohort study or RCT) | • Suitable for rare diseases or outcomes<br>• Less time consuming, cost effective<br>• Can evaluate multiple exposures | • Prone to recall bias and selection bias, which can limit causal inference<br>• Restricted to one outcome only<br>• Cannot measure incidence/risk<br>• Sometimes difficult to infer temporal association between the intervention and outcome | Guyatt 2015,[20]<br>Sedgwick 2015,[23]<br>Wiese 2018,[24]<br>Vergis 2001[25] |

| Type | Design | Brief Description | Control | Control Type | Followup | Temporality | Strengths | Weaknesses | Readings |
|---|---|---|---|---|---|---|---|---|---|
| **Controlled and either experimental or non-experimental (i.e., observational) NRSIs** | **Before-after study** *(Also known as pre-post study)*[†] | A study in which a single group of participants with outcomes evaluated before and after implementation of an intervention. | Yes | Historical (before versus after implementation of an intervention) | Longitudinal | Prospective or retrospective, or mixed | • Ease of participant enrollment | • Difficult to disentangle intervention effects from temporal changes (i.e., outcome changes irrespective of intervention) and "regression towards the mean" | Guyatt 2015,[20] Torgerson 2008,[26] Reignier 2009,[27] Austin 2003[28] |
| | **Interrupted time series** | A type of before-after study in which a single group of participants is observed multiple times (often at equally spaced intervals) before and after an intervention. Outcome of interest is measured as the difference of predicted changes before and after intervention. | Yes | Historical (before versus after intervention at multiple times) | Longitudinal | Prospective or retrospective | • Easier to adjust for temporal changes (i.e., outcome changes irrespective of intervention) | • Large sample size and number of time-points required<br>• Difficult to disentangle co-interventions when data collected are close in time | Ramsay 2003,[29] Hudson 2019,[30] Penfold 2013,[31] Lawes 2015,[32] Milder 2015,[33] Feldstein 2006[34] |

| Type | Design | Brief Description | Control | Control Type | Followup | Temporality | Strengths | Weaknesses | Readings |
|---|---|---|---|---|---|---|---|---|---|
| **Uncontrolled and non-experimental (i.e., observational) NRSIs** | **Case series** *(Also known as [uncontrolled] single-arm study)* | A study with outcomes evaluated after an intervention in a single group of participants. | No | N/A | May be longitudinal | Retrospective | • May be useful for rare diseases or newer interventions<br>• Ease of participant enrollment<br>• May provide best analyses of differential effects in different subgroups | • Can provide only event rates or changes in continuous measures on treatment<br>• Cannot infer association between intervention and outcome because of lack of a control group<br>• Prone to reporting bias | Buechner 2022[35] |
| | **Case study** *(Also known as case report)* | A description of a single participant who received an intervention. | No | N/A | May be longitudinal | Retrospective | • May be useful for rare diseases, atypical patients, or newer interventions | • Cannot infer association between intervention and outcome because of lack of control participants<br>• Prone to reporting bias<br>• Cannot measure incidence/risk<br>• Likely poorly applicable to most patients | Silva 2022[36] |

* The word "cohort" is sometimes used in the literature to describe a noncomparative study, but we are using it to denote comparative studies.

† A before/after study and a pre/post study are sometimes differentiated in that the former may refer to a study in which the same group of participants is assessed before and after they receive an intervention, while the latter refers to a study in which different groups of participants are assessed before and after an intervention is implemented. We are using these two study descriptors interchangeably.

This table does not include cross-sectional studies because we consider their methodology as indicative of a data collection approach and not as a study *per se*. Cross-sectional studies are non-experimental and may allow the comparison of groups of participants or may be uncontrolled case series. They lack participant follow-up and therefore do not allow inference regarding temporality of the association between interventions and outcomes. When cross-sectional data are used retrospectively to collect information on exposure, the potential for recall bias also limits the ability to infer an association between interventions and outcomes.

Abbreviations: N/A = not applicable, RCT = randomized controlled trial.

## 4.2 Caution Regarding NRSI Designs

When determining the eligibility of an NRSI, reviewers should not rely on study design labels provided by authors in reports of NRSIs.[4] Instead, reviewers should examine specific study methods (i.e., design features and analytic methods). Similarly, when assessing risk of bias, reviewers should avoid study design labels as surrogates for risk of bias.[37] Rather, reviewers should evaluate specific study methods that may have been used by the NRSI to mitigate specific types of bias (e.g., confounding, bias due to missing data).

# 5. Key Considerations for Including or Excluding NRSIs

Section 3 discussed the strengths and limitations of RCTs. A 2011 AHRQ methods report described a "best evidence" framework that outlined some important considerations when determining whether evidence from RCTs alone may meet the goals of an SR.[38] The report also noted that different topics may require different decisions regarding the inclusion of NRSIs.

We recommend that several factors be considered to help determine whether RCTs or NRSIs will likely form the primary evidence base or if RCTs and NRSIs may complement each other. These interrelated considerations include:

1. *What are the decisional dilemmas and Key Question being addressed, and how will the end user(s) of the SR use the evidence to inform decision making?*
2. *Is it logical and feasible for RCTs to address the Key Question either solely or primarily?*
    - *Population*: What are the condition and population of interest? Are they likely to be adequately represented in RCTs?
    - *Interventions and comparators*: Are they new or established? Do considerable variations exist in how they are implemented in clinical practice? Are the full ranges of interventions and comparators of interest likely to be adequately represented in RCTs?
    - *Outcomes*: What are the primary outcomes of interest, including benefits and harms? Are there outcomes of interest (e.g., long-term effectiveness, long-term harms) for which RCTs may not be suitable, available, or feasible? Particularly for harms, is a comparator group required to answer the Key Question?
    - *Settings*: What settings are of primary importance? Are they likely to be adequately represented in RCTs?
3. *How serious is the risk of bias in NRSIs that address the Key Question*?[38]
    - Is causal inference needed?
    - What level of methodologic rigor would be required of NRSIs to allow meaningful conclusions, meet the needs of the end-user, and comply with contemporary standards for SRs?
4. *To what extent do NRSIs and RCTs complement each other?*
    - Is randomization *required* to answer the Key Question, particularly regarding benefits?
    - Taken together, would the body of evidence from RCTs cover diverse populations and/or implementations of the intervention?

The above is not an exhaustive list of considerations. Useful information on many of these considerations may be apparent only during Topic Refinement or later. The decisions are not always straightforward and generally depend on the topic and Key Questions. In general, we agree with a strategy that considers risk of bias and applicability and assesses overall strength of evidence based on the best set of studies (i.e., the "best evidence" approach).[38] Whatever decision is made regarding including NRSIs, it is crucial for reviewers to be transparent in reporting the decision and a justification for it in the SR protocol and in the description of methods in the final report. Reporting study findings and considering their limitations (regardless of study design) remain of primary importance.

# 6. Applicability of NRSIs to the Key Questions

## 6.1 Addressing the SR's Decisional Dilemma

When considering whether to include NRSIs in an SR, it is important to assess the extent to which the specific research questions addressed in relevant NRSIs align with the SR's Key Questions. This assessment requires explicit consideration of whether the Key Questions, and hence the decisional dilemmas, concern efficacy, effectiveness, and/or harms.

As discussed in Section 1, RCTs, particularly explanatory RCTs, generally are designed to determine efficacy (i.e., the extent to which the intervention does more good than harm under the most *ideal* circumstances).[10] NRSIs that are comparative generally focus on effectiveness (i.e., whether the intervention does more good than harm under the *usual* circumstances of routine practice).[10] Comparative and single-group NRSIs also tend to provide evidence on the potential harms of an intervention that cannot be adequately gleaned from RCTs alone (see the EPC Methods Guide chapter Prioritization and Selection of Harms for Inclusion in Systematic Reviews[39]). Because most EPC Program SRs of interventions aim to address both effectiveness and harms, it is not appropriate to *routinely* exclude all NRSIs even when numerous RCTs are available.

## 6.2 Applicability of NRSIs to the Specific PICO

The Cochrane Non-Randomized Studies of Interventions Methods Group offers guidance on assessing how well NRSIs address SR questions defined in terms of the populations, interventions, comparators, and outcomes (PICO).[4] The assessment requires consideration of whether known available RCTs and NRSIs address a PICO-defined Key Question directly or indirectly. If RCTs address the Key Question directly and NRSIs address them only indirectly, it may be reasonable to exclude NRSIs from the SR. If both NRSIs and RCTs address the questions directly, it may be best to include NRSIs, especially if the RCT evidence is expected to contain gaps. However, NRSIs with high risk of bias may not be useful.

If the research question addressed in an NRSI (or RCT) deviates substantially from the PICO-defined SR question, it may be excluded. For example, in an SR of aspirin use for prevention of cardiovascular disease and colorectal cancer,[40] NRSIs (and RCTs) were excluded if they did not report on a comparison between aspirin and a non-aspirin control in a population eligible for primary prevention of cardiovascular disease.

Some outcomes may be particularly susceptible to measurement bias and confounding (see Section 7), so NRSIs that focus exclusively on such outcomes may be excluded if they report estimates of effectiveness or harms that are likely to be biased. For example, NRSIs, when compared with RCTs, have been shown to overestimate the benefits of treatments for pain.[41] In an EPC SR of treatments for acute and chronic pain,[41] reviewers focused on RCTs because of the susceptibility of NRSIs to confounding and bias for subjective outcomes, such as pain and function. However, the reviewers included large NRSIs for assessment of rare serious adverse events. For specific Key Questions where no RCTs were identified, the reviewers included cohort studies for evaluation of benefits.

## 6.3 Issues Related to Evolution of Evidence on a Topic

The decision regarding inclusion of NRSIs in an SR should consider the evolution of evidence in the topic area. Early in the evolution of evidence on a topic, when limited evidence from RCTs may be available, it is difficult to justify excluding NRSIs. In such a context, identification and synthesis of relevant NRSIs, perhaps with their limitations, could help articulate the need for RCTs.

As the evidence evolves, when RCT evidence accumulates, reviewers should consider whether and how NRSIs may have evolved to complement the evidence from the RCTs. The inclusion of NRSIs in SRs in this context could help (1) assess how the overall evidence applies to routine practice and specific subgroups of patients and (2) reveal the intervention's long-term effectiveness and/or harms.

When the evidence matures to a stage where large seminal RCTs are available, NRSIs may be excluded if the outstanding evidence gaps are either not particularly important or if the evidence from NRSIs is unlikely to alter conclusions gleaned from RCTs.

# 7. Threats to Internal Validity of NRSIs

Potential threats to the internal validity of NRSIs are important for deciding whether NRSIs should be included in an SR. Risk of bias refers to the likelihood that the estimate of an intervention's effect obtained from a study has a systematic error (i.e., bias) that leads to the estimate being different from the true effect.[42] Concerns around bias in estimates of effect are distinct from concerns around applicability, imprecision, and quality of reporting.[43] Applicability and imprecision are addressed in other SR processes in interpreting results, specifically, in strength of evidence assessments. Assessments of risk of bias in NRSIs should therefore focus on study design, conduct, and analysis,[43] with the caveat that poor study reporting can hinder risk of bias judgments.[43]

Sources of bias that are unique to NRSIs occur before or at the start of the intervention; sources of bias that occur after the intervention starts may be akin to those in RCTs.[13] Depending on the types and extent of biases and confounding in an NRSI, the magnitude or direction of the effects observed may be impacted, leading to spurious conclusions. Thus, it may not be helpful, and may even be problematic, to include an NRSI if its results are likely to be highly biased.[4] Results from biased studies can lead to misleading conclusions that could be used inappropriately by decision makers, especially if inadequate attention is paid to the potential biases.[44, 45] Therefore, it is reasonable to exclude NRSIs that do not adequately account for various potential biases.[16] Any such exclusions of NRSIs should be based on the most important design features for minimizing risk of bias. NRSIs should not be excluded based *solely* on study design labels (e.g., cohort study) because such labels are notoriously inconsistent (see Section 4).[46, 47]

The following subsections explore specific types of biases that are particularly relevant for *comparative* NRSIs. Assessing the quality of single-group NRSIs is beyond the scope of the current guidance.

## 7.1 Selection Bias

NRSIs may be subject to a high risk of *selection bias* if at study baseline some potentially eligible participants, or their followup time, were excluded from the treatment or comparator groups, and such exclusion may have led to a biased estimate of the treatment effect.[14] For example, consider a study using electronic health records, in which researchers defined the treatment group as those receiving a certain treatment for a certain disease and the comparator group as those receiving no treatment for the same disease. Selection bias could occur if the researchers excluded (for any reason) a larger proportion of patients with a greater risk of death due to comorbidities from the treatment group than the control group.

## 7.2 Confounding

NRSIs may be subject to a high risk of *confounding* if the treatment and comparator groups were imbalanced in terms of factors that were common causes of both the choice of treatment and the outcome. A confounder is a third variable that is associated with the treatment and a cause of the outcome but is not in the causal pathway between the treatment and the outcome (i.e., is *not* a mediator).[48] For example, in NRSIs of mental health treatments in pregnancy, women with greater symptom severity may be more likely to be treated with psychotropic medications and may also be more likely to experience adverse pregnancy outcomes (e.g., low infant birthweight or premature delivery) from the underlying condition (e.g., depression).[49] In this example, the effect of the intervention on pregnancy outcomes is confounded by the underlying severity of the condition.

## 7.3 Misclassification of Interventions

A source of bias that is unique to NRSIs, particularly retrospective NRSIs, relates to the misclassification of interventions. Intervention status may be misclassified (e.g., arising from an error in measurement) nondifferentially or it may be misclassified differentially, in terms of the outcome status. Differential misclassification is a particular problem because it can relate to the outcome.[48] If data on the intervention status are collected when the outcome or the risk of the outcome is known, differential misclassification of the intervention status may occur. Depending on knowledge or expectations of the outcome, participants or study personnel may overstate or understate participant exposure to the intervention. For example, participants in retrospective NRSIs of folic acid supplementation may be aware of the benefits of folic acid during early pregnancy.[50] Their experience of the pregnancy outcome may influence their recall of the extent of exposure to folic acid supplementation in early pregnancy (an example of "recall bias").

## 7.4 Other Sources of Bias

Other sources of bias in NRSIs are in common with RCTs. Some examples include bias due to missing data, bias in measurement of outcomes, and bias in selection of reported results.[13] However, assessment of some of these biases, such as bias due to missing data and bias in selection of reported results, may be challenging to evaluate for NRSIs because study protocols may not be available and reporting may be suboptimal.

# 8. Various Data Sources for NRSIs

In NRSIs, data sources vary and can include (1) *routinely collected data*, such as clinic records, electronic medical records, administrative claims data, and disease registries, and (2) *customized data*, such as study-specific visit data and patient-generated data, e.g., from fitness trackers or home medical equipment.

"Big data" is an ill-defined term that is increasingly used to describe large volumes of either routinely collected or customized data, as listed above. Studies that are conducted using big data offer the obvious advantage of very large sample sizes (often with many thousands of patients), potentially representing routine clinical practice well. Such data may permit evaluation of rare health outcomes or rare diseases and provide contextual information regarding effectiveness or harms. Because of large sample sizes, high precision of treatment effect sizes is often attained, which may or may not be clinically relevant. However, studies using big data are usually subject to the same sorts of threats to validity (e.g., confounding, selection bias) as studies using other data sources. Moreover, studies using big data are often prone to inaccuracies in diagnostic and intervention coding (i.e., misclassification of interventions and/or outcomes), inconsistent and/or incomplete follow-up data, and variability in reporting and interpretation.[51-54] For reviewers, an additional challenge is that subsets of the population in one big data study may overlap with other studies included in the SR. Reviewers may find it hard to detect and handle the potential double-counting arising from such overlap.

Understanding data sources and the context in which the data were generated can greatly help interpret the findings of NRSIs. For example, controlled clinical trials, in which participants are prospectively assigned to treatment groups by researchers without the use of randomization, usually obtain customized data using similar methods as in RCTs. However, NRSIs using administrative claims data, which are usually not gathered for the purposes of research, may lack important information regarding potential confounders. Additionally, there may be a substantial amount of missing data when sources such as patient-generated data are used. When such missing data are "informative," e.g., missing not at random (MNAR),[55] this can lead to findings that are subject to emigrative selection biases (i.e., bias due to post-baseline exclusion of some participants from the study for reasons related to both the exposure and outcome).[14, 56]

# 9. Advanced Analytic Methods for NRSIs

Previous guidance on the inclusion of NRSIs in EPC SRs recommended their inclusion as being conditional on RCTs being insufficient to address the research question and required NRSIs to fill the perceived gap.[1, 2] In effect, this approach handles NRSIs as a secondary source of information. Recent advances in causal inference call this handling into question. Specifically, improvements in design elements and analytic methods can support more sophisticated analyses. These methods may include *trial emulation* approaches, in which NRSIs carefully specify criteria to make the potential sources of bias transparent and clarify the concordance between results from NRSIs and RCTs, and *causal inference analytic* approaches, such as propensity scoring, instrumental variable, regression discontinuity, and difference-in-difference approaches, which facilitate causal inferences despite lack of randomization. This section explores these advanced analytic methods for comparative NRSIs.

## 9.1 Advanced Analytic Approaches: Trial Emulation Efforts

Trial emulation efforts aim to analyze NRSI data using designs that simulate a targeted, hypothetical RCT. Every attempt is made to emulate the features of the targeted RCT, except that randomization is not conducted.[57, 58] Early emulation studies have shown that discrepancies in effect estimates between RCTs and NRSIs that address a similar question are to a lesser extent the result of unmeasured confounding and can be largely explained by differences in study design and analytic methods, such as time since disease onset and duration of followup.[59-62] However, unmeasured confounding and other issues are still important and may not be adequately addressed in some trial emulation studies.[63, 64]

The Randomized Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology (RCT DUPLICATE) Initiative is a large systematic evaluation of the ability of NRSIs using routine clinical data to replicate RCTs.[65] The initiative aims to quantify differences between NRSIs that use routine clinical data and RCTs as well as the factors that may explain the differences. Results from the first 10 emulations focused on insurance claims data on cardiovascular outcomes of antidiabetic or antiplatelet medications. The results pertaining to the agreement between RCTs and NRSIs were mixed; 80 percent of the emulations achieved agreement in estimates[66]. Preliminary results in this limited clinical area support the conclusion that selection of active comparator therapies with similar indications and use patterns increases agreement between results of NRSIs and RCTs.[66]

Trial emulation studies have shown that NRSIs can be a supplemental source of high-quality evidence for answering questions of intervention effectiveness. Franklin and colleagues have provided general recommendations for evaluating quality and potential biases in trial emulation studies.[65] However, reliably distinguishing high-quality trial emulation studies from low-quality ones remains a challenge. Key considerations include the availability of data regarding confounders that may be imbalanced between study groups, whether the outcomes are defined similarly in each group, and whether the study power is sufficient to detect clinically meaningful differences. Confidence in the validity of trial emulation studies is increased if they report sensitivity analyses demonstrating minimal impact of the chosen design and analytic decisions.[65]

The inclusion of trial emulation studies in SRs is consistent with the goal of including the highest quality evidence. Some regulatory bodies, healthcare payors, health systems, and guideline developers consider trial emulation studies for drug approvals, drug labeling, formulary decisions, and evidence-based practice.[66-69] As data sources and statistical approaches improve, the opportunities for incorporating high-quality NRSIs that emulate the results of RCTs will continue to increase.[68] By understanding when and why results between NRSIs and RCTs might differ, reviewers can conduct "cross-design synthesis" by conducting meta-analyses across study designs to provide stakeholders with more pragmatic conclusions and valuable insights about the effectiveness of interventions that may not be gleaned by using RCT evidence alone.[68, 70]

## 9.2 Advanced Analytic Approaches: Causal Inference Methods

As noted in Section 3, randomization serves to prevent biases. Yet, NRSIs, under specific assumptions in which quasi-randomness occurs, can be analyzed in ways that facilitate causal inference. We discuss herein four approaches that, in our opinion, have the most merit for SRs: propensity scoring, instrumental variables, regression discontinuity, and difference-in-differences. The last three approaches capitalize on the existence of specific conditions that create quasi-random assignment of treatments, which allows for the estimation of causal effects even in the presence of selection bias and confounding. We highlight these methods because they are now commonly used advanced methods to evaluate causality. For each approach, we provide an explanation, the main assumption(s), and an example.

### 9.2.1 Propensity Scoring Methods

#### 9.2.1.1 Explanation

Propensity scoring is a set of analytic methods to adjust for observed confounders in an NRSI. The propensity score is defined as the conditional probability of receiving a certain intervention, given a set of covariates.[71] Yet, rather than maximizing the conditional probability of receiving the intervention, the primary purpose of propensity scoring is to balance the set of confounders between the two intervention groups.[72-76] Like all probabilities, the propensity score ranges from 0 to 1. The closer to 1, the stronger the probability that the participant would be in the intervention group; likewise, the closer to 0, the stronger the probability the participant would be in the control group.

There are two steps to a propensity score calculation. In the first step, an appropriate set of baseline covariates must be identified. Identification of relevant baseline covariates requires careful thought and should not include covariates simply because they are available in the dataset. Once the relevant covariates are identified, the second step involves estimation of each participant's probability of being treated (the propensity score). This can be done using such methods as binomial regression (using a logistic or probit model), statistical learning algorithms (classification trees or ensemble methods),[77, 78] and covariate balancing (which predicts treatment assignment while simultaneously optimizing covariate balance).[79]

Once the propensity score is calculated, one can use it as a covariate when estimating the treatment effect in a regression, matching, stratification, or weighting approach. A discussion of the strengths and weaknesses of each of these approaches is beyond the scope of this guidance.

#### 9.2.1.2 Assumptions

The major assumption required of a propensity score analysis is that of strongly ignorable treatment assignment (ignorability assumption). This assumption means that the treatment assignment and potential outcomes are conditionally independent, given the observed covariates.[71, 72] In other words, if the important confounders are identified, the only difference between the treatment and control groups is the treatment. If this assumption holds, the propensity score analysis is considered to produce unbiased estimates of the treatment effect. The ignorability assumption is fulfilled if all important covariates are identified (i.e., if there are no important unobserved confounders). Failure to identify all important confounders is a major limitation of propensity score analyses.[71, 72]

### 9.2.1.3 Example

Assessment of the effects of mental health treatments in pregnancy serves as a common example of the use of propensity score analysis.[80, 81] RCTs of psychotropic medications in pregnant women are rare. As a result, the evidence base relies on NRSIs. As discussed in Section 7, confounding is an important threat to validity of NRSIs. Propensity score analysis attempt to address confounding by modeling receipt of treatment on a wide range of covariates that are carefully selected. A key factor in predicting treatment receipt that may often be absent from large databases is the severity of the underlying condition. As a result, propensity score analyses may lack adjustment for severity or require the use of severity proxies, such as number of diagnoses. As with other analytic approaches, propensity score analyses are limited by the availability and completeness of data representing all confounders.

## 9.2.2 Instrumental Variable Approach

### 9.2.2.1 Explanation

NRSIs may choose to use an instrumental variable (IV) analytic approach when randomization is not feasible. In regression analysis of such studies, an IV refers to a variable that meaningfully relates to the intervention but affects the outcome only through the intervention (i.e., the IV has no direct effect on the outcome).[82] In such a context, an IV approach can simulate randomization because any variation in the outcome that is associated with variation of the IV is effectively due to the intervention. Randomization can be considered as the quintessential IV; random assignment occurs for the subset of the sample that received the intervention due to variation in the instrument and thus provides a causal treatment effect for this subset.

### 9.2.2.2 Assumptions

The IV analytic approach requires two main assumptions.[83] The relevance assumption refers to the existence of an association between the IV and the intervention variable. The exclusion assumption refers to the lack of a direct association between the IV and the outcome variable.[83]

### 9.2.2.3 Example

The effect of cardiac catheterization on mortality from acute myocardial infarction (AMI) is a well-known example of use of the IV analytic approach.[82, 84] The IV used was the *additional* distance that a patient must travel beyond the nearest hospital to get to a hospital that performs catheterization ("distance difference"). The relevance assumption holds because a smaller distance difference also makes it more likely that the patient received catheterization because the additional barrier of travel is lower. The exclusion assumption holds because the IV (i.e., the distance difference) is considered unrelated to mortality in the study, except through catheterization. Because the distance difference thus quasi-randomly "assigns" the intervention to patients, the IV approach can estimate the effect of catheterization on mortality in the subset of patients that received catheterization because they were closer to a catheterization hospital.

Regarding the assumptions in this example, the relevance assumption is fulfilled because a meaningful relationship can be demonstrated between the distance difference and catheterization. However, one cannot test the exclusion assumption due to the same reason that an IV is required, i.e., the relationship between the intervention variable and the outcome may be confounded by unobserved variables. Therefore, any test for a direct effect of the instrument on the outcome will also be confounded by these same unobserved variables. Here, expert knowledge should attempt to rule out the existence of any direct effect of differential distance on AMI mortality.

Two challenges with the IV approach worth noting are (1) it is often difficult to find an appropriate IV for a given research question, and (2) the results of IV analyses may be biased if the assumptions, which are often unverifiable, are not fulfilled.

## 9.2.3 Regression Discontinuity Design Approach

### 9.2.3.1 Explanation

A regression discontinuity design uses a threshold or cutoff point to assign an intervention to those on one side of the threshold and no intervention to those immediately on the other side of the threshold.

### 9.2.3.2 Assumption

The assumption is that study participants who are close to the threshold, on either side, are comparable in factors other than receipt of the intervention, making the intervention assignment arguably random.

### 9.2.3.3 Example

To investigate the effect of the human papilloma virus (HPV) vaccine on adolescent sexual behavior, researchers assessed a policy in Ontario, Canada, which made girls born after December 31, 1993, eligible for the vaccine and girls born on or before that date ineligible.[85] Girls born on or close to the date threshold (on either side) are arguably similar because small differences in birth date are not expected to affect sexual behavior. Because of the comparability in the populations before and after the policy change date, the regression discontinuity design could robustly estimate the effect of the HPV vaccine on health outcomes by comparing those born just before and just after the date threshold.

A limitation of this approach is that although known confounders (i.e., observable characteristics) can be compared between the two groups, unknown confounders (i.e., unobserved characteristics) cannot be compared to rule out unmeasured confounding. In this study, researchers tested whether girls born on either side of the date were dissimilar on observable characteristics but could not rule out the possibility of dissimilarities between the groups on unobservable characteristics.

### 9.2.4 Difference-in-Difference Approach

#### 9.2.4.1 Explanation

A difference-in-difference (DiD) approach compares the changes in an outcome over time between a group that received the intervention and a group that did not receive it. The approach adjusts the treatment effect estimate for factors other than the intervention that may also impact the outcome. The first difference (within-group) controls for time-invariant differences within the intervention group. The second difference (between-group) controls for time-varying factors that are common across groups.

#### 9.2.4.2 Assumption

The important assumption is that the outcome changes that occurred within the intervention group would have been the same for the control group had the intervention group not been treated (i.e., the counterfactual). If this assumption, known as the common trends assumption, is not fulfilled, the DiD approach would lead to a biased estimate of the treatment effect.

#### 9.2.4.3 Example

An example of use of the DiD approach was the estimation of the effect of the 2014 state-level expansions in Medicaid insurance coverage under the Affordable Care Act on clinical outcomes.[86] Researchers found that expanded coverage decreased the proportion of men with high prostate-specific antigen (PSA) results at the time of cancer diagnosis, which suggested that Medicaid expansion improved access to screening. Some states expanded Medicaid coverage, while others did not. The researchers estimated the difference between the proportion of men with high PSA scores after versus before 2014 and compared that difference between states that did and did not expand Medicaid coverage. While it is not possible to directly test the counterfactual, i.e., whether trends would be the same in Medicaid non-expansion states and expansion states in the absence of the expansion, it is possible to indirectly test the common trends assumption. For instance, a visual inspection of a graph of PSA trends for expansion and non-expansion states in the years before the policy implementation suggested common trends.

## 9.3 Summary of Considerations Regarding Advanced Methods

The advanced methods for NRSIs described above largely attempt to simulate randomization by balancing population characteristics for the intervention and the comparator groups on observed confounders (e.g., trial replication, propensity scores) and/or unobserved confounders (e.g., IV approach). In theory, such methods support making causal inferences regarding the impact of an intervention on outcomes of interest. As discussed, validity of the inferences made using these methods depends on the validity and robustness of underlying assumptions (e.g., the exclusion assumption in the IV approach) and modeling used. In practice, not all confounders are available or can be measured. Authors of NRSIs may lack the requisite expertise to appropriately apply the often-complex analytic methods and to verify all assumptions, some of which may be unverifiable. Moreover, evaluation of these assumptions and methods is often challenging, if not impossible, for most reviewers because it requires advanced knowledge of specific statistical modeling methods and a deep understanding of the study context and data structure. Another important consideration for reviewers is that modeling assumptions are often unverifiable due to lack of access to original datasets and/or inadequate reporting. We are not aware of risk of bias assessment tools that specifically evaluate the potential flaws of these advanced designs. Reviewers should acquire expertise or consult experts in these methods when including these types of NRSIs in an SR.

# 10. Incorporating NRSIs in Systematic Reviews

## 10.1 Planning for the Inclusion of NRSIs

When making the decision of whether to include NRSIs in an SR, traditional evidence hierarchies are not as relevant as specific study design features. As noted in Section 4, we agree with other methodologists that when considering NRSIs, reviewers should, instead of relying on study design labels (e.g., cohort studies), evaluate study methods and analytic approaches.[3, 4]

At a minimum, decisions to include or exclude NRSIs should be explained or justified as part of the SR protocol. The Cochrane Handbook for Systematic Reviews of Interventions describes some of the leading reasons to include NRSIs in SRs, many of which are related to the common limitations of RCTs or their absence.[4] For example, as we discuss in Section 3, large RCTs with long-term outcomes may not be conducted for ethical, practical, and/or resource-related reasons. Sometimes, even when RCTs are conducted, there may not be an adequate number of well-conducted RCTs addressing rare diseases or certain subpopulations. Even when well-conducted RCTs are available, they may not replicate typical clinical practice and outcomes as closely as NRSIs might. Low-quality evidence may often be better than no evidence at all for important clinical and policy decision making. However, as we discuss in Section 7, NRSIs have a higher susceptibility to confounding and other biases. A heavily biased estimate can be worse than no estimate because it could lead to erroneous conclusions and preclude higher quality and more reliable research. On the other hand, well-conducted NRSIs, such as those that are well-analyzed or use the advanced methods described in Section 9 carefully and with relevant assumptions fulfilled, may be less prone to confounding and other biases.

## 10.2 Developing Searches for NRSIs

"Hedges," also known as filters, are standardized search strategies that can be used to help retrieve relevant articles from electronic databases. Hedges are applied to improve the retrieval of various kinds of evidence, such as NRSIs. They are often used to identify study designs and, to a lesser extent, clinical concepts, such as treatment, diagnosis, and prognosis. To our knowledge, there are no published NRSI hedges with greater than 92-percent sensitivity,[87] so they should be used with some caution. See Appendix A for a sample of hedges for some common NRSI designs.

Another option is to use a hedge that eliminates unwanted publication types. These hedges retrieve records describing a broad range of study types while filtering out those that do not include primary research. We are not aware of published hedges for this type of exclusion. To maximize sensitivity, it may be best to not use hedges. In this case, machine-learning based screening tools, such as Abstrackr (http://abstrackr.cebm.brown.edu), that prioritize unscreened records based on the manual labels of previously screened citations, may be a particularly useful alternative.[88]

Often, NRSIs are used to identify evidence providing data on long-term adverse events. Adverse events hedges limit not by publication types but to studies reporting any type of adverse event. If NRSIs are being used only to evaluate harms, these hedges can be a good option. Appendix A includes examples of published adverse event hedges.

## 10.3 Assessing the Risk of Bias in NRSIs

Existing instruments for assessing the risk of bias in NRSIs[89, 90] vary in their (1) theoretical and empirical foundations; (2) comprehensiveness when considering sources of bias for a range of study designs and analytic approaches; (3) validation, documentation, and ease of use (length/complexity); and (4) presentation and transparency of the risk of bias assessments. Notably, although current AHRQ guidance does not recommend a specific tool for use in SRs, the guidance suggests that the chosen tool should:

- Be specifically designed for the study designs being evaluated
- Allow transparency in how assessments are made
- Be based on theory and supported by empirical evidence
- Avoid the presentation of risk of bias assessments as a numerical score.[43]

Sources of bias that are uniquely relevant to comparative NRSIs are described in Section 7 of this guidance. They include the potential for selection bias, confounding, and misclassification of interventions. It is worth noting that risk of bias assessments should focus on domains that contribute to bias and, as such, a well-conducted NRSI may sometimes be of better methodological quality than a poorly conducted RCT.

A key consideration in assessing the risk of bias in NRSIs is that topic-specific expertise is required to identify relevant confounders. Therefore, reviewer teams should include a mix of methodologic and content-specific expertise.

## 10.4 Interpreting Results From NRSIs

As discussed in Section 7, when interpreting results of NRSIs, it is important to remember that confounding is a key threat to validity. Across NRSI designs and analytic approaches, the ability to successfully account for confounding can vary greatly. For example, a before-after study (with a historic control group) may not be able to disentangle temporal changes of outcomes independent of the intervention being tested, while a prospective cohort study (with a concurrent control group) does not have the same level of challenge. The ability of an individual NRSI to adjust for confounding also depends on availability of data on the confounders, their precise and valid measurement, and analytic approaches used. As a result, poor adjustment (i.e., inadequate or overadjustment) of confounding in NRSIs can lead to bias, which can overestimate or underestimate the treatment effect, sometimes greatly.[13] Therefore, when including NRSIs, it is important to evaluate the extent to which confounding has been considered and effectively addressed. A statistic, known as the E-value, has been proposed to indicate the potential for an unmeasured confounder to have impacted the results for a given outcome in a study. The E-value has been defined as the minimum magnitude of association that an unmeasured confounder would need to have with both the intervention and the outcome to fully explain away a treatment effect, conditional on the other measured covariates.[91] The larger the E-value the larger the magnitude of unmeasured confounding would need to be to explain away an effect estimate.[91]

## 10.5 Incorporating Data From NRSIs Into Meta-Analyses

When interpreting data regarding a treatment effect from an NRSI for the purpose of meta-analysis, reviewers should consider: (1) whether the estimate was adjusted; (2) whether and how important confounders were handled in the design and analysis; (3) whether the confounders were measured in a precise and valid way; and (4) whether underlying assumptions of the analytic approach were evaluated and validated. Experts, such as clinical experts, statisticians, and SR methodologists, should assess whether the NRSI adequately adjusted for important confounding and whether estimates from NRSIs should be combined with those from other NRSIs and RCTs.

There sometimes is heterogeneity in estimates of treatment effect between RCTs and NRSIs. This methodological heterogeneity can be due to many factors, including differences in risks of selection bias, confounding, and other biases, potential treatment effect modifiers, and sampling error. However, empirical evaluation of such heterogeneity is limited. Heterogeneity statistical indicators (e.g., $I^2$, $H^2$) and statistical tests (e.g., Cochran's Q test) only evaluate statistical variations of *observed* treatment effects and do not capture the true uncertainty of the underlying true treatment effect.[92, 93] Investigating sources of heterogeneity through subgroup analysis or meta-regression is, by nature, exploratory and suffers from multiple potential shortcomings, such as the lack of sufficient detail reported in the included studies, small numbers of studies, and (in the case of meta-regression) collinearity.[92, 94]

When meta-analysis is deemed appropriate, reviewers should be cautious of combining different NRSI designs and/or analytic approaches or comparing NRSIs with RCTs. The advantages of conducting meta-analysis include the attainment of a singular overall estimate of the treatment effect that is relatively precise and based on a broader set of participants, with a potentially greater strength of evidence.[95] However, it is more likely than for meta-analyses of only RCTs that studies at higher risk of bias will be included with studies at lower risk of bias in the SR, which could lead to more biased results. Moreover, because of their generally large sample sizes, effect size estimates from individual NRSIs will generally be more precise and therefore will be assigned greater weights in a meta-analysis that weights studies using the inverse variance method.[95]

As an initial step, reviewers should examine the consistency of study findings between different NRSI designs/analytic approaches and between NRSIs and RCTs. Graphical displays, such as forest plots without an overall summary estimate, can be used to visually assess consistency in the direction and magnitude of treatment effect estimates and their confidence intervals. If applied, formal statistical tests, such as meta-regression and Cochran's Q test, should be used in concert with the considerations listed above (rather than as litmus tests to indicate the presence or absence of notable heterogeneity). Sensitivity analyses around study quality may be important.

Regardless of the study designs being analyzed, when deciding to combine study findings quantitatively (i.e., in a meta-analysis), considerations should include, but should not be limited to, similarity of the included studies in terms of population, intervention, comparator, outcome, timing, and settings (PICOTS) and type of NRSI design and analytic approach. When meta-analyses include studies of different designs, reviewers should present subgroup analyses by study design (at a minimum, RCTs vs. NRSIs). It may also be appropriate to conduct sensitivity analyses that exclude high-risk of bias NRSIs to avoid overestimating the strength of evidence. As a general rule, reviewers should not use statistical tests or indicators of heterogeneity (e.g., $I^2$, $H^2$, Cochran's Q test) purely as litmus tests to determine the appropriateness of conducting a meta-analysis.[95]

If different NRSI designs and analytic approaches present consistent effect estimates and confidence intervals, and if they are generally consistent with RCT effect estimates and confidence intervals, it may be appropriate to meta-analyze all the studies. However, reviewers should also present meta-analyzed results within subgroups by study design.[92]

If different NRSIs designs and analytic approaches and/or RCTs present inconsistent effect estimates and confidence intervals, in most instances, reviewers should avoid meta-analyzing estimates across study designs. Instead, evidence and associated heterogeneity should be reported separately, and the sources of inconsistency should be investigated, if possible, through subgroup analysis (or meta-regression). The "best evidence" approach is useful to select a body of evidence for investigation; investigators should decide whether RCTs, NRSIs in general, or specific NRSI design and analytic approaches represent lower risk of bias and better applicability to clinical practice. AHRQ's 2011 Methods Report–A Framework for "Best Evidence" Approaches in Systematic Reviews–provides detailed discussion regarding this approach.[38]

## 10.6 Grading the Strength of the Body of Evidence That Includes NRSIs

### 10.6.1 GRADE

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system is widely used to rate the certainty of evidence identified in SRs.[96] Domains of evidence considered in this system include study design, risk of bias, indirectness, inconsistency, imprecision, publication bias, dose-response, and magnitude of effect. Ratings for each domain feed into ratings for the certainty of the overall body of evidence for a given outcome; certainty may be rated as high, moderate, low, or very low.

### 10.6.2 AHRQ EPC Program Approach to Grading the Strength of a Body of Evidence That Includes NRSIs

The AHRQ EPC Program adopted a modified version of the GRADE system for use in EPC SRs.[97] The main modification is that applicability is separated from the indirectness domain to be an independent domain. This decision stems from the wide remit of EPC Program SRs that may have a diverse set of end-users with potentially unique parameters for judging applicability. The Program also uses the term "insufficient" rather than "very low" to describe the lowest level of evidence.[97]

Information in the rest of this section summarizes guidance specific to NRSIs that was provided in the 2013 AHRQ EPC Methods Guide Update.[97]

#### 10.6.2.1 Developing the Protocol

When developing the SR protocol, reviewers should establish *a priori* criteria to identify studies with design elements that would constitute an unacceptably high risk of bias (e.g., lack of adjustment for confounders). In addition to specifying the rationale, procedures, and decision rules, reviewers should explicitly describe the processes for synthesizing evidence from RCTs and NRSIs and determining overall strength of evidence.

#### 10.6.2.2 Rating Strength of Evidence Domains

For each outcome and intervention comparison of interest, when both RCTs and NRSIs are identified, reviewers should describe whether evidence from NRSIs agrees or conflicts with evidence from RCTs, provide potential reasons for any differences, and note pertinent limitations in both types of evidence. Reviewers do not need to assess the publication bias domain for NRSIs because methods to detect this among NRSIs are less certain than among RCTs. However, NRSIs may be susceptible to publication and other reporting biases because NRSIs are usually not registered *a priori*. The Real World Evidence Registry is a new registry that attempts to address this problem.[98]

The 2013 guidance also recommends the consideration of three additional domains for NRSIs: dose-response relationship, magnitude of treatment effect, and potential confounding that could impact the observed treatment effect (see Table 3 of the 2013 EPC Methods Guide).[97]

#### 10.6.2.3 Establishing an Overall Strength of Evidence

According to the 2013 guidance, evidence from NRSIs is generally assumed to suffer from a relatively higher risk of bias due to the lack of randomization and higher potential for confounding.[97] Thus, an initial provisional grade of low strength of evidence is assigned to evidence from NRSIs. Reviewers may increase the grade to moderate strength of evidence (although rarely high) if the evidence from NRSIs is rated as low for the Study Limitations domain (based on study conduct or analysis) or after assessing the additional domains.

When both NRSI and RCT evidence exist, reviewers may combine those design-specific strength of evidence grades into one overall strength of evidence grade or rely on one study design if it clearly provides stronger evidence. In general, the guidance allows reviewers the flexibility of using varied approaches to incorporate multiple domains into an overall strength of evidence grade as long as the rationale is clear and consistent and adheres to the important general principles of AHRQ EPC methods guidance.[97]

### 10.6.3 Other Approaches to Grading the Strength of a Body of Evidence

In addition to GRADE and the AHRQ EPC approach, various systems have been used for specific health topics or settings (e.g., Strength of Recommendation Taxonomy [SORT] for primary care,[99] Highest Attainable Standard of Evidence [HASTE] for HIV/AIDS,[100] Let Evidence Guide Every New Decision [LEGEND] for point-of-care[101]). A review of these systems is beyond the scope of the current guidance.

## 10.7 Reporting NRSI Evidence

When reporting findings from a synthesis of evidence that involves NRSIs, reviewers should be cautious and provide sufficient context regarding the strengths and limitations of all included studies. In general, making causal inferences from NRSIs should be done cautiously. We suggest that, unless there is substantial confidence in the NRSI design and analytic methods, their results should be interpreted as *associations* between an intervention and outcome and not as effects of the intervention on the outcome. Although well-designed NRSIs with adequate analytic methods (including multivariable regression or any of the advanced methods discussed in Section 9) reduce the potential impact of confounding and may come close to emulating an RCT, reviewers may be unlikely to encounter advanced analytic methods for most topics. Moreover, specialized expertise may be required for carefully interpreting findings of advanced methods. However, such methods as multivariable regression may be more common and will often be adequate for control of confounding. These approaches rely on the assumption that the full set of confounders is known and validly measured.

In general, NRSIs and, if any, RCTs, should be reported together when reporting findings for a given outcome for a given Key Question. In doing so, evidence with lower risk of bias and greater applicability to the population of interest should be prioritized. Regardless of whether meta-analysis is conducted, it is important to report consistency of the findings (in terms of direction and magnitude of treatment effects) among NRSI designs and between NRSIs and RCTs. Where inconsistencies are detected, their likely sources should be explored and discussed. Reviewers should also describe the extent to which NRSIs may have used appropriate analytic methods to address confounders and other important threats to validity.

# 11. Conclusion

## 11.1 Summary of Guidance

This updated AHRQ Program guidance on including NRSIs in SRs of interventions replaces AHRQ's 2010 guidance on selecting observational studies for such SRs.[1, 2] The main change from the previous guidance is the overall approach to the decision of including NRSIs. Instead of recommending NRSI inclusion *only* if RCTs are insufficient to address the Key Question, the current guidance considers NRSI evidence to have a prominent role in decision making.

This change has significant consequences. Although it may improve the utility of the final product to end-users, it is likely to require a greater scope, time, and resources needed to complete the SR than otherwise would have been needed if NRSIs were excluded from the SR. In balancing these tradeoffs, a crucial concern pertains to the potential threats to validity of NRSIs. This guidance document lays out key considerations in deciding whether and how to include NRSIs in an SR of interventions. Table 2 summarizes these considerations.

**Table 2. Key considerations for the inclusion and management of NRSIs in systematic reviews of interventions**

| SR Step | Key Considerations |
|---|---|
| **Topic scoping and refinement** | Consider the following questions at a minimum in evaluating the potential utility of RCTs and NRSIs:<br>1. What are the decisional dilemmas and KQs being addressed, and how will the end-user(s) of the SR use the evidence to inform decision making?<br>o Are the decisional dilemmas centered on efficacy, effectiveness, or harms? To what extent are RCTs and NRSIs likely to address these dilemmas? Are NRSIs likely to fill gaps in the RCT evidence base?<br>o To what extent do available RCTs and NRSIs address the populations, interventions, comparators, and outcomes of the KQs?<br>o Has the topic evolved in a way that increases or decreases the value of NRSIs?<br>2. Is it logical and likely for RCTs to have addressed the KQs adequately?<br>3. How serious is the risk of bias in NRSIs that address the KQs likely to be?<br>4. To what extent are NRSIs and RCTs likely to complement each other? |
| **SR team formation** | When NRSIs are included:<br>5. Ensure that the SR team members are familiar with topic-specific data source considerations and advanced analytic methods for NRSIs. |
| **Protocol development** | 6. Specify study design methods or features eligible for the SR.<br>7. Explain whether and why NRSIs will be included or excluded.<br>8. Explain the potential implications of the decision to include or exclude NRSIs.<br>9. When NRSIs are included, describe the processes for synthesizing evidence from RCTs and NRSIs and determining overall strength of evidence. |
| **Conduct of the SR** | When NRSIs are included,<br>10. Use appropriate search hedges to capture all relevant study designs.<br>11. Assess risk of bias using an appropriate tool for NRSIs.<br>12. Account for risk of bias when interpreting NRSI findings.<br>13. Evaluate and explain heterogeneity between RCTs and NRSIs, regardless of whether meta-analysis is conducted.<br>14. Avoid meta-analyzing results from one type of NRSI with results from other types of NRSIs or RCTs if there is considerable heterogeneity in findings.<br>15. Consider conducting sensitivity analyses when meta-analyzing high risk of bias NRSIs with other NRSIs.<br>16. Grade strength of evidence for NRSIs using established guidance. |
| **Reporting the SR** | When NRSIs are included:<br>17. Prioritize lower risk of bias and higher applicability findings in the synthesis regardless of study design, but consider presenting findings from RCTs and NRSIs in the same section for a given outcome for a given KQ.<br>18. Report on consistency between RCTs and NRSIs. Explore and discuss sources of heterogeneity in findings across study designs.<br>19. Discuss strengths, limitations, and caveats of including, or not including, NRSIs. |

Abbreviations: KQ = Key Question, NRSI = nonrandomized study of interventions, RCT = randomized controlled trial, SR = systematic review.

## 11.2 Context for Guidance and Challenges Encountered

AHRQ convened the EPC Program workgroup that developed this guidance with the goals of (1) improving the consistency of methodological approaches and (2) reporting a framework to guide the decision of inclusion of NRSIs for addressing the benefits and intended effects of interventions in EPC comparative effectiveness SRs.

The workgroup's discussions and the drafting and revising of this report, with subsequent input from directors of various EPCs, occurred in the context of an evolving climate regarding how the global SR community and, more broadly, the global intervention research community view studies in which participants are not randomized. This presented challenges because, naturally, there were various points of view on this issue within and beyond the workgroup. We view this as a strength of the process. Members of the workgroup had engaged in months of rigorous discussions and reading related to how NRSIs should be handled in SRs of intervention effectiveness.

Relatively early in the workgroup's discussions, it became clear that a one-size-fits-all approach to this guidance was neither desired nor appropriate. The current guidance therefore does not require that *all* SRs follow the same decision pathway. The workgroup decided to abandon the decisional framework (flow diagram) recommended in the 2010 guidance because the workgroup believes that the considerations for the decision regarding the inclusion of NRSIs in SRs are currently too complex to be fully captured in a figure. Instead, we call for flexibility in the decision making and lay out considerations that are intended to guide the decision. Different topics may require different decisions regarding NRSI inclusion

Another challenge worth repeating, as discussed in Section 1, is that AHRQ EPC Program SRs, taken together, inform decisions of a diverse set of decision-making bodies. However, many SRs are conducted to inform decisions of specific end-users, such as a guideline developing body or an LHS. End-users of most SRs are likely to be interested in the question of whether an intervention works in the "real world," which, due to limitations of RCTs, may necessitate the inclusion of NRSIs. Such variability in end-users across individual SRs would make universal guidance regarding NRSI inclusion in EPC SRs inappropriate.

The workgroup anticipates that, as is generally true, this guidance will need updating in a few years. In the interim, we hope this document is useful to facilitate decision making regarding the inclusion or exclusion of NRSIs in intervention SRs within the AHRQ EPC Program and beyond.

# References

1. Norris S, Atkins D, Bruening W, et al. Selecting Observational Studies for Comparing Medical Interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD); 2010.

2. Norris SL, Atkins D, Bruening W, et al. Observational studies in systematic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program [corrected] J Clin Epidemiol. 2011 Nov;64(11):1178-86. doi: 10.1016/j.jclinepi.2010.04.027. PMID: 21636246.

3. The Joanna Briggs Institute. Joanna Briggs Institute Reviewers' Manual: 2014 edition. 2014 ed. South Australia: The Joanna Briggs Institute.

4. Reeves BC, Deeks JJ, Higgins JPT, et al. Including non-randomized studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions. version 6.2 (updated February 2021) ed.: Cochrane; 2021.

5. Franklin JM, Pawar A, Martin D, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. Clin Pharmacol Ther. 2020 04;107(4):817-26. doi: 10.1002/cpt.1633. PMID: 31541454.

6. Franklin JM, Platt R, Dreyer NA, et al. When Can Nonrandomized Studies Support Valid Inference Regarding Effectiveness or Safety of New Medical Treatments? Clinical Pharmacology & Therapeutics. 2022;111(1):108-15. doi: 10.1002/cpt.2255. PMID: 33826756.

7. Institute of Medicine. Best care at lower cost: the path to continuously learning health care in America. In: Smith M, Saunders R, Stuckhardt L, McGinnis JM, eds. Washington (DC): National Academies Press; 2013.

8. Borsky AE, Savitz LA, Bindman AB, et al. AHRQ Series on Improving Translation of Evidence: Perceived Value of Translational Products by the AHRQ EPC Learning Health Systems Panel. Joint Commission Journal on Quality and Patient Safety. 2019;45(11):772-8. doi: 10.1016/j.jcjq.2019.08.002.

9. Paez K, Shapiro R, Thompson L, et al. Health System Panel To Inform and Encourage Use of Evidence Reports: Findings From the Implementation and Evaluation of Two Evidence-Based Tools. Methods Research Report. (Prepared by American Institutes for Research under Contract No. HHSP23320150014I/HHSP23337004T.) AHRQ Publication No. 22-EHC005. Rockville, MD: Agency for Healthcare Research and Quality. .

10. Haynes B. Can it work? Does it work? Is it worth it? The testing of healthcareinterventions is evolving. BMJ. 1999 Sep 11;319(7211):652-3. doi: 10.1136/bmj.319.7211.652. PMID: 10480802.

11. Zwarenstein M, Thorpe K, Treweek S, et al. PRECIS-2 for retrospective assessment of RCTs in systematic reviews. Journal of clinical epidemiology. 2020;126. doi: 10.1016/j.jclinepi.2020.06.023. PMID: 32565215.

12. Gartlehner G, Hansen RA, Nissman D, et al. Criteria for Distinguishing Effectiveness From Efficacy Trials in Systematic Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US) CTI - AHRQ Technical Reviews; 2006.

13. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355:i4919. doi: 10.1136/bmj.i4919. PMID: 27733354.

14. Hernán, M. A., Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25. doi: 10.1097/01.ede.0000135174.63482.43. PMID: 15308962.

15. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention

studies. Health Technol Assess. 2003;7(27):iii-x, 1-173. doi: 10.3310/hta7270. PMID: 14499048.

16. Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series-paper 5: a checklist for classifying studies evaluating the effects on health interventions-a taxonomy without labels. J Clin Epidemiol. 2017 Sep;89:30-42. doi: 10.1016/j.jclinepi.2017.02.016. PMID: 28351692.

17. Hartling L, Bond K, Santaguida PL, et al. Testing a tool for the classification of study designs in systematic reviews of interventions and exposures showed moderate reliability and low accuracy. J Clin Epidemiol. 2011 Aug;64(8):861-71. doi: 10.1016/j.jclinepi.2011.01.010. PMID: 21531537.

18. Saldanha IJ, Cao W, Broyles JM, et al. AHRQ Comparative Effectiveness Reviews. Breast Reconstruction After Mastectomy: A Systematic Review and Meta-Analysis. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021.

19. Dávila-Fajardo CL, Sánchez-Ramos J, Villamarín XD, et al. The study protocol for a non-randomized controlled clinical trial using a genotype-guided strategy in a dataset of patients who undergone percutaneous coronary intervention with stent. Data Brief. 2017 Feb;10:518-24. doi: 10.1016/j.dib.2016.12.019. PMID: 28066799.

20. Guyatt G, Rennie D, Meade M, et al. Users' guides to the medical literature : a manual for evidence-based clinical practice. 3rd ed. New York: McGraw-Hill Education Medical; 2015.

21. Giovannucci E, Ascherio A, Rimm EB, et al. A prospective cohort study of vasectomy and prostate cancer in US men. JAMA. 1993 Feb 17;269(7):873-7. PMID: 8426446.

22. Go AS, Singer DE, Toh S, et al. Outcomes of Dabigatran and Warfarin for Atrial Fibrillation in Contemporary Practice: A Retrospective Cohort Study. Ann Intern Med. 2017 Dec 19;167(12):845-54. doi: 10.7326/M16-1157. PMID: 29132153.

23. Sedgwick P. Bias in observational study designs: case-control studies. BMJ. 2015 Jan 30;350:h560. doi: 10.1136/bmj.h560. PMID: 25636996.

24. Wiese AD, Griffin MR, Schaffner W, et al. Opioid Analgesic Use and Risk for Invasive Pneumococcal Diseases: A Nested Case-Control Study. Ann Intern Med. 2018 03 20;168(6):396-404. doi: 10.7326/M17-1907. PMID: 29435555.

25. Vergis EN, Brennen C, Wagener M, et al. Pneumonia in long-term care: a prospective case-control study of risk factors and impact on survival. Arch Intern Med. 2001 Oct 22;161(19):2378-81. doi: 10.1001/archinte.161.19.2378. PMID: 11606155.

26. Torgerson DJ, Torgerson CJ. The Limitations of Before and After Designs. Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction. London: Palgrave Macmillan; 2008:9-16.

27. Reignier J, Dimet J, Martin-Lefevre L, et al. Before-after study of a standardized ICU protocol for early enteral feeding in patients turned in the prone position. Clin Nutr. 2009 Aug;29(2):210-6. doi: 10.1016/j.clnu.2009.08.004. PMID: 19709786.

28. Austin PC, Mamdani MM, Tu K, et al. Prescriptions for estrogen replacement therapy in Ontario before and after publication of the Women's Health Initiative Study. JAMA. 2003 Jun 25;289(24):3241-2. doi: 10.1001/jama.289.24.3241. PMID: 12824204.

29. Ramsay CR, Matowe L, Grilli R, et al. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care. 2003;19(4):613-23. doi: 10.1017/s0266462303000576. PMID: 15095767.

30. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. BMC Med Res Methodol. 2019 07 04;19(1):137. doi:

10.1186/s12874-019-0777-x. PMID: 31272382.

31. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr. 2013 2013 Nov-Dec;13(6 Suppl):S38-44. doi: 10.1016/j.acap.2013.08.002. PMID: 24268083.

32. Lawes T, Lopez-Lozano JM, Nebot CA, et al. Effects of national antibiotic stewardship and infection control strategies on hospital-associated and community-associated meticillin-resistant Staphylococcus aureus infections across a region of Scotland: a non-linear time-series study. Lancet Infect Dis. 2015 Dec;15(12):1438-49. doi: 10.1016/S1473-3099(15)00315-1. PMID: 26411518.

33. Milder EA, Rizzi MD, Morales KH, et al. Impact of a new practice guideline on antibiotic use with pediatric tonsillectomy. JAMA Otolaryngol Head Neck Surg. 2015 May 01;141(5):410-6. doi: 10.1001/jamaoto.2015.95. PMID: 25719954.

34. Feldstein AC, Smith DH, Perrin N, et al. Reducing warfarin medication interactions: an interrupted time series evaluation. Arch Intern Med. 2006 May 08;166(9):1009-15. doi: 10.1001/archinte.166.9.1009. PMID: 16682575.

35. Buechner A, Lesinski-Schiedat A, Becker P, et al. Real-world clinical experience with bimodal neuromodulation for the treatment of tinnitus - A case series. Brain Stimul. 2022 Feb 2. doi: 10.1016/j.brs.2022.01.022. PMID: 35123145.

36. Silva LC, Faustino ISP, Cantadori GR, et al. Adenocarcinoma not otherwise specified (NOS) arising in the sublingual gland: Rare case report and follow-up. Oral Oncol. 2022 Feb 2;126:105754. doi: 10.1016/j.oraloncology.2022.105754. PMID: 35123257.

37. Viswanathan M, Patnode CD, Berkman ND, et al. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. J Clin Epidemiol. 2018 May;97:26-34. doi: 10.1016/j.jclinepi.2017.12.004. PMID: 29248724.

38. Treadwell JR, Singh S, Talati R, et al. A Framework for “Best Evidence” Approaches in Systematic Reviews. Methods Research Report. (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. HHSA 290-2007-10063-I.) AHRQ Publication No. 11-EHC046-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm. doi: NBK56653. PMID: 21834173.

39. Chou R, Baker WL, Bañez L, et al. Prioritization and Selection of Harms for Inclusion in Systematic Reviews. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Scientific Resource Center under Contract No. 290-2012-0004-C). AHRQ Publication No. AHRQ Pub No. 17(18)-EHC-034-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2018. www.effectivehealthcare.ahrq.gov/reports/final.cfm DOI: https://doi.org/10.23970/AHRQEPCMETHGUIDE1. February 2018.

40. Guirguis-Blake JM, Evans CV, Perdue LA, et al. Aspirin Use to Prevent Cardiovascular Disease and Colorectal Cancer: An Evidence Update for the U.S. Preventive Services Task Force. (Prepared by Kaiser Permanente Evidence-based Practice Center under Contract No. HSA-290-2015-00007-I). AHRQ Publication No. 21-05283-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; September 2021.

41. Chou R, Fu R, Dana T, et al. Interventional Treatments for Acute and Chronic Pain: Systematic Review. Comparative Effectiveness Review No. 247. (Prepared by the Pacific Northwest Evidence-based Practice Center under Contract No. 75Q80120D00006.) AHRQ Publication No. 21-EHC030. Rockville, MD: Agency for Healthcare Research and Quality; September 2021. DOI: https://doi.org/10.23970/AHRQEPCCER247. PMID: 34524764.

42. Higgins JPT, Savović J, Page MJ, et al. Assessing risk of bias in a randomized trial. In Cochrane Handbook for

Systematic Reviews of Interventions. Cochrane. 2021. Available from: www.training.cochrane.org/handbook.

43. Viswanathan M, Patnode CD, Berkman ND, et al. Assessing the Risk of Bias in Systematic Reviews of Health Care Interventions. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Scientific Resource Center under Contract No. 290-2012-0004-C). AHRQ Publication No. 17(18)-EHC036-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2017. doi: https://doi.org/10.23970/AHRQEPCMETHGUIDE2. PMID: 30125066.

44. Doll R. Doing more good than harm: The evaluation of health care interventions: Summation of the conference. Annals of the New York Academy of Sciences. 1993;703(1):310-3.

45. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol. 1995 Jan;48(1):23-40. doi: 10.1016/0895-4356(94)00150-o. PMID: 7853045.

46. Polus S, Pieper D, Burns J, et al. Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in Cochrane reviews. Journal of Clinical Epidemiology. 2017;91:56-69.

47. Shadish W, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference: Houghton Mifflin Boston, MA; 2002.

48. Celentano DD, Szklo M. Gordis Epidemiology. 6th ed: Elsevier.

49. Viswanathan M, Middleton JC, Stuebe A, et al. AHRQ Comparative Effectiveness Reviews. Maternal, Fetal, and Child Outcomes of Mental Health Treatments in Women: A Systematic Review of Perinatal Pharmacologic Interventions. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021.

50. Viswanathan M, Treiman KA, Doto JK, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Folic Acid Supplementation: An Evidence Review for the US Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.

51. Kruse CS, Goswamy R, Raval Y, et al. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform. 2016 Nov 21;4(4):e38. doi: 10.2196/medinform.5359. PMID: 27872036.

52. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323-37. doi: 10.1016/j.jclinepi.2004.10.012. PMID: 15862718.

53. Verheij RA, Curcin V, Delaney BC, et al. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018 05 29;20(5):e185. doi: 10.2196/jmir.9134. PMID: 29844010.

54. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol. 2021 10 27;21(1):234. doi: 10.1186/s12874-021-01416-5. PMID: 34706667.

55. Mack C, Su Z, Westreich D. Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide. Third ed. Rockville (MD): Agency for Healthcare Research and Quality; 2018.

56. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology. 2003 May;14(3):300-6. PMID: 12859030.

57. Gomes M, Latimer N, Soares M, et al. Target Trial Emulation for Transparent and Robust Estimation of Treatment Effects for Health Technology Assessment Using Real-World Data: Opportunities and Challenges. Pharmacoeconomics. 2022 Mar 25. doi: 10.1007/s40273-022-01141-x. PMID: 35332434.

58. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758-64.

doi: 10.1093/aje/kwv254. PMID: 26994063.

59. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008 Nov;19(6):766-79. doi: 10.1097/EDE.0b013e3181875e61. PMID: 18854702.

60. Hernán MA, Sauer BC, Hernández-Díaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016 11;79:70-5. doi: 10.1016/j.jclinepi.2016.04.014. PMID: 27237061.

61. Dickerman BA, García-Albéniz X, Logan RW, et al. Avoidable flaws in observational analyses: an application to statins and cancer. Nat Med. 2019 10;25(10):1601-6. doi: 10.1038/s41591-019-0597-x. PMID: 31591592.

62. Lodi S, Phillips A, Lundgren J, et al. Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples With Apples. Am J Epidemiol. 2019 08 01;188(8):1569-77. doi: 10.1093/aje/kwz100. PMID: 31063192.

63. García-Albéniz X, Hsu J, Bretthauer M, et al. Estimating the Effect of Preventive Services With Databases of Administrative Claims: Reasons to Be Concerned. Am J Epidemiol. 2019 10 01;188(10):1764-7. doi: 10.1093/aje/kwz049. PMID: 30869122.

64. Forbes SP, Dahabreh IJ. Benchmarking Observational Analyses Against Randomized Trials: a Review of Studies Assessing Propensity Score Methods. J Gen Intern Med. 2020 May;35(5):1396-404. doi: 10.1007/s11606-020-05713-5. PMID: 32193818.

65. Franklin JM, Glynn RJ, Martin D, et al. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. Clin Pharmacol Ther. 2019;105(4):867-77. doi: 10.1002/cpt.1351. PMID: 30636285.

66. Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation. 2021 03 09;143(10):1002-13. doi: 10.1161/CIRCULATIONAHA.120.051718. PMID: 33327727.

67. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther. 2017 Dec;102(6):924-33. doi: 10.1002/cpt.857. PMID: 28836267.

68. Beaulieu-Jones BK, Finlayson SG, Yuan W, et al. Examining the Use of Real-World Evidence in the Regulatory Process. Clin Pharmacol Ther. 2020 04;107(4):843-52. doi: 10.1002/cpt.1658. PMID: 31562770.

69. Visvanathan K, Levit LA, Raghavan D, et al. Untapped Potential of Observational Research to Inform Clinical Decision Making: American Society of Clinical Oncology Research Statement. J Clin Oncol. 2017 Jun 01;35(16):1845-54. doi: 10.1200/JCO.2017.72.6414. PMID: 28358653.

70. Labrecque JA, Swanson SA. Target trial emulation: teaching epidemiology and beyond. Eur J Epidemiol. 2017 06;32(6):473-5. doi: 10.1007/s10654-017-0293-4. PMID: 28770358.

71. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

72. Shadish WR, Steiner PM. A primer on propensity score analysis. Newborn and Infant Nursing Reviews. 2010;10(1):19-26. doi: https://doi.org/10.1053/j.nainr.2009.12.010.

73. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. Am J Epidemiol. 2011 Sep 1;174(5):613-20. doi: 10.1093/aje/kwr143. PMID: 21749976.

74. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Stat Methods Med Res. 2009 Feb;18(1):67-80. doi: 10.1177/0962280208092347. PMID: 18562398.

75. Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol. 2013 Aug;66(8 Suppl):S84-S90 e1. doi: 10.1016/j.jclinepi.2013.01.013. PMID: 23849158.

76. Tadrous M, Gagne JJ, Sturmer T, et al. Disease risk score as a confounder summary method: systematic review and recommendations. Pharmacoepidemiol Drug Saf. 2013 Feb;22(2):122-9. doi: 10.1002/pds.3377. PMID: 23172692.

77. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2001.

78. Berk RA. Statistical learning from a regression perspective. New York: Springer; 2008.

79. Imai K, Ratkovic M. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014;76(1):243-63.

80. Andrade C. Propensity Score Matching in Nonrandomized Studies: A Concept Simply Explained Using Antidepressant Treatment During Pregnancy as an Example. J Clin Psychiatry. 2017 Feb;78(2):e162-e5. doi: 10.4088/JCP.17f11446. PMID: 28234438.

81. Huybrechts KF, Palmsten K, Avorn J, et al. Antidepressant use in pregnancy and the risk of cardiac defects. N Engl J Med. 2014;370(25):2397-407. doi: 10.1056/NEJMoa1312828. PMID: 24941178.

82. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health. 1998;19:17-34. doi: 10.1146/annurev.publhealth.19.1.17. PMID: 9611610.

83. Lousdal ML. An introduction to instrumental variable assumptions, validation and estimation. Emerg Themes Epidemiol. 2018;15:1. doi: 10.1186/s12982-018-0069-7. PMID: 29387137.

84. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA. 1994 Sep 21;272(11):859-66. PMID: 8078163.

85. Smith LM, Kaufman JS, Strumpf EC, et al. Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study. CMAJ. 2015 Feb 03;187(2):E74-E81. doi: 10.1503/cmaj.140900. PMID: 25487660.

86. Weiner AB, Vo AX, Desai AS, et al. Changes in prostate-specific antigen at the time of prostate cancer diagnosis after Medicaid expansion in young men. Cancer. 2020 07 15;126(14):3229-36. doi: 10.1002/cncr.32930. PMID: 32343403.

87. Hausner E, Metzendorf MI, Richter B, et al. Study filters for non-randomized studies of interventions consistently lacked sensitivity upon external validation. BMC Med Res Methodol. 2018 Dec 18;18(1):171. doi: 10.1186/s12874-018-0625-4. PMID: 30563471.

88. Wallace BC, Trikalinos TA, Lau J, et al. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010 Jan 26;11:55. doi: 10.1186/1471-2105-11-55. PMID: 20102628.

89. Ma LL, Wang YY, Yang ZH, et al. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res. 2020 Feb 29;7(1):7. doi: 10.1186/s40779-020-00238-8. PMID: 32111253.

90. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed on February 6, 2022.

91. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med. 2017 Aug 15;167(4):268-74. doi: 10.7326/m16-2607. PMID: 28693043.

92. Borenstein M, Higgins JP. Meta-analysis and subgroups. Prev Sci. 2013 Apr;14(2):134-43. doi: 10.1007/s11121-013-0377-7. PMID: 23479191.

93. Valentine JC, Thompson SG. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. Res Synth Methods. 2013 Mar;4(1):26-35. doi: 10.1002/jrsm.1064. PMID: 26053537.

94. Hedges LV, Pigott TD. The power of statistical tests for moderators in meta-analysis. Psychol Methods. 2004 Dec;9(4):426-45. doi: 10.1037/1082-989X.9.4.426. PMID: 15598097.

95. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. . In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021) ed.: Cochrane; 2021.

96. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015. PMID: 21208779.

97. Berkman ND, Lohr KN, Ansari M, et al. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality; 2013.

98. Open Science Framework. Real World Evidence Registry (RWE). https://osf.io/registries/rwe/discover. Accessed on May 1, 2022.

99. Ebell MH, Siwek J, Weiss BD, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. Am Fam Physician. 2004 Feb 01;69(3):548-56. PMID: 14971837.

100. Baral SD, Wirtz A, Sifakis F, et al. The highest attainable standard of evidence (HASTE) for HIV/AIDS interventions: toward a public health approach to defining evidence. Public Health Rep. 2012 2012 Nov-Dec;127(6):572-84. doi: 10.1177/003335491212700607. PMID: 23115382.

101. Clark E, Burkett K, Stanko-Lopp D. Let Evidence Guide Every New Decision (LEGEND): an evidence evaluation system for point-of-care clinicians and guideline development teams. J Eval Clin Pract. 2009 Dec;15(6):1054-60. doi: 10.1111/j.1365-2753.2009.01314.x. PMID: 20367705.

102. Li L, Smith HE, Atun R, et al. Search strategies to identify observational studies in MEDLINE and Embase. Cochrane Database Syst Rev. 2019 Mar 12;3:MR000041. doi: 10.1002/14651858.MR000041.pub2. PMID: 30860595.

103. Waffenschmidt S, Navarro-Ruan T, Hobson N, et al. Development and validation of study filters for identifying controlled non-randomized studies in PubMed and Ovid MEDLINE. Res Synth Methods. 2020 Sep;11(5):617-26. doi: 10.1002/jrsm.1425. PMID: 32472632.

104. BMJ Best Practice. Study design search filters. https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/study-design-search-filters/. Accessed on February 6, 2022.

105. Golder S, Peryer G, Loke YK. Overview: comprehensive and carefully constructed strategies are required when conducting searches for adverse effects data. J Clin Epidemiol. 2019 09;113:36-43. doi: 10.1016/j.jclinepi.2019.05.019. PMID: 31150833.

# Appendix A. Hedges

The suggested hedges here are for MEDLINE®, but the cited papers and the InterTASC Information Specialists Sub-Group Search Filter Resource can be used to find other similar hedges and those for Embase® and other databases.

**Observational studies** (Li 2019[102])
MEDLINE (Ovid)
1. Epidemiologic Studies/
2. exp Case-Control Studies/
3. exp Cohort Studies/
4. Cross-Sectional Studies/
5. (epidemiologic adj (study or studies)).ab,ti.
6. case control.ab,ti.
7. (cohort adj (study or studies)).ab,ti.
8. cross sectional.ab,ti.
9. cohort analy$.ab,ti.
10. (follow up adj (study or studies)).ab,ti.
11. longitudinal.ab,ti.
12. retrospective$.ab,ti.
13. prospective$.ab,ti.
14. (observ$ adj3 (study or studies)).ab,ti.
15. adverse effect?.ab,ti.
16. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15
17. medline.ti.
18. embase.ti.
19. pubmed.ti.
20. (database? and searching).ti.
21. *MEDLINE/
22. *PubMed/
23. *Databases, Bibliographic/
24. 17 or 18 or 19 or 20 or 21 or 22 or 23
25. 16 and 24
26. ((identify$ or develop$ or design$ or test$ or assess$ or evaluat$ or robust$ or optim$ or effic$ or effect$ or sensitiv$ or simpl$ or specific$ or precis$) adj3 ("search strat$" or "search filter?")).ab,ti.
27. 16 and 26
28. 25 or 27

*Appendix A references can be found in the main reference list above.*

**Controlled nonrandomized studies** (Waffenschmdt 2020[103])

MEDLINE (PubMed)
cohort[all] OR (control[all] AND study[all]) OR (control[tw] AND group*[tw]) OR epidemiologic studies[mh] OR program[tw] OR clinical trial[pt] OR comparative stud*[all] OR evaluation studies[all] OR statistics as topic[mh] OR survey*[tw] OR follow-up*[all] OR time factors[all] OR ci[tw]) NOT ((animals[mh:noexp] NOT humans[mh:noexp]) OR comment[pt] OR editorial[pt] OR review[pt] OR meta analysis[pt] OR case report[tw] OR consensus[mh] OR guideline[pt] OR history[sh]

**Cohort, case-control, and case series** (BMJ[104])

MEDINE (OVID)

1. exp cohort studies/
2. cohort$.tw.
3. controlled clinical trial.pt.
4. epidemiologic methods/
5. limit 4 to yr=1966-1989
6. exp case-control studies/
7. (case$ and control$).tw.
8. (case$ and series).tw.
9. or/1-3,5-8

**Adverse events** (Golder 2019[105])

MEDLINE (OVID) Medical devices
complicat*.ti,ab. OR ae.fs. [adverse effects] OR safe*.ti,ab. OR exp postoperative complications/OR failure*.ti,ab. OR adverse.ti,ab. OR co.fs. [complications] OR failed.ti,ab. OR exp equipment failure/OR removal.ti,ab. OR equipment safety/OR problem*.ti,ab. OR side effect*.ti,ab. OR harmful.ti,ab. OR tolerated.ti,ab. OR loosen*.ti,ab. OR Intraoperative complications/OR migration.ti,ab. OR breakag*.ti,ab. OR discomfort.ti,ab. OR displacement.ti,ab. OR detrimental adj2 effect*.ti,ab. OR untoward effects.ti,ab.

MEDLINE (OVID) Surgical procedures
complication*.ti,ab. OR ae.fs. [adverse effects] OR safe*.ti,ab. OR co.fs. [complications] OR postoperative complications/

MEDLINE (OVID) Drug interventions
ae.fs. OR co.fs. OR de.fs. OR safe.ti,ab. OR safety.ti,ab. OR side-effect*.ti,ab. OR undesirable effect*.ti,ab. OR treatment emergent.ti,ab. OR tolerability.ti,ab. OR toxicity.ti,ab. OR adrs OR (adverse adj2 (effect OR effects OR reaction OR reactions OR event OR events OR outcome OR outcomes)).ti,ab.

*Appendix A references can be found in the main reference list above.*