

*Draft Methods Guide
for Comparative Effectiveness Reviews*

Updating Quantitative Synthesis

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

This information is distributed solely for the purposes of prepublication peer review. It has not been formally disseminated by the Agency for Healthcare Research and Quality. The findings are subject to change based on the literature identified in the interim and peer-review/public comments and should not be referenced as definitive. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services (AHRQ) determination or policy.

Contract No.

Prepared by:

Investigators:

**AHRQ Publication No. xx-EHCxxx
<Month Year>**

This report is based on research conducted by the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers' 2016 Methods Workgroup. The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

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

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 .

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

This research was funded through contracts from the Agency for Healthcare Research and Quality to the following Evidence-based Practice Centers:

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

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

Suggested citation [pending]:

Preface

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

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. We welcome 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, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Andrew Bindman, M.D.
Director
Agency for Healthcare Research and Quality

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

Stephanie Chang, M.D., M.P.H.
Director, EPC Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: [will be included in the final report]

Peer Reviewers

Prior to publication of the final evidence report, EPCs sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than \$10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified.

The list of Peer Reviewers follows: [will be included in the final report]

Contents

Preface	iii
Peer Reviewers	iv
Contents	v
Introduction	6
Methods	6
Chapter I: Decision to Combine Trials of Treatment Efficacy or Harm	9
Chapter II: Optimizing Use of Effect Size Data	14
Chapter III: Choice of Statistical Model for Combining Studies	25
Chapter IV: Quantifying, Testing and Exploring Statistical Heterogeneity	33
Chapter V: Network Meta-Analysis (Mixed Treatment Comparisons/Indirect Comparisons)	59
Chapter VI: Stability and Sensitivity Analyses in Evidence Synthesis	70
Future Research Suggestions	78
References	80

*signifies chapter group leadership

Introduction

Background

The purpose of this document is to consolidate and update quantitative synthesis guidance provided in three previous methods guides¹ (Fu, 2012)^{2,3} We address comparative effectiveness reviews (CERs), which are systematic reviews that summarize comparative effectiveness and harms of alternative clinical options, and aim to help clinicians, policy makers, and patients make informed treatment choices. We focus on interventional studies of efficacy and do not address diagnostic studies, individual patient level analysis, or observational studies, which are addressed elsewhere.⁴ We differentiate between what is new guidance in this document from what was contained in the previous three guides.

Quantitative synthesis, or meta-analysis, is often essential for CERs to provide scientifically rigorous summary information. Quantitative synthesis should be conducted in a transparent and consistent way with methodologies reported explicitly. This guide provides practical recommendations on conducting synthesis. The guide is not meant to be a textbook on meta-analysis nor is it a comprehensive review of methods, but rather we address situations and decisions that are commonly faced by Evidence-based Practice Centers (EPCs). The goal is not to state requirements but rather to describe choices as explicitly as possible, with an appropriate degree of confidence.

EPC investigators are encouraged to follow these recommendations but may choose to use alternative methods if deemed appropriate, and after discussion with their AHRQ project officer. If alternative methods are used, the investigators are required to provide rationales for their choice, and if appropriate, to state the strengths and limitations of the chosen method in order to promote consistency and transparency. In addition, as elaborated in later sections of this document, several steps in conducting a meta-analysis require subjective decisions such as the decision to combine studies or the decision to incorporate indirect evidence. For each subjective decision, investigators should fully explain how the decision was reached.

This guide addresses issues in the order that they are usually addressed in a synthesis, though we acknowledge that the process is not always linear. We first consider the decision of whether or not to combine studies quantitatively. The next section addresses how to extract and utilize data from individual studies to construct effect sizes, followed by a section on statistical model choice. The fourth section considers quantifying and exploring heterogeneity. The fifth section describes an indirect evidence technique that has not been included in previous guidance – network meta-analysis, also known as mixed treatment comparisons. In the final section, we address the special topic of conducting stability and sensitivity analyses.

Methods

This guide has taken form through efforts of a workgroup comprised of members from across the EPCs, as well as from the Scientific Resource Center (SRC) of the AHRQ Effective Healthcare Program. Through surveys and discussions between AHRQ, Directors of Evidence-based Practice Centers, the Scientific Resource Center, and the Methods Steering Committee, quantitative synthesis was identified as a high-priority methods workgroup topic and a need was identified to update the original guidance.^{1,3,5} Once confirmed as a Methods Workgroup,

the SRC emailed EPCs soliciting workgroup volunteers for those with quantitative methods expertise including statisticians, librarians, thought leaders, and methodologists. Charged by AHRQ to update current guidance, the workgroup consists of members from eight of thirteen EPCs, the SRC, and AHRQ, and commenced in the fall of 2015. We conducted regular workgroup teleconference calls over the course of 14 months to discuss project direction and scope, assign and coordinate tasks, collect and analyze data, and discuss and edit draft documents. After constructing a draft table of contents, we surveyed all EPCs to ensure no topics of interest were missing.

The initial teleconference meeting was used to outline the draft, discuss the timeline, and agree upon a method for reaching consensus. The larger workgroup was split into subgroups each taking responsibility for a different chapter and topic focus. The larger group participated in biweekly discussions via teleconference and email communication. Subgroups communicated separately (in addition to the larger meetings) to coordinate tasks, discuss the literature review results and draft their respective chapters. Later, chapter drafts were combined into a larger document for workgroup review and discussion on the bi-weekly calls.

Literature Search and Review

A medical research librarian searched the ARHQ SRC Methods Library, a bibliographic database curated by the SRC currently containing over 16,000 citations of methodological works for systematic reviews and comparative effectiveness reviews. Key words and descriptors used were empirical guidance and research articles, determined on calls and email correspondence with the research librarian and used to search the AHRQ SRC Methods Library. The date was limited to 2012 and after to capture new and current guidance or methods documents and anything prior to 2012 would have already been reflected in the original guidance on quantitative synthesis, limited since last guidance.

The search yield was 1,358 titles and abstracts which were reviewed by all workgroup members using ABSTRACTR software (available at <http://abstrackr.cebm.brown.edu>). Each subgroup preferred the responsibility of including articles relevant to their own groups, thus each person in the workgroup reviewed the abstracts looking specifically at relevance to their respective subgroups. Reviews were done by single review as the inclusion criteria were very generous and fluid, investigators would include anything that may be potentially relevant and not necessarily used. Full text articles were pulled for each group leaving the decision of inclusion/exclusion to the authors of the subgroups/chapters.

Consensus and Recommendations

Reaching consensus is of great importance for AHRQ methods guidance as before the workgroup product can be submitted for peer review and publication, Directors from each EPC must agree with the proposed recommendations. This workgroup recognized this importance and on the first call and agreed on a process for consensus and conflict resolution: state disagreements in the document; conflicts within the smaller groups will be taken back to the larger groups for discussion and resolution. If a resolution isn't met, the groups will present all different ways/methods, be explicit; all members are encouraged to be candid on calls and emails about any concerns, ensuring voices are heard.

Following extensive drafting, a report of the workgroup's key conclusions and recommendations was circulated for comment by EPC and AHRQ officers at a biannual EPC Director's meeting in October 2016. In addition a full draft was circulated to AHRQ and EPC Director's prior to submission for peer review, and the manuscript was made available for public review; all these comments have been considered by the team in the final preparation of this report.

Chapter I: Decision to Combine Trials of Treatment Efficacy or Harm

1.1 Goals of the Meta-Analysis

In addition to supporting the goal of the larger review in which it is embedded, the overarching goal of a quantitative synthesis is generally to provide the best estimate of the effect of an intervention. As part of that aspirational goal, results of a meta-analysis may inform a number of related questions, such as whether that best estimate represents something other than a null effect (is this intervention beneficial?), the range in which the true effect likely lies, whether it is appropriate to provide a single best estimate, and what study-level characteristics may influence the effect estimate. Before tackling these questions, it is necessary to answer a preliminary but fundamental question: Is it appropriate to pool the results of the identified studies?⁶

1.2 Clinical and Methodological Heterogeneity

Studies must be reasonably similar to be pooled in a meta-analysis.¹ Even when the study protocol identifies a coherent and fairly narrow body of literature, the actual included studies may represent a wide range of specific population, intervention, and study characteristics. Variations in these factors are referred to as clinical heterogeneity and methodological heterogeneity.^{7,8} A third form of heterogeneity, statistical heterogeneity, will be discussed later.

Clinical heterogeneity refers to characteristics related to the participants, interventions, outcomes, and study setting, while **methodological heterogeneity** refers to variations in study methods (e.g., study design and study conduct). Exploring these two types of heterogeneity will inform the decision to combine studies, both because they may reveal a non-cohesive body of evidence that should not be pooled all together, and because variations in these factors may be associated with variations in treatment effect.⁸

Deciding whether studies are “similar enough” to pool is inherently a matter of judgement; there is no universally accepted standard.⁶ Some have suggested that pooling may be acceptable when it is plausible that the underlying effects could be similar across subpopulations and variations in interventions and outcomes.⁹ Others suggest that it may be acceptable to combine interventions with likely similar mechanisms of action.⁶ Verbeek and colleagues suggest working through key sources of variability in sequence, beginning the clinical variables of intervention/exposure, control condition, and participants, before moving on to methodological areas of study design, outcome, and follow-up time. Where variability on some dimensions is very high, reviewers should consider whether there are coherent subgroups of trials that can be pooled.⁶ In this way, the assessment of heterogeneity may actually inform the goal/question of the quantitative synthesis iteratively, such that the goal is refined until a reasonably low level of heterogeneity is achieved (or there is a decision not to combine studies).

Again, what constitutes an acceptably low level of clinical and methodological heterogeneity is not objectively defined, and investigators should be transparent about this assessment.

Methodological heterogeneity presents some common challenges:

Sample size. Sometimes the body of evidence comprises one or two very large trials and many small trials. If the best evidence is derived from large trials, it may be appropriate to focus on those trials rather than combining the best data with inferior data, particularly when addressing rare events that small studies are underpowered to examine.^{10, 11} Relatedly, when the body of evidence is limited only to small studies, pooling should be undertaken with caution. Results from small trials are less likely to be reliable than results of large trials, even when the risk of bias is low.¹²

First, with small samples it is difficult to balance the proportion of patients in potentially important subgroups across treatment conditions, and a difference between conditions of just a very few persons falling into a subgroup can result in a large proportional difference between groups. Characteristics that are rare are particularly at risk of being unbalanced in studies with small samples. In such situations there is no way to know if study effects are due to the intervention or to differences in the study group samples. In addition, samples are generally drawn from a narrower geographic range in small studies, making replication in other samples more uncertain. Finally, although it is not always the case, large trials are more likely to involve a level of scrutiny and standardization to ensure low risk of bias than small studies. Therefore, when there are only a handful of studies, and many have small samples, pooled effects are less likely to reflect the true effects of the intervention. In this case, methods such as trial sequential analysis to estimate the required or optimal information size can help the reviewer determine whether the sample size is sufficient to conclude that results are likely to be stable and not due to random heterogeneity (i.e., a truly significant or truly null results; not a type I or type II error).^{13, 14} A similar case can be made for meta-analysis of rare events: a small difference in absolute numbers of events can result in large relative differences, usually with low precision (i.e., wide confidence intervals). This could result in misleading effect estimates if the analysis is limited to trials that are underpowered for the rare outcomes.¹¹ An option here would be to pool the studies and acknowledge imprecision or other limitations when rating the strength of evidence.

Level of randomization. Another question is whether it is acceptable to combine individually-randomized and cluster randomized trials. We believe this is generally acceptable, with appropriate adjustment for cluster randomization as needed.¹⁵ However, closer examination may show that the cluster randomized trials also tend to systematically differ on population or intervention characteristics from the individually-randomized trials. If so, subgroup analyses may be considered.

Outcomes. There are also a number of challenges commonly encountered related to outcomes:

1. **Outcome definition.** Studies may have a wide array of specific instruments and cut-points for a common outcome. For example, a review considering pooling a binary depression prevalence outcome may find specific measures that range from a formal depression diagnosis based on a clinical interview to scoring above a cut-point on a wide variety of

specific instruments. One guiding principle is to consider pooling only when it is plausible that the underlying relative effects are consistent across specific definitions of an outcome.

2. **Reported statistics.** There is also typically substantial variability in which statistics are reported (e.g., baseline and mean followup scores, change scores for each condition, between-group differences at followup, etc.). Methods to calculate or estimate missing statistics are available,⁵ however the investigators must ultimately weigh the tradeoff of risking less accurate results (due to assumptions required to estimate missing data) with the potential advantage of pooling a more complete set of studies.
3. **Sparsely reported outcomes.** Another difficult but common situation is when a relatively small subset of the included studies report an important review outcome. For example, a reviewer may have 30 trials of weight loss interventions, of which only 10 reported blood pressure, which is considered an important outcome for the review. This pattern of results may indicate reporting bias, that trials finding group differences in blood pressure were more likely to report blood pressure findings. On the other hand, perhaps most of the studies limited to patients with elevated cardiovascular disease (CVD) risk factors did report blood pressure. In this case the reviewer may decide to combine the studies reporting blood pressure that were conducted in high CVD risk populations. However, investigators should be clear about the subset of the target population the meta-analysis is applicable to. An examination of the clinical and methodological features of the subset of trials where blood pressure was reported is necessary to make an informed judgement about whether to conduct a meta-analysis or not.
4. **Rare outcomes.** As noted above, meta-analyses of rare binary outcomes are frequently underpowered, and tend to overestimate the true effect size, so pooling should be undertaken with caution.¹⁰ One example is all-cause mortality, which is frequently provided as part of the participant flow methods, but may not be a primary study outcome, may not have adjudication methods described, and typically occurs very rarely. Studies are often underpowered to detect differences in mortality if it is not a primary outcome.

1.3 Statistical Heterogeneity

Once clinical and methodologic heterogeneity have been deemed acceptable for pooling, reviewers should next consider statistical heterogeneity. This is accomplished by looking at the consistency and precision of results of the included studies, i.e., conducting a preliminary meta-analysis. The decision that this process informs is whether the results of the meta-analysis are valid and should be presented, or should instead be shown without the pooled results, either in a forest plot or a table. If statistical heterogeneity is very high, the investigator may question whether an “average” effect is really meaningful or useful. If there is a reasonably large number of trials the reviewer may shift to exploring effect modification with high heterogeneity, however this may not be possible if few trials are being pooled. While many would likely agree that pooling (or reporting pooled results) should be avoided when there are few studies and statistical heterogeneity is high, what constitutes “few” studies and “high” heterogeneity is a matter of judgement.

While there are a variety of methods for characterizing statistical heterogeneity, one common method is with the I^2 statistic. I^2 is the proportion of inter-study variance in the pooled trials that is due to heterogeneity in the studies, as opposed to random variation.¹⁶ The Cochrane

manual proposes ranges for interpreting I^2 , but notes that I^2 should be interpreted in light of factors such as the magnitude and direction of effects, so did not provide discrete (non-overlapping) categories for low, medium, and high heterogeneity.¹⁵ They propose that the statistical heterogeneity associated with I^2 values of 0-40% might not be important, 30-60% may represent moderate heterogeneity, 50-90% may represent substantial heterogeneity, and 75-100% is considerable heterogeneity. Other measures of statistical heterogeneity include Cochrane's Q and τ^2 , but these heterogeneity statistics do not have intrinsic standardized scales that allow any specific values to be characterized as "small", "medium" or "large" in any meaningful way.¹⁷

Although widely used in quantitative synthesis, the I^2 has come under criticism in recent years. One important challenge with I^2 is that it has low power to detect statistical heterogeneity when there are few studies to pool and high statistical heterogeneity.^{18, 19} Further, in random effects models (but not fixed effects models), calculations demonstrate that I^2 tends to underestimate true statistical heterogeneity when there are fewer than about 10 studies and the true I^2 value is 50% or more.²⁰ Complicating this, meta-analyses of continuous measures tend to have higher heterogeneity than those of binary outcomes, and I^2 tends to increase as the number of studies increases when analyzing continuous outcomes, but not binary outcomes.^{21, 22} This has prompted some authors to suggest that different standards may be considered for interpreting I^2 for meta-analyses of continuous and binary outcomes, but I^2 should only be considered reliable when there are a sufficient number of studies.²² Unfortunately there is not clear consensus regarding what constitutes a sufficient number of studies for a given amount of statistical heterogeneity, nor is it possible to be entirely prescriptive, given the limits of I^2 as a measure of heterogeneity

1.4 Other Factors to Consider

Small Studies Effect. Another factor to consider is whether there is a small studies effect; that is, are smaller studies more likely to show larger effects than larger studies? If so, the assumption is that the pooled results may overestimate the true effect size. Reviewers should examine small studies effects using standard statistical tests such as the Egger test.²³ If there appears to be a small studies effect, the reviewer may decide not to report pooled results since they could be misleading.

Consistency of Effects. The consistency of the effects is also important to consider when deciding whether meta-analysis results are valid and should be reported. A reviewer may decide that it is acceptable to combine 3 large similar studies with fairly consistent effect sizes, but not to combine 10 small studies with high heterogeneity in effects, especially if results are not consistently in the same direction. In other words, if results are scattered on both sides of the null, indicating that the intervention could be either beneficial or harmful, a pooled effect may not be valid; there may be factors that influence whether an intervention is helpful or harmful that are not captured in the analysis.

1.5 Conclusion

In the end, the decision to pool boils down to the question: will the results of a meta-analysis help you find an answer to a meaningful question? That is, will the meta-analysis provide something in addition to what can be understood from looking at the studies individually? There is broad guidance to inform investigators in making this decision, but ultimately the choice is subjective, and requires careful consideration of the body of literature identified. To provide a meaningful result, the trials must be similar enough in content, procedures, and implementation to represent a cohesive group that is relevant to real practice/decision-making.

Chapter II: Optimizing Use of Effect Size Data

2.1 Nuances of Binary Effect Sizes

Data Needed for Binary Effect Size Computation

Under ideal circumstances, minimal data necessary for the computation of effect sizes of binary data would be available in published trial documents or otherwise available from the original source(s). Specifically, risk difference (RD), relative risk (RR) and odds ratios (OR) can be computed when the number of events (technically the number of cases in whom there was an event) and sample sizes are known for treatment and control groups. A schematic of one common approach to assembling binary data from trials for effect size computation is presented in Table 2.1 and will facilitate easy conversion to analysis using commercially-available software such as Stata (College Station, TX) or Comprehensive Meta-Analysis (Englewood, NJ).

Table 2.1: Assembling binary data for effect size computation

	Treatment Events	Treatment n	Control Events	Control n
Study X	5	25	6	25
Study Y	23	194	21	189

In many instances, however, a single or subset of studies to be included in the meta-analysis only report a single measure of association (for example, an odds ratio), and the sample size and event count are not available. Hence, the effect size chosen for the meta-analysis can be dictated by whichever was used in studies wherein essential raw metrics are not reported and not available. It should be noted that trial data published using CONSORT guidance should include the number of events and sample sizes for treatment and control groups groups.²⁴ The flexibility of choosing the most appropriate effect size is important to the integrity and transparency of meta-analyses; hence, every effort should be made to obtain all data presented in Table 2.1. In the event that data are only available in an effect size from the original reports, it is important assemble both the mean effect sizes and the associated 95% confidence intervals.

Choosing Among Effect Size Options

There is one absolute measure and two relative measures that are commonly used in meta-analyses involving binary data. The absolute measure - RD - is a simple metric and therefore most easily understood by clinicians and other key stakeholders including patient and lay groups. The relative measures – RR and OR – are also used frequently.

Risk Difference

The RD is most easily understood by clinicians and patients alike, and therefore most useful to aid decision making. But the RD tends to be less consistent than relative measures of effect size (RR and OR) across studies. Hence the RD may be a preferred measure in meta-analyses whenever the proportions of events among control groups are relatively common and similar across studies. When events are rare and when event rates differ across studies, however, the RD is not the preferred effects size to be used in meta-analysis because combined estimates based on RD in such instances have conservative confidence intervals and low statistical power. The calculation of RD and other effect size metrics using binary data from clinical trials can be performed considering the following labeling (**Table 2.2**).

Table 2.2: Organizing binary data for effect size computation

	Events	No Events	N
Treatment	A	B	n_1
Control	C	D	n_2

Equation Set 2.1: Risk Difference

$$RD = \left(\frac{A}{n_1}\right) - \left(\frac{C}{n_2}\right)$$

$$V_{RD} = \frac{AB}{n_1^3} + \frac{CD}{n_2^3}$$

$$SE_{RD} = \sqrt{V_{RD}}$$

$$LL_{RD} = RD - 1.96 * SE_{RD}$$

$$UL_{RD} = RD + 1.96 * SE_{RD}$$

Where,

RD = risk difference

V_{RD} = variance of the risk difference

SE_{RD} = standard error of the risk difference

LL_{RD} = lower limit of the 95% confidence interval of the risk difference

UL_{RD} = upper limit of the 95% confidence interval of the risk difference

Risk Ratio

It is important to note that the RR and OR are effectively equivalent for event rates below 10%-15%; hence, in such cases the RR is chosen over the OR simply for interpretability and not substantive difference. A potential drawback to the use of RR over OR (or RD) is that the RR of an event is not the reciprocal of the RR for from the non-occurrence of that event (e.g. using

survival as the outcome instead of death). In contrast, switching between events and non-occurrence of events is reciprocal in the metric of OR and only entails a change in the sign of RD. Hence, if switching between death and survival from death, as an example, in central to the meta-analysis then the RR is likely not the binary effect size metric of choice unless all raw data are available and re-computation is accessible. Moreover, investigators should be particularly attentive to the definition of an outcome event when using a RR.

The calculation of RR using binary data from clinical trials can be performed considering the labeling listed in **Table 2.2**. Of particularly note, the metrics of dispersion related to the RR are first computed in a natural log metric and then converted to the metric of RR.

Equation Set 2.2: Risk Ratio

$$RR = \frac{A/n_1}{C/n_2}$$

$$\ln_{RR} = \ln(RR)$$

$$V_{\ln_{RR}} = \frac{1}{A} + \frac{1}{C} - \frac{1}{n_1} - \frac{1}{n_2}$$

$$SE_{\ln_{RR}} = \sqrt{V_{\ln_{RR}}}$$

$$LL_{\ln_{RR}} = \ln_{RR} - 1.96 * SE_{\ln_{RR}}$$

$$UL_{\ln_{RR}} = \ln_{RR} + 1.96 * SE_{\ln_{RR}}$$

$$RR = \exp(\ln_{RR})$$

$$LL \text{ of the } 95\%CI = \exp(LL_{\ln_{RR}})$$

$$UL \text{ of the } 95\%CI = \exp(UL_{\ln_{RR}})$$

Where,

RR = risk ratio

\ln_{RR} = natural log of the risk ratio

$V_{\ln_{RR}}$ = variance of the natural log of the risk ratio

$SE_{\ln_{RR}}$ = standard error of the natural log of the risk ratio

$LL_{\ln_{RR}}$ = lower limit of the 95% confidence interval of the natural log of the risk ratio

$UL_{\ln_{RR}}$ = upper limit of the 95% confidence interval of the natural log of the risk ratio

LL_{RR} = lower limit of the 95% confidence interval of the risk ratio

UL_{RR} = upper limit of the 95% confidence interval of the risk ratio

Therefore, while the definition of the outcome event needs to be consistent among the included studies when using any measure, the investigators should be particularly attentive to the definition of an outcome event when using a RR.

Odds Ratios

An alternative relative metric for use with binary data is the OR. The calculation of OR using binary data from clinical trials can be performed considering the labeling listed in **Table 2.2**.

Similar to the computation of RR, the metrics of dispersion related to the OR are first computed in a natural log metric and then converted to the metric of OR.

Equation Set 2.3: Odds Ratios

$$\begin{aligned} \text{OR} &= \frac{AD}{BC} \\ \ln_{\text{OR}} &= \ln(\text{OR}) \\ V_{\ln_{\text{OR}}} &= \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \\ \text{SE}_{\ln_{\text{OR}}} &= \sqrt{V_{\ln_{\text{OR}}}} \\ \text{LL}_{\ln_{\text{OR}}} &= \ln_{\text{OR}} - 1.96 * \text{SE}_{\ln_{\text{OR}}} \\ \text{UL}_{\ln_{\text{OR}}} &= \ln_{\text{OR}} + 1.96 * \text{SE}_{\ln_{\text{OR}}} \\ \text{OR} &= \exp(\ln_{\text{OR}}) \\ \text{LL of the 95\%CI} &= \exp(\text{LL}_{\ln_{\text{OR}}}) \\ \text{UL of the 95\%CI} &= \exp(\text{UL}_{\ln_{\text{OR}}}) \end{aligned}$$

Where,

OR = odds ratio

\ln_{OR} = natural log of the odds ratio

$V_{\ln_{\text{OR}}}$ = variance of the natural log of the odds ratio

$\text{SE}_{\ln_{\text{OR}}}$ = standard error of the natural log of the odds ratio

$\text{LL}_{\ln_{\text{OR}}}$ = lower limit of the 95% confidence interval of the natural log of the odds ratio

$\text{UL}_{\ln_{\text{OR}}}$ = upper limit of the 95% confidence interval of the natural log of the odds ratio

LL_{OR} = lower limit of the 95% confidence interval of the odds ratio

UL_{OR} = upper limit of the 95% confidence interval of the odds ratio

A variation on the calculation of OR is the Peto OR that is commonly referred to as the assumption free method of calculating OR. The two key differences between the standard OR and the Peto OR is that the latter takes into consideration the expected number of events in the treatment group and also incorporates a hypergeometric variance. Because of these difference, the Peto OR is preferred for binary studies with rare events, especially when less than 1%. But in contrast, the Peto OR is biased when treatment effects are large and in the instance of imbalanced treatment and control groups.²⁶

Equation Set 2.4: Peto Odds Ratios

$$\text{OR}_{\text{peto}} = \exp\left[\frac{A - E(A)}{v}\right]$$

where $E(A)$ is the expected number of events in the treatment group calculated as:

$$E(A) = \frac{n_1(A + C)}{N}$$

and v is hypergeometric variance, calculated as:

$$v = \{n_1 n_2 (A + C)(B + D)\} / \{N^2 (N - 1)\}$$

There is no perfect effect size of binary data to choose because each has benefits and disadvantages. Criteria used to compare and contrast these measures include consistency over a set of studies, statistical properties, and interpretability. Key benefits and disadvantages of each are presented in **Table 2.3**.

Table 2.3: Benefits and Disadvantages of Binary Data Effect Sizes²⁷		
<i>Effects Size</i>	<i>Benefits</i>	<i>Disadvantages</i>
Risk Difference	-may be more easily interpretable among lay audiences -absolute metric	- not consistent between studies with differing baseline risks. -precision can be misleading when events are rare -not commonly reported in individual trials.
Relative Risk	-easily interpretable -commonly reported in individual trials considered in meta-analyses -consistent even with differing baseline risks	-values of “death” and “survival” are not reciprocals of each other as would be intuitively expected
Odds Ratio	-consistent even with differing baseline risks -commonly reported in individual trials considered in meta-analyses	-not easily interpretable - can be misleading when interpreted like relative risks -widespread use in meta-analyses may be because of convenience and history as opposed to mathematical properties.

Time-to-Event and Count Outcomes

For time to event data, the effect size measure is hazard ratio (HR), and most commonly estimated from the Cox proportional hazards model. In the best case scenario, HR and associated 95% confidence intervals are available from all sources, the time horizon was similar across

studies, and there is evidence that the proportional hazards assumption was met in each study to be included in a meta-analysis. When these conditions are not met, however, a HR and associated dispersion can be extracted with difficulty and concern over reproducibility due to observer variation.²⁸

Incident rate ratio (IRR) is used for count data and can be estimated from a Poisson or negative binomial regression models. It is important to consider how IRR estimates were derived in individual studies particularly with respect to adjustments for zero-inflation and/or over-dispersion; these can be sources of between-study heterogeneity.

Special Topics

Control Rate Meta-regression for Lower Underlying Risk.

For studies with binary outcomes, the “control rate” refers to the proportion of subjects in the control group who experienced the event. The control rate is viewed as a surrogate for covariate differences between studies because it is influenced by illness severity, concomitant treatment, duration of follow-up, and other factors that differ across studies.^{29,30} Patients with higher underlying risk for poor outcomes may experience different benefits and/or harms from treatment than patients with lower underlying risk.³¹ Hence, the control rate can be used to test for interaction between underlying population risk and treatment benefit, particularly in the setting of significant heterogeneity (see Chapter IV) or otherwise known differences in control rate across studies. To examine for an interaction between underlying population risk and treatment benefit, we recommend the following approach:

1. Generate a scatter plot of treatment effect against control rate as a useful preliminary approach to visually assess whether there may be a relation between the two. The RD is more highly correlated with the control rate compared with the RR or OR. Similarly, the relationship between the treatment effect and control rate is inflated using the RD;³⁰ hence RR or OR should be used when examining for a treatment effect against the control rate in visual assessment and subsequent steps.
2. Generate a simple weighted regression of the effect size on the control rate. Simple weighted regressions tend to identify a significant relation between control rate and treatment effect twice as often compared with more suitable approaches (below).^{30,32} A negative finding based on a simple weighted regression (i.e. slope not significantly different from zero) would be most likely replicated by the more complicated methods, and a positive finding (i.e. slope significantly different from zero) would need to be verified by a more comprehensive method.
3. If there is a positive finding based on a simple weighted regression, consider using hierarchical meta-regression models³⁰ or Bayesian meta-regression³² models to validate and refine the presence of an interaction between underlying population risk and treatment benefit using formal control rate meta-regression. These approaches incorporate the covariate of control rate in explaining variance in the treatment effect

under the hypothesis that the control rate is a surrogate for covariate differences between the studies.³³

Zero Cell Counts. In a study with zero events in one arm, estimation of effect measures (RR and OR) or their standard errors needs the addition of a correction factor, most commonly, 0.5 added to all cells. This is often done in the background; hence, the default handling of zero cells by statistical packages should be investigated thoroughly prior to use with binary data (See **Box 2.1** for examples default zero-cell count handling).

Box 2.1: Examples of Background Default Zero Cell Count Handling

Equation messages from Comprehensive Meta-Analysis (version 3.3). “One or more cells was empty, so 0.5 was added to each cell for computation of Log odds ratio and its variance,” and “one or more cells was empty, so 0.5 was added to each cell for computation of the Risk difference variance but not for computation of the Risk difference itself.”

Equation notes from metan package in Stata: “Here the default is to add 0.5 to all cells of the 2×2 table for the study (except for the Peto method, which does not require a correction).”

It has been shown, however, that the Mantel-Haenszel method with the 0.5 correction does not perform as well as the uncorrected Mantel-Haenszel method or with alternative correction factors. Hence, we advise against the use of the Mantel-Haenszel method with the 0.5 correction.³⁴ The investigators could choose adding no correction factors or exploring alternative correction factors using sensitivity analyses. Moreover, the Peto method of OR calculation does not require correction for zero cells counts, and it is only the variance of RD that requires non-zero cells and not the RD itself. Hence, Peto OR and RD should be considered as reasonable alternatives to RR (with or without adding 0.5 to each cell) in the setting of zero cell counts.

When both study arms have zero events, the relative measures (OR and RR) are not defined which can be problematic. These studies are usually excluded from the analysis as they do not provide information on the direction and magnitude of the effect size. Others consider including studies without events in the analyses to be important and choose to include them using correction factors. The Peto method and the Mantel-Haenszel method effectively exclude these studies from the analysis by assigning them zero weight. On the other hand, when the investigators estimate a combined control event rate, the zero events studies should be included, and we recommend the random effects logistic model that directly models the binomial distribution. As an alternative, the RD can be computed if that is an acceptable metric because zero cell counts influence the computation of variance but not the effect size directly. See **Table 2.4** for implications of the various methods of handling zero cell counts. Finally, if studies are excluded because of zero events in both study arms, they should be qualitatively summarized in the narrative section of the meta-analysis.

Table 2.4: Interpretive implications for handling of zero cell counts

Study	Intervention		Control		RR*	RR (0.9 added)	OR*	Peto OR	RD
	Events	No Events	Events	No Events					
Study A	0	100	5	99	0.095	0.159	0.090	0.135	-0.048
Study B	1	98	7	100	0.154	0.154	0.146	0.229	-0.055
Study C	10	400	18	396	0.561	0.561	0.550	0.559	-0.019
Study D	0	54	1	58	0.364	0.516	0.358	0.147	-0.017
Study E	0	210	0	200	Excluded	Excluded	Excluded	Excluded	-0.000
Study F	2	82	4	79	0.494	0.494	0.482	0.497	-0.024

* = 0.5 added to event and no event cells in the case of zeros in one study arm

2.2 Continuous Outcomes

Assembling Data Needed for Effect Size Computation

Once one has determined that a meta-analysis of a continuous outcomes will be performed, one must assemble from each included trial what will be needed to compute a pooled estimate. Regardless of which method one uses, this will boil down to acquiring an estimate of the difference between the two groups being compared, and an estimate of the standard error around the difference.

Estimating the difference between the two groups can be done most easily if the study gives us directly what we need. This would most commonly be the mean difference; although both standardized mean difference and ratio of means could possibly be given by the study authors. Most often, though we are given the means for each group from which we can readily compute a mean difference or ratio of means; combined with other pieces of information (see below) we can also compute standardized mean difference.

Estimates of the standard error around the mean difference can come from many sources. These include direct reporting of standard error or confidence interval of mean differences or other estimates. More commonly, you will be given confidence intervals, standard deviations, p-values, z-statistics, and t-statistics for which it will be possible to compute the standard error of the estimate of mean difference. In the absence of any of these statistics, other methods are available to estimate standard error (see recommendations for handling missing data).

More details of what precisely is needed for computations will be given in each of the corresponding sections below:

Choosing Among Effect Size Options

Details on how to compute these estimates can be found in previous AHRQ guidance document on continuous methods.⁵

(Weighted) Mean Difference

The mean difference (formerly known as weighted mean difference, but the “weighted” is usually dropped since it contains a subtle false implication that a pooled standardized mean difference is not weighted) is the most common way of summarizing and pooling a continuous outcome in a meta-analysis. Pooled mean differences can be computed when every study in the analysis measured the outcome on the same scale or on scales that can be easily converted. For example total weight could be pooled using mean difference even if different studies reported weights in kilograms and pounds; however it would not be possible to pool quality of life measured in both Self Perceived quality of life scale (SPQL) and the 36-item Short Form Survey Instrument (SF-36), since these are not readily convertible to one format.

Computation of mean difference is quite straightforward and readily explained elsewhere.⁵ Most software programs will require the mean, standard deviation and sample size from each intervention group and for each study in the meta-analysis, although other pieces of information will make the computation possible.

Some studies may report values as change from baseline, or alternatively present both baseline and final values. In these cases it is possible to pool final values with change from baseline values, although if baseline values are unbalanced it may be better to perform ANCOVA analysis (see below).⁵

Standardized Mean Difference

Sometimes different studies will assess the same outcome using different scales or metrics that cannot be readily converted to a common measure. In such instances computing a standardized mean difference (SMD) for each study and then pooling these across all studies in the meta-analysis is the most common method of dealing with this situation. By essentially dividing the mean difference by a pooled estimate of the standard deviation, we theoretically put all scales in the same unit (that being standard deviation), and are then able to statistically combine all the studies.

There are several methods that have been used to compute SMDs. The most frequent ones encountered are Cohen’s d , Hedges’ g , and Glass’ Δ .

Cohen’s d

Cohen’s d is the simplest SMD computation; it is defined as the mean difference divided by the pooled standard deviation.⁵ It has been shown that this estimate is biased in estimating the true population SMD, and the bias decreases as the sample size increases (small sample bias).³⁵ For this reason Cohen’s d is not used as much as Hedges’ g .

Hedges' *g*

Hedges' *g* is a transformation of Cohen's *d* that attempts to adjust for the small sample bias inherent in the latter. The transformation involves multiplying Cohen's *d* by a function of the total sample size.⁵ This generally results in a slight decrease in value of Hedges' *g* compared to Cohen's *d*, but the reduction lessens as the total sample size increases. For very large sample sizes the two will be very similar. The resultant reduction in bias of Hedges' *g* has made it generally more preferred compared to Cohen's *d*.

Back Transformation of Pooled SMD

One of the biggest disadvantages of the standardized mean difference is its lack of transparency in terms of being easily understandable by clinicians. SMDs are in units of standard deviation which makes it difficult to interpret clinically. Guidelines do exist but these are often thought to be arbitrary and not applicable to all situations.³⁶

An alternative solution is to back transform the pooled SMD into the original scales used in the one of the analyses. In theory, by multiplying the SMD (and its upper and lower confidence bounds) by the standard deviation of the original scale, one can obtain a pooled estimate in that original scale. The difficulty is that the true standard deviation is unknown and must be estimated from available data. Suggested methods for this include using the SD from the largest study or using a pooled estimate of the SDs across studies.⁵

Ratio of Means

Ratio of Means (RoM) has been presented as an alternative to the SMD when outcomes are reported in different non-convertible scales. As the name implies the RoM divides the treatment mean by the control mean rather than taking the difference between the two. The ratio can be interpreted as the percentage change in the mean value of the treatment group relative to the control group. By meta-analyzing across studies we are making the assumption that the relative change will be homogeneous across all studies, regardless of which scale was used to measure it. Similar to the risk ratio and odds ratio, the RoM is pooled on the log scale; computational formulas are readily available.⁵

For the RoM to have any clinical meaning, it is required that in the scale being used, the values are always positive (or always negative) and that a value of "zero" truly means zero. For example if the outcome were patient temperature, RoM would be a poor choice since a temperature of 0 degrees does not truly represent what we would think of as zero. As a result the same temperatures measured in degrees Celsius or degrees Fahrenheit would have different ratios when converted.

There is currently research being done to compare clinical interpretability of RoM versus SMD.

Special Topics

Crossover Trials

A crossover trial is one where all patients receive, in sequence, both the treatment and control interventions. This results in the final data having the same group of patients represented

with both their outcome values while in the treatment and control groups. When computing the standard error of the mean difference of a crossover trial, one must consider the correlation between the two groups.⁵ For most variables, the correlation will be positive, resulting in a smaller standard error than would be seen with the same values in a parallel trial.

To compute the correct pooled standard error requires an estimate of the correlation between the two groups. Most studies do not give the correlation or enough information to compute it, and thus it often has to be estimated based on investigator knowledge or imputed.⁵

Cluster Randomized Trials

Cluster trials occur when patients are randomized to treatment and control in groups (or clusters) rather than individually. If the units/subjects within clusters are positively correlated (as they usually are), then there is a loss of precision compared to a standard (non-clustered) parallel design. The design effect (DE) of a cluster randomized trial reflects this increase. Reported results from cluster trials may not reflect the design effect, and thus it will need to be computed by the reviewer.

Computation of the design effect involves a quantity known as the intra-class correlation coefficient (ICC), which is defined as the proportion of the total variance (i.e. within cluster variance plus between cluster variance) that is due to between cluster variance.⁵ ICC's are often not reported by cluster trials and thus a value must be obtained from external literature or a plausible value must be assumed by the investigator.

Mean Difference and Baseline Imbalance

Baseline imbalance in trials occurs when an important variable shows clinically important differences (by chance) between the intervention and control groups. If one is given both baseline and follow up times, there are three possible ways to compute a mean difference between groups:

1. Use follow up data to compute mean difference.
2. Use change from baseline data to compute mean difference.
3. Use an ANCOVA model to compute a mean difference that adjusts for the effects of baseline imbalance.³⁷

As long as trials are balanced at baseline, all three methods will give similar unbiased estimates of mean difference.⁵ When baseline balance is present, it can be shown that using the ANCOVA will give the best estimate of the true mean difference; however the parameters required to perform this analysis (mean and standard deviations of baseline, follow-up and change from baseline values) are usually not provided by the study authors.³⁸ If it is not feasible to perform an ANCOVA analysis, the choice of whether to use follow up or change from baseline values depends on the amount of correlation between baseline and final values. If correlation is greater than 0.5, then change from baseline values will be less biased—otherwise the follow up values will have less bias. There is evidence that these correlations are more often greater than 0.5, so the change from baseline means will usually be preferred if estimates of correlation are totally unobtainable.³⁹ A recent study⁴⁰ showed that three methods were unbiased when there were both few trials and small sample sizes within the trials.

Chapter III: Choice of Statistical Model for Combining Studies

3.1 Introduction

Meta-analysis can be performed using either a fixed or a random effects model to provide a combined estimate of effect size. A fixed effects model assumes that there is one single treatment effect across studies, and any differences between observed effect sizes are due to sampling error. Under a random effects model, the treatment effects across studies are assumed to vary from study to study and follow a random distribution. The differences between observed effect sizes are not only due to sampling error, but also to variation in true treatment effects. A random effects model usually assumes that the treatment effects across studies follow a normal distribution, though the validity of this assumption may be difficult to verify, especially when the number of studies is small. Alternative distributions⁴² or distribution free models^{43, 44} have also been proposed.

Recent advances in meta-analysis include the development of alternative models to fixed or random effects model. For example, Doi et al. proposed an inverse variance heterogeneity model (the IVhet model) for the meta-analysis of heterogeneous clinical trials that uses an estimator under the fixed effect model assumption with a quasi-likelihood based variance structure.⁴⁵ Stanley and Doucouliagos (2015) proposed an unrestricted weighted least squares estimator for meta-analysis and claimed superiority to both conventional fixed and random effects.⁴⁶ However, these methods have not been compared to the many estimators developed within the framework of the fixed and random effects model and are not readily available in most statistical packages; thus will not be further considered in the current guidance.

3.2 General Considerations for Model Choice

Considerations for model choice include many factors including but not limited to heterogeneity across treatment effects, the number and size of included studies, the type of outcomes, and potential bias. Generally, a fixed effects model is not advised in the presence of significant heterogeneity. We recommend against choosing a statistical model based on the significance level of heterogeneity test, for example, picking a fixed effect model when the *p*-value for heterogeneity is more than 0.10 and a random effects model when $P < 0.10$.

In practice, clinical and methodological diversity are always present across a set of included studies. Variation among studies is inevitable whether or not the test of heterogeneity detects it. Therefore, we recommend random effects models, with exceptions for rare binary outcomes (discussed in more details under combining rare binary outcomes). The considerations for the choice of random effects models and alternative estimators for the effect measures will be discussed in the next section. When the estimate of between-study heterogeneity is zero, for common binary outcomes, a fixed effects model (e.g., the Mantel-Haenszel method, inverse variance method, Peto method (for OR), or fixed effects logistic regression) could be used and provide similar estimates to the random effects model. Peto method requires that no substantial imbalance exists between treatment and control group sizes within trials and treatment effects are

not exceptionally large.

When a system review include both small and large studies, and the results of small studies are systematically different from those of the large ones, publication bias may be present and the assumption of a random distribution assumption is not justified. Other potential reasons that may lead to this systematic difference should also be examined. In this case, neither the random effects model nor the fixed effects model would provide an appropriate estimate and reviewers may choose not combining all the studies.¹⁵ Investigators can choose to combine the large studies if they are well conducted with good quality and expected to provide unbiased effect estimates.

3.3 Choice of Random Effects Model and Estimate

The most commonly used random effects model is based on an estimator developed by DerSimonian and Laird (DL) due to its simplicity and ease of implementation.⁴⁷ It is well recognized that the estimator does not adequately reflect the error associated with parameter estimation, in particular, when the number of studies is small, and between-study heterogeneity is high.⁴⁰ Refined estimators have been proposed by the original authors.⁴⁸⁻⁵⁰ Other estimators have also been proposed to improve the DL estimator. Sidik and Jonkman (SJ) and Hartung and Knapp(HK) independently proposed a non-iterative variant of the DL estimator using t-distribution and an adjusted confidence interval for the overall effect.^{51,52, 53} Biggerstaff–Tweedie (BT) proposed another variant of the DL method by building error in the point estimate of between study heterogeneity into the estimation of the overall effect.⁵⁴ There are also many other likelihood based estimators such as maximum likelihood estimate, restricted maximum likelihood estimate and profile likelihood (PL) methods, which account better for the uncertainty in the estimate of between-study variance.⁴⁸

Several simulation studies have been conducted to compare the performance of the different estimators,^{48, 55-59} and most of these studies compared a few selected methods. For example, Brockwell et al. (2001) showed that the PL method provides an estimate with better coverage probability than the DL method. Jackson et al. (2010) showed similar results that when the number of studies is small, the DL method does not provide adequate coverage probability, in particular, when there is moderate to large heterogeneity.⁵⁵ However, their results supported the usefulness of the DL method for larger samples. In contrast, the PL estimates result in coverage probability closer to nominal values. IntHout et al. (2014) compared the performance of the DL and HKSJ methods and showed that the HKSJ method consistently results in more adequate error rates than the DL method, especially when the number of studies is small, though they did not evaluate coverage probability and power.⁵⁹ Nevertheless, Kontopantelis and Reeves (2012a and 2012b)^{56, 58} conducted the most comprehensive simulation studies to compare the performance of nine different methods and evaluated multiple performance measures including coverage probability, power and overall effect estimation (accuracy of point estimates and error intervals). When the goal is to obtain an accurate estimate of overall effect size and the associated error interval, and by balancing the multiple performance measures, they recommend using the DL method when the heterogeneity is low and using the PL method when the heterogeneity is high, where the threshold of heterogeneity varies by the number of studies. PL method overestimates coverage probability in the absence of between-study heterogeneity. Methods like BT and SJ, despite being developed to address the limitation of the DL method,

were noted to be frequently outperformed by the DL method when considering the multiple measures. Encouragingly, Kontopantelis and Reeves also showed that regardless of the estimation method, results are highly robust against even very severe violations of the assumption of normally distributed effect sizes.

Recently there has been a call to use alternative random-effects estimators to replace the universal use of the DerSimonian-Laird random effects model.⁶⁰ Based on the results from the simulation studies, the PL method appears to generally perform best, and provides best performance across more scenarios than other methods, though it may overestimate the confidence intervals in small studies with low heterogeneity. It is also appropriate if the EPC investigators choose to use the DL method when the heterogeneity is low and use the PL method when the heterogeneity is high and determines the threshold of heterogeneity based on the number of studies.⁵⁶ The disadvantage of the PL method is that it does not always converge and produce valid estimate. In those situations, investigators may choose the DL method with sensitivity analyses using other methods, such as the HKSJ method. If the non-convergence is due to a lot of heterogeneity, the investigators should also reevaluate the appropriateness of combining the studies. The PL method (and the DL method) could be used to combine measures for continuous, count and time to event data, as well as binary data when the events are common. Also note that the confidence interval produced by the PL method may not be symmetric. For OR, RR, HR and incidence rate ratio, they should be analyzed on the logarithmic scale. In addition, we also support the use of the full Bayesian method to combine estimates since it takes the variations in all parameters into account (see more in the section of Bayesian methods).

Role of Generalized Linear Mixed Effects models

The different methods and estimators discussed above are generally used to combine effect measures directly (for example, mean difference, SMD, OR, RR, HR and incidence rate ratio). For study-level aggregated binary data and count data, we also support the use of generalized linear mixed effects model assuming random treatment effects. For aggregated binary data, a combined OR could be generated by assuming the binomial distribution with a logit link. It is also possible to generate a combined RR with the binomial distribution and a log link, though the model does not always converge and produce a valid combined estimate. For aggregated count data, a combined rate ratio could be generated by assuming the Poisson distribution with a log link. The advantage of such models is to model the exact likelihood of aggregated binary data and count data. Results from using the generalized linear models and directly combining effect measures are similar when the number of studies is large and/or the sample sizes are large.

3.4 A Special Case: Combining Rare Binary Outcomes

When comparing rare binary outcomes (such as adverse event data), few or zero events often occur in one or both arms in some of the included studies. The normal approximation of the binomial distribution does not hold well and choice of model becomes complicated. The DerSimonian-Laird (DL) method does not perform well with low-event rate binary data.^{61,62} A fixed effects model often outperforms the DL method for rare events based on simulation study,

even under the conditions of heterogeneity,³⁴ Within the past few years, many methods have been proposed to analyze sparse data from simple average,⁶³ exact methods,^{73, 74} Bayesian approach^{64, 65} to various parametric models (e.g. generalized linear mixed effect models, beta-binomial model, Gamma-Poisson model, bivariate Binomial-Normal model etc). Two dominating opinions are to 1) move away from the use of continuity corrections, and 2) include studies with zero events in both arms in the meta-analysis. Great efforts have been made to the development of methods that could include studies with zero events in both arms in the meta-analysis.

In earlier simulation studies, when event rates are less than 1 percent, the Peto OR method has been show to provide the least biased, most powerful combined estimates with the best confidence interval coverage,⁶¹ if the included studies have moderate effect sizes and the treatment and control group are of relatively similar sizes. The Peto method does not perform well when either the studies are unbalanced or the studies have large ORs (outside the range of 0.2-5).^{66, 67} Otherwise, when treatment and control group sizes are very different or effect sizes are large, or when events become more frequent (5 percent to 10 percent), the Mantel-Haenszel method (without correction factor) or a fixed effects logistic regression provide better combined estimates.

Dealing with studies with zero events in one or both arms

Bhaumik et al. (2012) proposed the simple (unweighted) average (SA) treatment affect with the 0.5 continuity correction, and found out that the bias of the SA estimate in the presence of even significant heterogeneity is minimal compared to MH estimates (with 0.5 correction), and a simple average was also advocated by Shuster (2010).^{63, 68} However, the issue of confounding always remains for an unweighted estimate. Spittal (2015) showed that Poisson regression works better than the inverse variance method for rare events but the inverse variance method is generally not preferred anyway.⁶⁹ Kuss et al. (2015) conducted a comprehensive simulation of eleven methods that could combine rare binary events including most recent developed methods, and recommend the use of beta-binomial model for the three common estimators (OR, RR and RD) as the preferred meta-analysis methods for rare binary events with studies of zero events in one or both arms.⁷⁰ They examined methods that could incorporate data from studies with zero events from both arms, and do not need any continuity correction, and only compared the Peto and MH methods as reference methods. Ma et al. (2016) also showed that the binomial-beta approach tends to have substantially smaller bias and mean squared error than Binomial-Normal approach for rare events < 5%.⁷¹

Given the development of methods that could handle studies with zero events in both arms, we agree to avoid methods that use continuity corrections should be avoided. Investigators should use valid methods that include studies with zero events in one or both arms. For studies with zero events in one arm, or studies with sparse binary data but no zero events, an estimate can be obtained using the Peto method, the Mantel-Haenszel method, or a logistic regression approach, without adding a correction factor, when the between study heterogeneity is small. These methods are simple to use and more readily available in standard statistical package. When the between study heterogeneity is large, and there are studies with zero events in both arms, the more recently developed methods, such as beta-binomial model could be explored and used. However, investigators should note that no method gives completely unbiased estimates

when events are rare. The issue of sparse data could never be completely solved by statistical methods; investigators should always conduct sensitivity analysis using alternative methods to check the robustness of results to different methods, and acknowledge the inadequacy of data sources when presenting the data synthesis results, in particular, when the proportion of studies with zero events in both arms are high.

A risk-difference (RD) may be favored because it includes zero-event studies and is easily interpretable.⁷² The RD can also be used to calculate the Number Needed to Harm (NNH).⁷³ It is not preferred when there is heterogeneity between studies in duration and incident rates. The RD has been shown to lack power.⁶¹

If double-zero studies are to be excluded, they should be qualitatively summarized, by providing information on the confidence intervals for the proportion of events in each arm.

A note on exact method for sparse binary data

For rare binary events, the normal approximation and asymptotic theory for large sample size don't work satisfactorily and exact inference has been developed to overcome these limitations. Also exact methods don't need continuity corrections. However, simulation analyses did not identify a clear advantage of early developed exact methods^{74, 75} over a logistic regression or the Mantel-Haenszel method even in situations where these exact methods would theoretically be advantageous.⁶¹ Recent developments of exact method includes Tian et al.'s method of combining confidence intervals⁷⁶ and Liu et al.'s method of combining p-value functions.⁷⁷ Yang et al.⁷⁸ developed a general framework for meta-analysis of rare events by combining confidence distributions (CDs), and showed that Tian's and Liu's methods could be unified under the CD framework. Liu showed that exact methods performed better than the Peto method (except when studies are unbalanced) and the Mantel-Haenszel method,⁷⁷ though the comparative performance of these methods has not been thoroughly evaluated. Therefore the investigators may choose to use exact methods with considerations for the interpretation of effect measures but we don't specifically recommend exact methods over other models discussed above.

3.5 Bayesian Methods

The Bayesian framework can provide a unified and comprehensive approach to meta-analysis. This framework accommodates a wide variety of reported outcomes, and allows a common approach in which pairwise meta-analysis is a special case of network meta-analysis. Rather than thinking of meta-analysis as a collection of procedures, this approach emphasizes that meta-analysis is "just regression" where outcomes are nested within trials.⁷⁹ Generalized linear modeling (GLM) theory provides for normal, binomial, Poisson and multinomial likelihoods, with various link functions, providing common core models for the linear predictor. This leads to a modular approach: different likelihoods and link functions are employed, but the "synthesis" operation, which occurs at the level of the linear predictor, takes the exact same form in every case.⁸⁰ This flexibility has a number of advantages, detailed below:

1. It is not necessary to use approximate normal likelihoods. ‘Exact’ likelihoods (e.g. binomial) can be specified. Thus, the number of events and number of individuals in each study arm are specified, without the need for continuity correction.³⁴
2. Within this modular framework, generalization to meta-regression models is easily accomplished by adding study level covariates to the linear predictor. However, centering of covariates may be needed to facilitate estimation using MCMC sampling.⁸¹

It should be noted that these GLM models are routinely implemented in the frequentist framework, and not specific to the Bayesian framework. However, extensions to more complex challenges are most approachable using the Bayesian framework, for example, allowing for mixed treatment comparisons involving repeated measurements of a continuous outcome that varies over time.⁸²

There are several specific advantages inherent to the Bayesian framework:

The Bayesian posterior parameter distributions fully incorporate the uncertainty of all parameters. These posterior distributions need not be assumed to be normal.⁸³ In random-effects meta-analysis, standard methods use only the most likely value of the between-study variance,⁴⁷ rather than incorporating the full uncertainty of each parameter. Thus, Bayesian credible intervals will tend to be wider than confidence intervals those from a classical random-effects analysis.⁸⁴ However, when small numbers of studies are available, the between study variance will be poorly estimated by both traditional and Bayesian methods. Indeed, this is an example where the use of vague priors can be problematic, and can lead to a marked variation in results,⁸⁵ particularly when the model is used to predict the treatment effect in a future study.⁸⁶ If a meta-analysis is to be undertaken, plausible values for τ^2 may be preferable to values estimated from very few studies. A natural alternative is to use an informative prior distribution, based on observed heterogeneity variances in other, similar meta-analyses.⁸⁷⁻⁸⁹

Further, full posterior distributions provide a more informative summary of the likely value of parameters than do point estimates. Another advantage is that posterior distributions of functions of model parameters can be easily obtained. Thus, when communicating results of meta-analysis to clinicians, the Bayesian framework allows direct probability statements to be made such as the rank probability that a given treatment is best, second best, or worst. Distinct from the choice of scale used for modeling treatment effects (e.g. odds ratios/logit scale), given information on the absolute effect of one treatment, it is possible to derive treatment effects (with credible intervals) on other scales such as risk difference, relative risk, or number(s) needed to treat.⁸⁰ Finally, the Bayesian approach allows full incorporation of parameter uncertainty from meta-analysis into decision analyses.⁹⁰

Until recently, Bayesian meta-analysis required specialized software such as WinBUGS,⁹¹ OpenBUGS,⁹² and JAGS.^{93,94} Newer open source software platforms such as Stan⁹⁵ and Nimble^{96,97} provide additional functionality and use BUGS-like modeling languages.

For analysts working in general Statistics programs (e.g. Stata and R) there are user written commands that allow data processing in a familiar environment which then can be passed to WinBUGS, or JAGS for model fitting.⁹⁸ In R, the package *bmeta* currently generates JAGS code to implement 22 models.⁹⁹ The R package *gemtc* similarly automates generation of JAGS code for network meta-analysis models and facilitate assessment of model convergence and inconsistency.^{100,101}

On the other hand, Bayesian meta-analysis could be implemented in commonly used statistical packages. For example, SAS PROC MCMC can now estimate at least some Bayesian hierarchical models¹⁰² directly, as can Stata, version 14, via the *bayesmh* command.¹⁰³

Both fixed and random effects models have been developed within a Bayesian framework for various types of outcomes. The Bayesian fixed effects model provides good estimates when events are rare for binary data.³⁴ We support the use of Bayesian methods with vague priors in CERs, if the investigators choose Bayesian methods. When the prior distributions are vague, Bayesian estimates are usually similar to estimates using the above methods, though choice of vague priors could lead to a marked variation in the Bayesian estimate of between-study variance when the number of studies is small.⁸⁵ Use of informative prior is not prohibited but it takes careful considerations to avoid introducing biases into the posterior estimates. Investigations should provide adequate justifications for the choice of priors and conduct sensitivity analyses. The basic principle to guide the choice between a random effects and a fixed effect model is the same as that for the above non-Bayesian methods, though the Bayesian methods needs more work in programming, simulation and simulation diagnostic.

A Note on Bayesian Method for Sparse Binary Data

There was argument that Bayesian method might be a valuable alternative for sparse event data since Bayesian inference does not depend on asymptotic theory and take into account all uncertainty in the model parameters.¹⁰⁴ However, the choice of prior distribution, even non-informative ones, could have a big impact on the results, in particular, when a big proportion of studies have zero events in one or two arms.^{85, 105, 106} Nevertheless, other simulation studies found that when overall baseline rate is very small and there is moderate or large heterogeneity, the Bayesian hierarchical binomial - normal random-effect models can provide less biased estimates for the effect measures and the heterogeneity parameters.⁶⁵ To reduce the impact of the prior distributions, objective Bayesian methods have been developed^{107, 108} with special attention paid to the coherence between the prior distributions of the study model parameters and the meta-parameter,¹⁰⁸ though the Bayesian model is developed outside the usual hierarchical normal random-effect framework. Further evaluations of these methods are required before recommendations for using these methods could be made.

3.6 Multivariate Meta-Analysis

Medical studies often examine multiple, and correlated, outcomes of interest to the meta-analyst. For example, in clinical trials of treating hyperlipidemia, the treatment effects on HDL, LDL and triglyceride are often evaluated; or both systolic and diastolic blood pressures would be measured in a hypertension study. Another situation is longitudinal outcome that were measured at multiple time points. These outcomes may be combined using multivariate meta-analysis.

The multivariate meta-analysis is a generalization of the standard univariate meta-analysis model. There has been substantial advancement in multivariate meta-analysis methods in both frequentist and Bayesian framework in recent years,¹⁰⁹⁻¹¹⁶ and some of these methods are readily available in statistical packages (for example, Stata *mvmeta*). The multivariate meta-analysis allows estimation of multiple effects in a single modeling framework while taking into account the correlation among multiple outcomes with parameter estimates likely more precise.¹¹⁰

Further, the multivariate methods may have the potential to reduce the impact of outcome reporting bias.^{110, 117, 118} However, multivariate meta-analysis requires estimates of within-study correlations which are typically unknown. This remains the greatest difficulty of multivariate meta-analysis methods in practice and needs assumptions that may not always result in better inference. The gain in precision of parameter estimates is often marginal and the conclusions from the multivariate meta-analysis are often the same as those from the univariate meta-analysis, which may not justify the increased complexity and difficulty of utilizing multivariate meta-analysis.

With the exception of diagnostics testing studies (which provides a natural situation to meta-analyze sensitivity and specificity simultaneously, but out of scope for this guidance) and network meta-analysis (a special case of multivariate meta-analysis with its unique challenges, see Chapter V on network meta-analysis), multivariate meta-analysis has not been widely used in practice. While multivariate meta-analysis holds its potentials, we currently don't recommend its routine use. However, investigators are encouraged to explore multivariate meta-analysis when it likely brings considerable advantages, for example, when there is a large amount of missing data and borrowing of strength from correlated outcomes could lead to significant gain in the precision of the combined estimates, or reduced impact from outcome reporting bias.

Chapter IV: Quantifying, Testing and Exploring Statistical Heterogeneity

4.1 Concepts of heterogeneity in meta-analysis

In this chapter, it is assumed that a well-specified research question has been posed, the relevant literature has been reviewed, a set of trials meeting selection criteria have been identified and data from them has been abstracted and verified. Even when the review selection criteria are aimed toward identifying studies that are adequately homogenous, however, it is common for trials included in meta-analyses to differ considerably as a function of clinical and/or methodological heterogeneity. Clinical heterogeneity^{8, 119} and methodological heterogeneity^{5, 10-12, 15} have been reviewed in Chapter I. However, even when these sources of heterogeneity have been accounted for, there remains statistical heterogeneity that is the consequence of the degree of inconsistency in intervention effects among studies. Since statistical heterogeneity must be expected, quantified and sufficiently addressed in meta-analyses;¹²⁰ therefore, we sought to lay the foundation for investigators to address the following questions:

- 1) Is there evidence of heterogeneity in effect sizes across studies?
- 2) What is the ratio of total and excess dispersion in effect sizes?
- 3) What is the among-study variance of the true effects?
- 4) What proportion of the observed heterogeneity is real vs. spurious?
- 5) What can be done in instances of significant heterogeneity?

4.2 Causes of heterogeneity

As described in Chapter 1, clinical heterogeneity refers to characteristics related to the participants, interventions, outcomes and study setting, while methodological heterogeneity refers to variations in study methodology (e.g., study design, level of randomization, outcomes definition, statistics reported and study conduct). Clinical and methodological heterogeneity should inform the decision to combine studies prior to formal meta-analysis.⁸ Once the decision to combine studies is deemed acceptable, our attention must turn toward quantifying, exploring and explaining statistical heterogeneity as a product of meta-analysis.

4.3 Quantifying heterogeneity

Quantitative methods can assist in many analytic goals related to heterogeneity: they can provide evidence of heterogeneity in effect sizes across studies, the ratio of total and excess dispersion in effect sizes, the among-study variance of the true effects, and proportion of the observed heterogeneity that is real vs. spurious.

Heterogeneity in effect sizes across studies

DerSimonian and Laird proposed an estimator of Q that can and should be computed as a metric of heterogeneity in effect sizes across studies.⁴⁷ Although Q formula notation can vary, the following provides an example formula that is useful for direct computation:

$$Q = \sum w_i (y_i - \hat{\mu}_F)^2$$

Where Q is the metric of heterogeneity in effect sizes across studies,
 w is the study weight based on inverse variance weighting,
 y is the observed effect size in each trial, and
 $\hat{\mu}_F$ is the summary estimate in a fixed-effect meta-analysis.

This heterogeneity estimator is very popular in applied research and has been implemented in practically all meta-analysis software as a default. Since Q is derived from fixed-effect meta-analysis there are underlying assumptions that studies share a common effect size, and that all variation between studies is due to sampling error. Hence, there is an expected degree of variation that is equivalent to the degrees of freedom, and all variation in effect sizes beyond what is expected can be considered excess dispersion. Therefore, a simple but informative metric of excess dispersion is:

$$Q - (k - 1)$$

Where Q is the metric of heterogeneity in effect sizes across studies, and
 $k - 1$ is the degrees of freedom.

Assuming that Q follows a χ^2 distribution with $k - 1$ degrees of freedom, it also can be used in a null hypothesis test of homogeneity vs. an alternative hypothesis of heterogeneity in intervention effects across studies. Hence, Q is typically presented along with a p-value. Interpretation of a Q statistic in isolation is not advisable however, because it has low statistical power in meta-analyses involving a limited number of studies^{121, 122} and may detect unimportant heterogeneity when the number of studies included in a meta-analysis is large. Most importantly, since heterogeneity is expected in meta-analyses irrespective of whether or not we have statistical tests to support that claim, non-significant Q statistics must not be interpreted as the absence of heterogeneity. Instead, the Q statistic in each instance must be interpreted along with other heterogeneity statistics and under full consideration of its strengths and limitations.

Among-study variance and standard deviation

DerSimonian and Laird also proposed a non-iterative method-of-moments estimator of between-study variance (τ^2)⁴⁷ that was later described more appropriately by Higgins et al.¹⁶ as ‘among-study variance:’

$$\hat{\tau}_{DL}^2 = \frac{Q - (k - 1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}$$

Where τ^2 is the estimate among-study variance of the true effects,
DL is the DerSimonian and Laird approach to τ^2 estimation,
Q is the heterogeneity in effect sizes across studies (as above),
k - 1 is the degrees of freedom, and
w is the weight applied to each study based on inverse variance weighting.

The interpretation of τ^2 is the among-study variance of the true effects. Since variance in true effects cannot be less than zero, estimates of τ^2 that are less than zero are set to zero. The value of τ^2 is integrated into the weights of random-effects meta-analysis as presented in Chapter III. An additional benefit of the τ^2 estimate is that it can help in the interpretation of the relevance of the variance in effects across studies because unlike the *Q* statistic, τ^2 remains in the metric of the effect size (e.g. RR, SMD et al.). Moreover, since the τ^2 estimate is a variance statistic, we can compute the standard deviation in effect sizes across studies by taking the square root of the τ^2 estimate (i.e. τ). Since our estimates of τ^2 and τ remain in the metric of the effect size, investigators can comment on the among-study variance of the true effects and the standard deviation of true effects observed across studies using similar language as can be used to discuss the overall pooled estimates. Using heterogeneity statistics that are in the metric of the effect size are particularly helpful in instances where minimally important/clinically meaningful differences are well established.

Inconsistency across studies

Another statistic that should be generated and interpreted even in cases where *Q* is not statistically significant is the proportion of variability in effect sizes across studies that is explained by heterogeneity vs. sampling error (i.e. beyond chance) or I^2 .^{16, 123}

$$I^2 = \frac{Q - (k - 1)}{Q} * 100$$

Where *Q* is the heterogeneity in effect sizes across studies, and
k - 1 is the degrees of freedom.

More than other heterogeneity statistics presented in this chapter, the I^2 is metric of how much heterogeneity is influencing the meta-analysis. With a range from 0% (indicating no heterogeneity) to 100% (indicating that all of the observed variance is real vs. spurious), the I^2 statistic has several advantages over other heterogeneity statistics including its relative simplicity, ease of interpretation across meta-analyses, lack of direct dependence on the number of studies included in the meta-analysis, and focus on how heterogeneity may be influencing interpretation of the meta-analysis.⁹ By various means, confidence/uncertainty intervals can be estimated for I^2 including the following common example based on Higgins' test-based method:^{12,17}

$$B = 0.5 \times \frac{\ln(Q) - \ln(df)}{\sqrt{2Q} - \sqrt{2 \times (df) - 1}} \text{ when } Q > (df + 1), \text{ or}$$

$$B = 0.5 \times \sqrt{\frac{1}{2 \times (df - 1) \times (1 - (\frac{1}{3 \times (df - 1)^2})^2)}} \text{ when } Q \leq (df + 1)$$

$$L = \exp\left(0.5 \times \ln\left(\frac{Q}{df}\right)\right) - 1.96 \times B$$

$$U = \exp\left(0.5 \times \ln\left(\frac{Q}{df}\right)\right) + 1.96 \times B$$

$$LL_{I^2} = \left(\frac{L^2 - 1}{L^2}\right) \times 100\%$$

$$UL_{I^2} = \left(\frac{U^2 - 1}{U^2}\right) \times 100\%$$

Where Q is the heterogeneity in effect sizes across studies, and df is the degrees of freedom.

Hence, another benefit of the I^2 over other heterogeneity statistics is the ability to provide confidence intervals as a means of being transparent about inherent uncertainty in the variability in effect sizes across studies that is explained by heterogeneity vs. sampling error.⁹ It is important to note that since the I^2 is based on Q , any problems that influence Q (most notably the number of trials included in the meta-analysis) will also indirectly interfere with the computation of I^2 . Although it is important to note that assumptions involved in the construction of 95% confidence intervals cannot be justified in all cases, I^2 confidence intervals based on frequentist assumptions generally provide sufficient coverage of uncertainty in meta-analyses.¹²³

Based primarily on the observed distributions of I^2 across meta-analyses, there are ranges that are commonly used to further categorize heterogeneity. That is, I^2 values of 25%, 50% and 75% have been proposed as working definitions of what could be considered low, moderate and high proportions, respectively, of variability in effect sizes across studies that is explained by heterogeneity vs. sampling error.⁹ Currently, the Cochrane manual also includes ranges for interpreting I^2 (0%-40% might not be important, 30%-60% may represent moderate heterogeneity, 50-90% may represent substantial heterogeneity and 75-100% may represent considerable heterogeneity).¹⁵ Irrespective of which categorization of I^2 is used, this statistic must be interpreted with the understanding of several nuances, including issues related to a small number of studies (i.e. fewer than 10),^{18, 19, 20} and inherent differences in I^2 comparing binary and continuous effect sizes.^{21, 22} Moreover, I^2 of zero is often misinterpreted in published reports as being synonymous with the absence of heterogeneity despite upper confidence interval limits that most often would exceed 33% when calculated.¹²⁴ Finally, a high I^2 does not necessarily mean that dispersion occurs across a wide range of effect sizes, and a low I^2 does not necessarily mean that dispersion occurs across a narrow range of effect sizes; the I^2 is as signal-to-noise metric not a statistic about the magnitude of heterogeneity.

4.4 Tests for the Null Hypothesis of Homogeneity

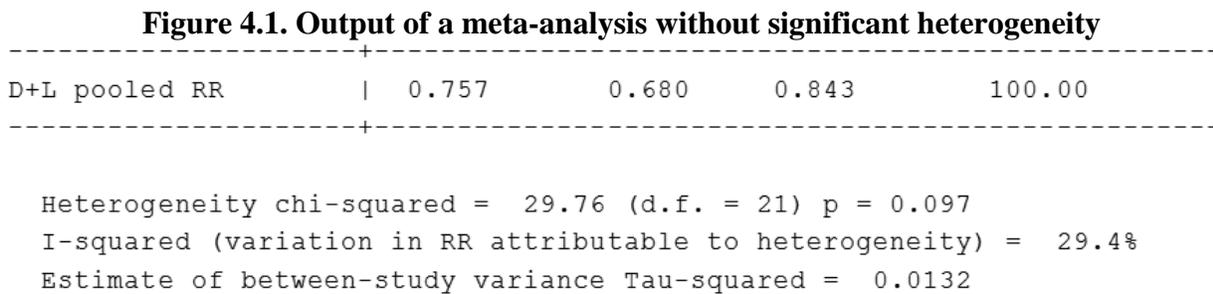
The most commonly assessed hypothesis for heterogeneity in meta-analysis is:

$$H_0 : \theta_i = \theta \text{ for all } i,$$

Where the null hypothesis is homogeneity (i.e. all studies have the same true effect parameter that may or may not be equivalent to zero), and the implicit alternative hypothesis that at least one study has an effect that is different from θ . It also is possible to apply directional alternative hypotheses involving the direction of the relationship (e.g. at least one study had a larger effect than θ). As described above, assumptions of certain heterogeneity statistics (like Q that follows a χ^2 distribution with $k - 1$ degrees of freedom) can be tested statistically. But, the Q must be interpreted in consideration of the other heterogeneity metrics because alone the Q statistic needs to be interpreted carefully (as described above).

Example of non-significant heterogeneity

The figure below (**Figure 4.1**) is an example of heterogeneity statistics generated with a DerSimonian and Laird (D+L) random effects meta-analysis of 22 trials with a binary outcome estimating risk ratio (RR). Note that although the heterogeneity statistics are associated with a random effects meta-analysis, they are in fact generated from a fixed-effect model using the inverse variance method of estimation.



Is there evidence of heterogeneity in effect sizes? In this instance, the heterogeneity statistic (Q) is 29.76, and we do not have sufficient evidence to refute the null hypothesis of homogeneity with a p-value that exceeds 0.05. Hence, we can conclude that there is homogeneity in effect sizes across studies or in other words that there is no significant heterogeneity. We must use caution and avoid interpreting the non-significant heterogeneity as no heterogeneity because the number of trials in this meta-analysis is quite low and because of empirical evidence that heterogeneity exists whether or not we can detect it statistically.

What is the ratio of total to excess dispersion in effect sizes? In this instance, the Q of 29.76 is greater than the model degrees of freedom (i.e. 21). Hence we have some excess dispersion in effect sizes given the number of studies included. But, we have already learned from prior testing that there is homogeneity in effect sizes. Therefore, in this instance the excess to total dispersion is not very informative.

What is the between-study variance of the true effects? The τ^2 remains in the metric of the effect size. In this example, the τ^2 of 0.0132 indicates limited variance in risk ratio across studies. The small value of τ^2 also is concordant with findings of homogeneity and limited excess dispersion from the statistics above.

Are there substantive implications of this heterogeneity? Although much more informative in instances of significant heterogeneity, we can convert the τ^2 to τ by taking the square root of τ^2 . Doing so allows us to consider the standard deviation of dispersion in effects sizes across studies (as opposed to variance). In this example, the standard deviation of dispersion in effects sizes across studies (τ) in the metric of risk ratio is 0.11. In combination, the pooled estimate in risk ratio of 0.757 and the standard deviation of dispersion in effects sizes across studies in risk ratio of 0.11 may help making claims about the substantive implications of heterogeneity. In this example, however, we have evidence of homogeneity in effect sizes across studies. Thus, reporting additional metrics beyond the heterogeneity chi-square (Q) and associated p-value would facilitate transparency in presentation but may otherwise be minimally informative.

What proportion of the observed heterogeneity is real vs. spurious? Although the variation in effect sizes across studies was minimal (non-significant Q, minimal excess dispersion (Q-df), and small τ^2 and τ), the proportion of the variation in risk ratio across studies attributable to heterogeneity (I^2) in this example was 29.4%. Even in instances of significant heterogeneity in effect sizes across studies, an I^2 of 29.4% would be considered low. In this example, there was not significant heterogeneity in effect sizes across studies; but, knowing the proportion of variance in effects that was due to heterogeneity is informative.

Example of significant heterogeneity

The figure below (**Figure 4.2**) is an example of heterogeneity statistics generated with a DerSimonian and Laird (D+L) random effects meta-analysis of 20 trials with a continuous outcome estimating standardized mean difference (SMD). Note that although the heterogeneity statistics are associated with a random effects meta-analysis, they are in fact generated from a fixed-effect model using the inverse variance method of estimation.

Figure 4.1. Output of a meta-analysis with significant heterogeneity

D+L pooled SMD	0.472	0.310	0.635	100.00
Heterogeneity chi-squared = 54.39 (d.f. = 19) p = 0.000				
I-squared (variation in SMD attributable to heterogeneity) = 65.1%				
Estimate of between-study variance Tau-squared = 0.0703				
Test of SMD=0 : z= 5.70 p = 0.000				

Is there evidence of heterogeneity in effect sizes? In this instance, the Q is 54.39, and we have sufficient evidence to reject the null hypothesis of homogeneity with a p-value that is less than 0.05. Hence, we can conclude that there is significant heterogeneity in effect sizes across studies. In the case of significant heterogeneity like as in this example, the subsequent statistics become much more informative compared with homogeneity of effect sizes.

What is the ratio of total to excess dispersion in effect sizes? In this instance, the Q of 54.39 is much greater than the model degrees of freedom (i.e. 19). Hence we have evidence of excess dispersion in effect sizes given the number of studies included in the meta-analysis that is much greater compared with the prior example. Examining $Q-df$ is most helpful in understanding the magnitude of heterogeneity given the scope of the analysis.

What is the between-study variance of the true effects? The τ^2 remains in the metric of the effect size. In this example, the τ^2 of 0.0703 indicates that variance in standardized mean difference (i.e. Hedges' g) across studies is larger compared with our prior example. The value of τ^2 also is concordant with findings of heterogeneity and excess dispersion from the statistics above compared with the prior example.

Are there substantive implications of this heterogeneity? We can convert the τ^2 to τ by taking the square root of τ^2 . Doing so allows for us to consider the standard deviation of dispersion in effects sizes across studies (as opposed to variance). In this example, the standard deviation of dispersion in effects sizes across studies (τ) in the metric of standardized mean difference is 0.265. In combination, the pooled estimate in standardized mean difference of 0.472 and the standard deviation of dispersion in effects sizes across studies in standardized mean difference of 0.265 may help making claims about the substantive implications of heterogeneity. In this example, however, we have evidence of heterogeneity in effect sizes across studies and have evidence that the heterogeneity is substantive given the pooled estimate derived from the model.

What proportion of the observed heterogeneity is real vs. spurious? The variation in effect sizes across studies was significant (significant Q , excess dispersion ($Q-df$), and larger τ^2 and τ), and from the I^2 we learn that the proportion of the variation in standardized mean difference across studies attributable to heterogeneity in this example was 65.1%. An I^2 of 65.1% would be considered moderate to large in most circumstances. Therefore, in this example, there was not significant heterogeneity in effect sizes across studies. In addition to being transparent in presentation by reporting these heterogeneity statistics, the potential sources and means of reducing heterogeneity should be considered.

4.5 Exploring Heterogeneity

Investigators are frequently interested in understanding which studies, and subsequently which study-level factors, may be associated with statistical heterogeneity that is estimated in meta-analyses. Minimally, heterogeneity should be explored descriptively and graphically as described below. In the special circumstances that specified, scientifically-defensible and

hypothesis-based study-level factors are identified, investigators may choose to complement preliminary descriptive and graphical approaches with subgroup or other meta-analytic approaches also described below.

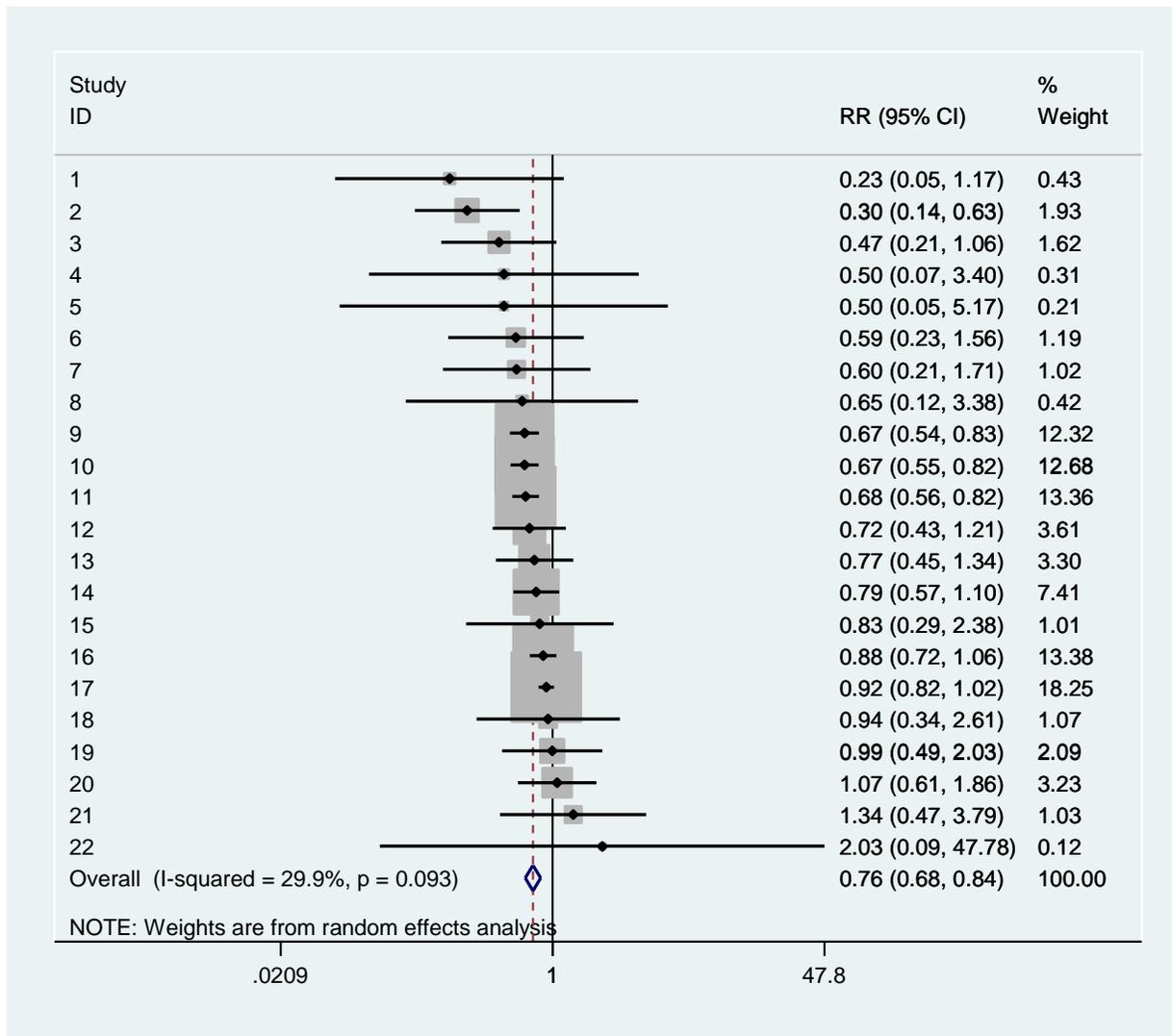
Descriptive and Graphical Approaches

Although simple histograms, box plots and other related graphical methods of depicting effect estimates across studies may be helpful preliminarily, these approaches do not necessarily provide insight into statistical heterogeneity; but, there are several graphics designed specifically for the interpretation of meta-analytic results.¹²⁵

Forest plots

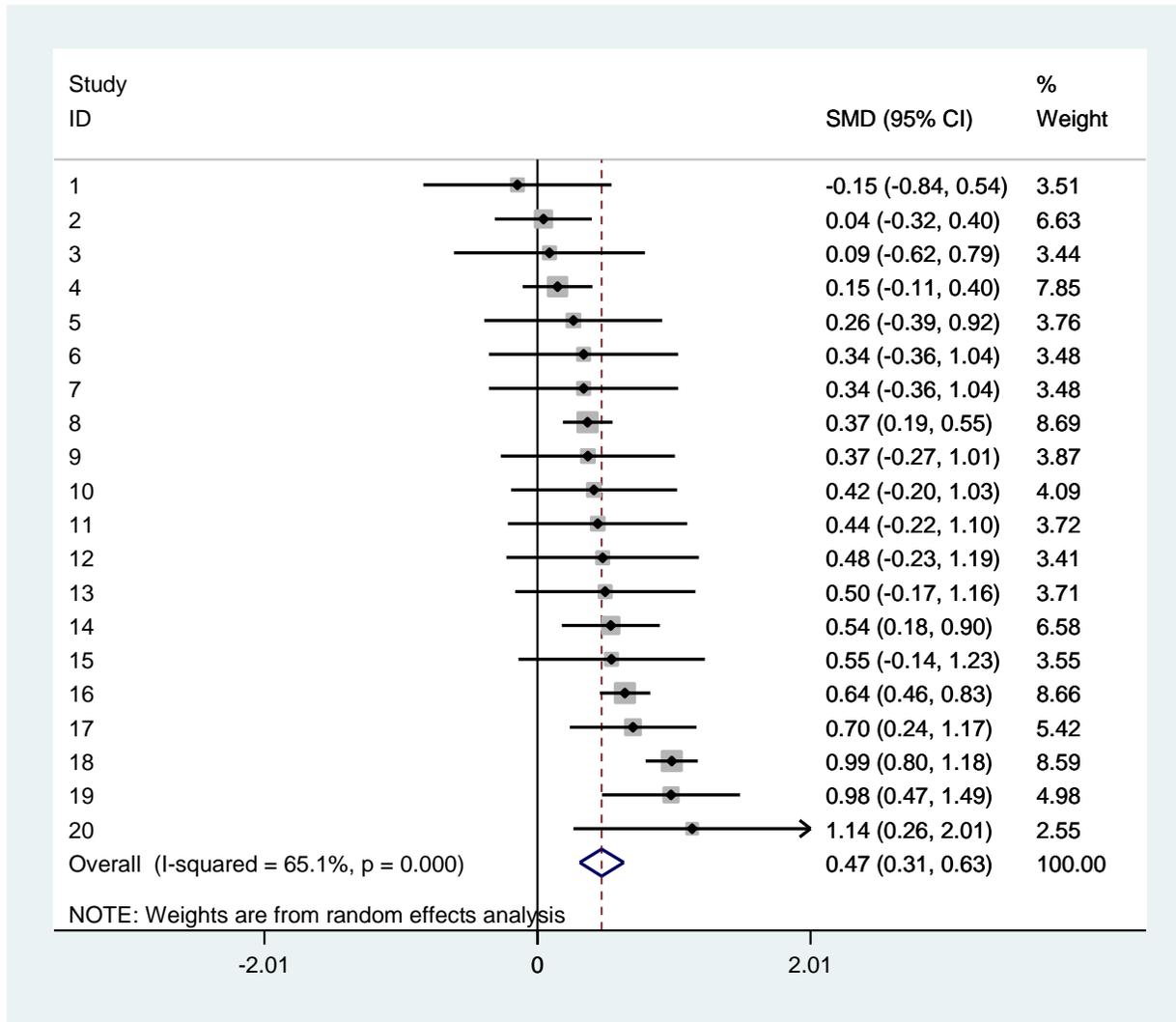
Forest plots themselves can help identify potential sources and the extent of statistical heterogeneity. Meta-analyses with limited heterogeneity will produce forest plots with significant and grossly visual overlap among confidence intervals of the studies included. In contrast, a sign of statistical heterogeneity would be poor overlap among confidence intervals of the studies included in a meta-analysis.¹²⁵ An example of a forest plot that shows considerable overall in the confidence intervals of each study (in risk ratio (RR)) is presented in the figure below (**Figure 4.2**) ($Q = 29.76$, $p=0.097$; $I^2 = 29.9\%$).

Figure 4.3. Forest plot without significant heterogeneity



An example of a forest plot that shows poor overall in the confidence intervals of each study in standardized mean difference (SMD) is presented in the figure below (**Figure 4.4**) ($Q = 54.39, p < 0.0001; I^2 = 65.1\%$).

Figure 4.4. Forest plot with significant heterogeneity



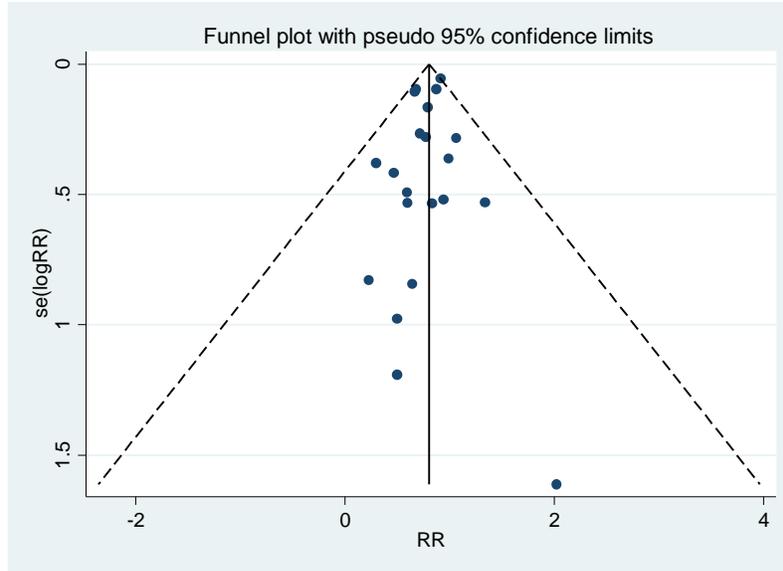
It is important to note that in each example the graphic depiction matches the heterogeneity statistics derived from the meta-analysis. We recommend performing graphic and quantitative exploration of heterogeneity in combination.

Funnel plots

It is often the case that funnel plots are thought of representing bias but they also can aid in detecting sources of heterogeneity as it may not be possible to distinguish between bias and heterogeneity using graphical means alone. Funnel plots are essentially the plotting of effect sizes observed in each study (x-axis) around the summary effect size vs. the degree of precision of each study (typically by standard error, variance or precision on the y-axis). A meta-analysis that includes studies that estimate the same underlying effect across a range of precision and has limited bias and heterogeneity would result in a funnel plot that resembles a symmetrical inverted funnel shape with increasing dispersion ranging with less precise (i.e. smaller) studies.¹²⁵ Funnel plots should be considered for the preliminary analysis of statistical

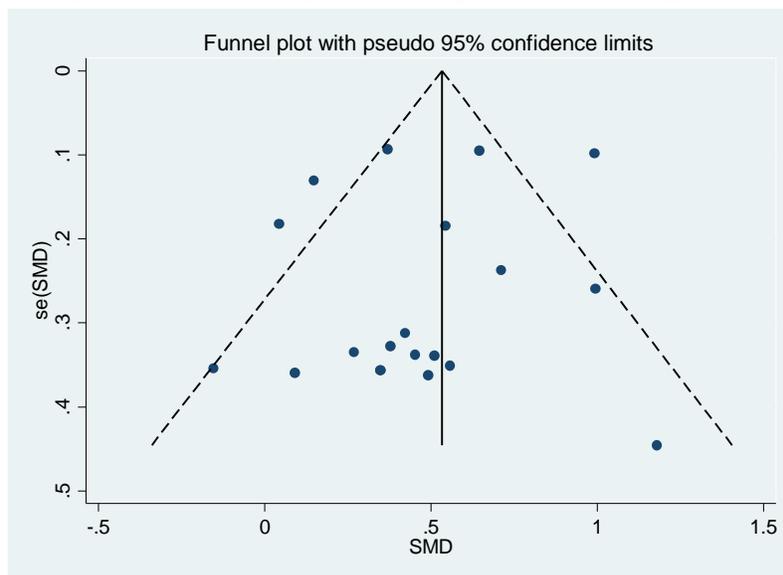
heterogeneity. An example of a symmetric funnel plot around the summary estimate (in risk ratio (RR)) with increasing dispersion with decreasing precision is presented in the figure below (**Figure 4.5**) ($Q = 29.76$, $p=0.097$; $I^2 = 29.4\%$).

Figure 4.5. Funnel plot without significant heterogeneity



In the event of heterogeneity and/or bias, funnel plots will take on an asymmetric pattern around the summary effect size and also provide evidence of scatter outside of the bounds of the 95% confidence limits. An example of a funnel plot with evidence of asymmetry around the summary estimate (in standardized mean difference (SMD)) and studies falling outside of the confidence limits is presented in the figure below (**Figure 4.6**) ($Q = 54.39$, $p<0.0001$; $I^2 = 65.1\%$).

Figure 4.6. Funnel plot with significant heterogeneity



Normal probability plots

Normal probability plots may be used in order to check the distributional assumptions made under random-effects models. Hardy and Thompson first proposed investigating the contribution that each study makes to the overall test statistic for heterogeneity (i.e. Q).¹²⁶ A normal probability plot is generated presenting q_i against $\Phi^{-1}(F_k(q_{(i)}))$, when under a random effects model:

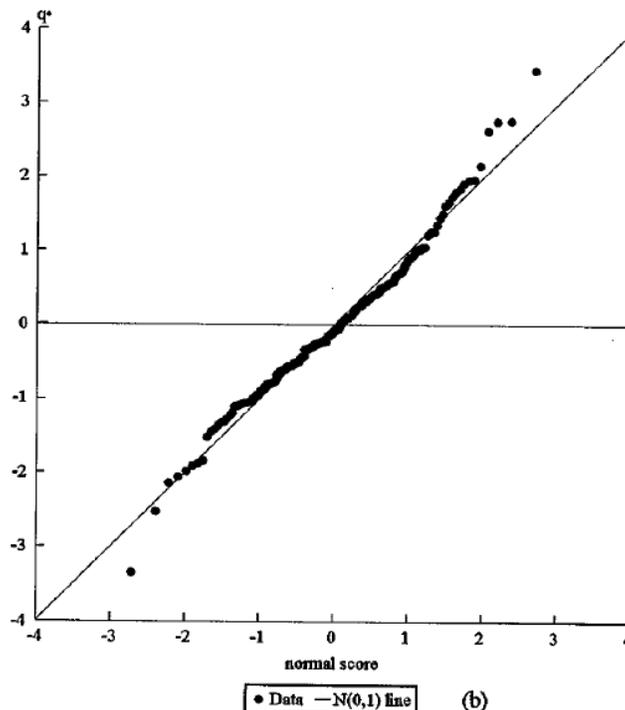
$$q_i^* = (\hat{\theta}_i - \hat{\theta}^*) / \sqrt{(\hat{v}_i + \hat{\sigma}_B^2)},$$

$$F_k(q_{(i)}) = (i - 3/8) / (k + 1/4)$$

where Φ is the standard distribution function,
 $\hat{\theta}^*$ is the random-effects estimate
 \hat{v}_i is the within-study variance,
 $\hat{\sigma}_B^2$ is the between-study variance, and
 $F_k(q_{(i)})$ is approximated are depicted above.

Using this approach, it is assumed that q_i will have an approximate standard normal distribution under a normally distributed random-effects model. An example of a normal probability plot from the original paper¹²⁶ is presented in the figure below (**Figure 4.7**). Normal probability plots can be adapted so that studies are labeled for visual identification of those that deviate from the normal distribution and are therefore exerting greater influence on overall heterogeneity.

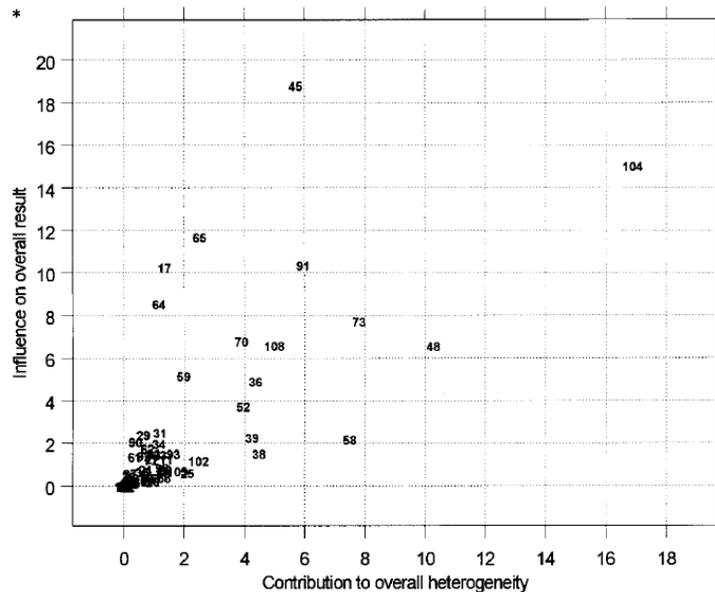
Figure 4.7. Normal probability plot



Baujat plot

Baujat and colleagues proposed a graphical method to identify studies that have the greatest impact on heterogeneity.¹²⁷ Baujat proposed plotting the contribution to the heterogeneity statistic for each study (q_i^F) on the horizontal axis, and the squared difference between meta-analytic estimates with and without the i^{th} study divided by the estimated variance of the meta-analytic estimate without the i^{th} study along the vertical axis. Because of Baujat plot presentation, studies that have the greatest influence on heterogeneity will be located in the upper right corner for easy visual identification. An example of a normal probability plot from the original paper¹²⁷ is presented in the figure below (**Figure 4.8**). Note that studies in the example Baujat plot are labeled for easy identification. Also see the more recent paper by Bowden and colleagues¹²⁸ for additional applied examples of Baujat plots in meta-analyses of randomized trials.

Figure 4.8. Baujat plot



Meta-Regression

Meta-regression is a common approach employed to examine the degree to which study-level factors explain statistical heterogeneity.¹²⁹ Random-effects meta-regression, as compared with fixed-effect meta-regression, allows for residual heterogeneity or in other words among-study variance that is not explained by study-level factors incorporated into the model.¹³⁰ Because of this feature, among other benefits described below, random-effects meta-regression is preferred over fixed-effect meta-regression.¹³¹ Random-effects meta-regression takes on the following general form under the assumption that true effects follow a normal distribution around the linear prediction:

$$y_i = x_i\beta + \mu_i + \epsilon_i \text{ when}$$

$$\mu_i \sim N(0, \tau^2) \text{ and}$$

$$\epsilon_i \sim N(0, \sigma_i^2)$$

Where y_i is the linear prediction for the effect size of study i ,
 x_i is a $1 \times k$ vector of covariates values in study i ,
 β is a $k + 1$ vector of coefficients in study i ,
 τ^2 is the estimate of among study variance, and
 σ_i is the standard error in study i .

It is the default of several statistical packages to use a modified estimator of variance in effect estimates generated by random-effects meta-regression that also employs a t distribution in lieu of a standard normal distribution when calculating p-values and confidence intervals (i.e. the Knapp-Hartung modification).¹³² This approach is recommended to help mitigate false-positive rates that are common in meta-regression.¹³¹ Since the earliest papers on random-effects meta-regression there has been general caution about the inherent low statistical power in analyses when there are fewer than 10 studies for each study-level factor modelled.¹³⁰ Currently, the Cochrane manual recommends that there be at least 10 studies per characteristic modelled in meta-regression¹⁵ over the enduring concern about inflated false-positive rates with too few studies.¹³¹ Another consideration that is reasonable to endorse is adjusting the level of statistical significance to account for making multiple comparisons in cases where more than one characteristic is being investigated in meta-regression.

Beyond statistical considerations important in meta-regression, there are also several important conceptual considerations. First, study-level characteristics to be considered in meta-regression should be pre-specified, scientifically defensible and based on hypotheses.^{8, 15} This first consideration will allow investigators to focus on factors that are believed to modify the effect of intervention as opposed to clinically meaningless study-level characteristics. Arguably, it may not be possible to identify all study-level characteristics that may modify intervention effects until all of the worlds' evidence is collected and synthesized. Minimally, however, the focus of meta-regression should be on factors that are plausible and based on scientifically-defensible hypotheses. Second, meta-regression should be carried out under full consideration of ecological bias (i.e. the inherent problems associated with aggregating individual-level data).¹³³ As classic examples, the mean study age or the proportion of study participants who were female may result in different inferences when included in meta-regression as opposed to the intervention effect modifying relationships that was observed in each trial.¹²⁹ Hence, using study-level characteristics as opposed to summary measures of individual-level data (e.g., average age, percent female) should be avoided.

Example random-effects meta-regression

The following is a simulated example of a random-effects meta-analysis followed by a random-effect meta-regression of the study-level characteristic of mean sample age. Based on the random-effects meta-analysis of 49 trials, the intervention effect was estimated as a risk ratio of 1.033 with a confidence interval that crossed to neutral value of 1.000 (i.e. 0.906-1.178); there also was significant and moderate heterogeneity based on the Q and I^2 statistics presented in the figure below (**Figure 4.9**).

Figure 4.9. Meta-analysis output indicating significant heterogeneity

```

-----+-----
D+L pooled RR      |  1.033      0.906      1.178      100.00
-----+-----

Heterogeneity chi-squared =  71.86 (d.f. = 48) p = 0.014
I-squared (variation in RR attributable to heterogeneity) =  33.2%
Estimate of between-study variance Tau-squared =  0.0516

Test of RR=1 : z=  0.49 p = 0.627

```

Based on extensive preliminary comparisons of trial data indicating that the intervention under study was employed across a wide range of patient age as well as anecdotal published evidence of an age-related treatment effect (works well in younger but not older patients), it was hypothesized prior to the analysis that age would moderate the treatment effect in the event of significant statistical heterogeneity. Note that the consideration of including study age as a factor in subsequent meta-regression was pre-specified, scientifically defensible and based on a specific hypothesis. A random-effects meta-regression was performed using mean study age as the single study-level characteristic (the statistical output (generated using Stata's **metareg** command) is presented in the figure below (**Figure 4.10**)).

Figure 4.10. Meta-regression output

```

Meta-regression                               Number of obs =      49
REML estimate of between-study variance       tau2                = .006654
% residual variation due to heterogeneity     I-squared_res      =  0.00%
Proportion of between-study variance explained Adj R-squared      =  91.10%
With Knapp-Hartung modification

```

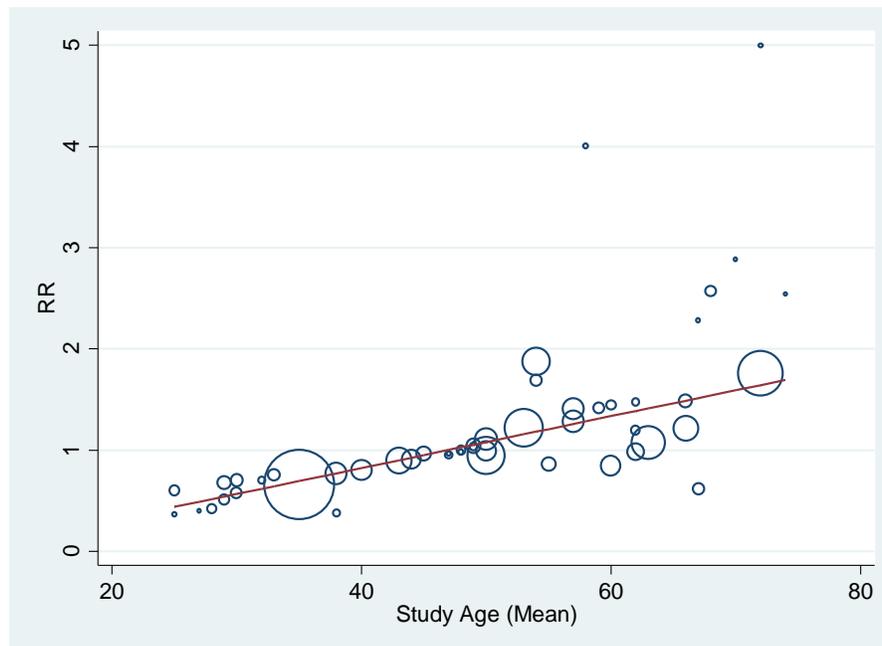
_ES	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
StudyAge	.0256257	.0038948	6.58	0.000	.0177904 .033461
_cons	-.2022906	.203356	-0.99	0.325	-.61139 .2068089

Based on the meta-regression, we observe that there is a precise linear association between the study-level characteristic of mean age and the intervention effect indicating a higher risk associated with studies involving a higher mean patient age. Although there are several alternatives including method of movements and empirical Bayes, the default estimator of τ^2 using this command is REML that is based on maximization of the residual log likelihood. As discussed earlier, the Knapp-Hartung modification was applied. One additional key finding from this simulated meta-regression is that the residual I^2 was reduced to 0.0% by including mean study age as the single study-level characteristic, and that nearly all of the observed heterogeneity was explained by this single factor (adjusted R^2 of 91.10%).

Like most other elements of meta-analysis and the extensions thereof, graph are helpful in interpreting the findings of the analyses. The figure below depicts the risk ratio (RR) of each study on the y-axis (with random-effect weights depicted as the volume of circles (bigger circles = more weight)), the mean study age on the x-axis, and the linear prediction line generated from the random-effects meta-regression. What can be particularly helpful with a graph of meta-regression is insight into when benefit changes to risk. That is, given the neutral value of risk

ratio of 1.00 we have evidence of intervention benefit in trials of younger persons on average, and we have evidence of intervention risk in trials of older persons on average.

Figure 4.11. Meta-regression graph



Subgroup Analysis

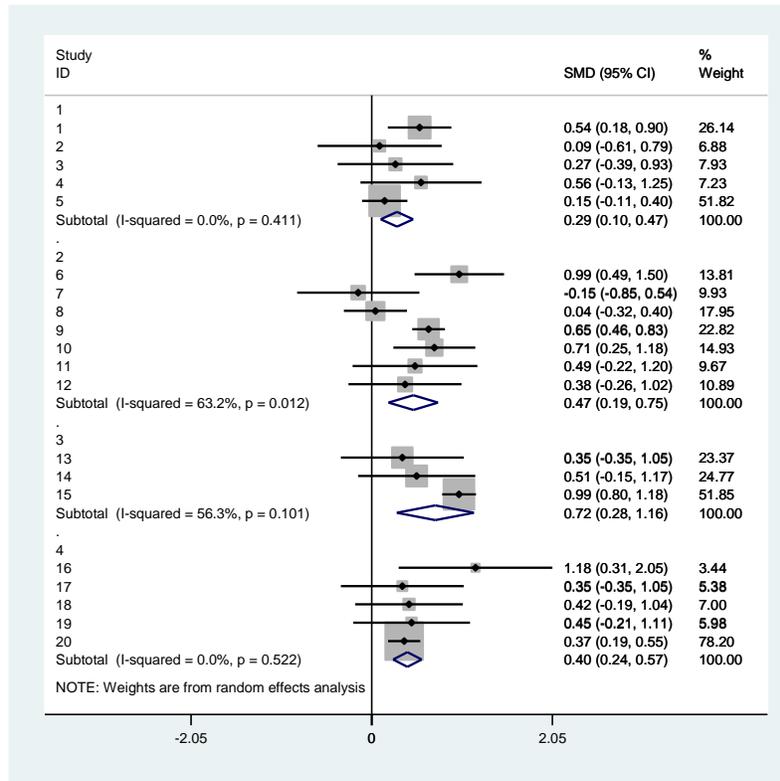
Subgroup analysis is another common approach employed to examine the degree to which study-level factors explain statistical heterogeneity. Since subgroup analysis is a type of meta-regression that incorporates a categorical study-level factor as opposed to a continuous study-level factor, it is similarly important that the grouping of studies to be considered in subgroup analysis be pre-specified, scientifically defensible and based on hypotheses.^{8, 15} Like other forms of meta-regression, subgroup analyses have a high false-positive rate¹³¹ and may be misleading when few studies are included in the meta-analysis. There are two general approaches to handling subgroups in meta-analysis. First, it is an option to perform meta-analyses within subgroups without any statistical among-group comparisons. A central problem with this approach is the tendency to misinterpret results from within separate groups as being comparative. That is, identification of groups wherein there is a significant summary effect and/or limited heterogeneity and others wherein there is no significant summary effect and/or substantive heterogeneity does not necessarily indicate that the subgroup factor explains heterogeneity.¹⁵ Second, it is recommended to incorporate the subgrouping factor into a meta-regression framework. Doing so allows for quantification of both within and among subgroup quantification of overall summary effects and heterogeneity as well as well as formal statistical testing that informs whether or not the summary estimates are different across subgroups. Moreover, subgroup analysis in a meta-regression framework will allow for formal testing of residual heterogeneity in a similar fashion compared with meta-regression using a continuous

study-level factor. The example below presents data within subgroups but then shifts focuses towards subgroup analysis using meta-regression.

Example subgroup analysis

The figure below is an example of a subgroup meta-analysis of 20 drug trials with a continuous outcome estimating using standardized mean difference. In this example, reductions in the continuous outcome were examined in four drug classes vs a comparator drug. If our only interest was how the comparator drug compared with each inherent subgroup of drug classes, we could perform a meta-analysis within each drug class (graph and output below (**Figure 4.12**) generated using Stata's **metan** command with the **by(subgroup)** and **nooverall** options).

Figure 4.12. Subgroup meta-regression output



Test(s) of heterogeneity:							
	Heterogeneity statistic	degrees of freedom	P	I-squared**	Tau-squared	Significance test(s) of SMD=0	
1	3.96	4	0.411	0.0%	0.0000	z= 3.04	p = 0.002
2	16.33	6	0.012	63.2%	0.0806	z= 3.30	p = 0.001
3	4.58	2	0.101	56.3%	0.0867	z= 3.23	p = 0.001
4	3.22	4	0.522	0.0%	0.0000	z= 4.90	p = 0.000

Based on our interest into how the comparator drug compared with each inherent subgroup of drug classes, we can have evidence that there were some subgroups where there was significant heterogeneity (i.e. drug class 2 versus the comparator drug) while there was no significant heterogeneity in others. We also have evidence that the estimate in standardized mean difference was precise within each subgroup of drug classes. Based on this information we may

want to comment generally on the difference in magnitude of the effect sizes between groups. For example, we could comment that the effect observed in trials of drug class 3 vs the comparator was more than double the effect observed in trials of drug class 1 vs the comparator. We must resist the temptation to interpret other findings as being comparative, however, and also resist indicating whether or not examining subgroups reduced overall heterogeneity because that is not formally tested using this approach.

The recommended alternative to within-subgroup analysis is incorporating the subgrouping factor into a meta-regression framework. The output below (generated using Comprehensive Meta Analysis v3) provides detailed information on both fixed effect and mixed effects subgroup meta-analysis in a meta-regression framework (**Figure 4.13**).

Figure 4.13. Subgroup meta-regression output (meta-analytic framework)

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared
Fixed effect analysis												
1.000	5	0.287	0.094	0.009	0.102	0.471	3.039	0.002	3.965	4	0.411	0.000
2.000	7	0.536	0.071	0.005	0.397	0.675	7.547	0.000	16.337	6	0.012	63.274
3.000	3	0.915	0.091	0.008	0.736	1.094	10.036	0.000	4.581	2	0.101	56.344
4.000	5	0.405	0.083	0.007	0.243	0.567	4.907	0.000	3.261	4	0.515	0.000
Total within									28.144	16	0.030	
Total									26.773	3	0.000	
Overall	20	0.533	0.042	0.002	0.452	0.615	12.815	0.000	54.917	19	0.000	65.402
Mixed effects analysis												
1.000	5	0.287	0.094	0.009	0.102	0.471	3.039	0.002				
2.000	7	0.472	0.143	0.020	0.192	0.753	3.302	0.001				
3.000	3	0.722	0.223	0.050	0.284	1.160	3.230	0.001				
4.000	5	0.405	0.083	0.007	0.243	0.567	4.907	0.000				
Total									3.769	3	0.287	
Overall	20	0.394	0.055	0.003	0.286	0.502	7.133	0.000				

For each drug class, the number of studies included, and estimate, standard error, variance and confidence interval in the metric of standardized mean difference are presented. Note that these results are slightly different from those without the meta-regression framework (Figure 4.12). The z and p-values of summary estimates by drug class vs. the comparator drug are presented under the Test of null column. In this example, precise estimates were derived for each drug class comparison. Heterogeneity statistics for each drug class vs. the comparator drug are presented under the heterogeneity column. The overall model heterogeneity was significant ($Q = 54.9$, $p < 0.001$; $Q-df = 35.9$ (excess); $I^2 = 65.4\%$ (moderate-to-high)). When examining heterogeneity by drug class, however, we observe that there was homogeneity with certain drug classes (class 1 and 4 vs the comparator), and significant heterogeneity with others (i.e. drug class 2 vs. the comparator). Thus, with this case as an example we can use subgroup analysis as a means to explore the reasons for significant overall model heterogeneity.

Additional metrics derived from subgroup meta-analyses are the total between chi-square tests (total Q in the output above) that inform whether or not the pooled summary estimates are different across subgroups. In this example, the chi-square test of between-group differences in summary estimates by **fixed effect** analysis was 26.773 and significant ($p < 0.001$). The literal interpretation in this example would be that reductions in the continuous outcome are different

depending on which drug class is being tested against the comparator (i.e. the observed effect was not constant across drug classes). The chi-square test of between-group differences in summary estimates by **mixed effects** analysis, however, was 3.769 and not significant ($p=0.287$). The literal interpretation in this example would be that reductions in the continuous outcome are constant across drug classes tested against the comparator drug. The biggest difference between analyses is that the Q statistic generated from a fixed effect subgroup analysis is cumulative and it is not cumulative in the mixed effects subgroup analysis. Hence, the mixed effects subgroup analysis allows us to consider that the true effects vary within groups and that within-group random effects weights should be applied.

Finally, we also can incorporate the categorical variable indicating subgroup into a meta-regression to assess the influence on I^2 and residual I^2 . The related output (generated from Stata's **metareg** command) is presented in the figure below (**Figure 4.14**).

Figure 4.14. Subgroup meta-regression output

Meta-regression	Number of obs =	20
REML estimate of between-study variance	tau2 =	.04431
% residual variation due to heterogeneity	I-squared_res =	43.03%
Proportion of between-study variance explained	Adj R-squared =	26.10%
Joint test for all covariates	Model F(3,16) =	1.29
With Knapp-Hartung modification	Prob > F =	0.3132

Although we have reduced I^2 from 65.5% by incorporating the subgrouping variable into meta-regression, we still have moderate residual I^2 of 43.0% and evidence of a poor fitting model. Hence, under a meta-regression framework we cannot consider that subgroup analysis reduced heterogeneity in the overall meta-analysis in this example.

4.6 Detecting Outlying Studies

Under full consideration that removal of one or more studies from a meta-analysis may interject bias in the results,¹⁵ clear identification of outlier studies may help build the evidence necessary to justify study removal. Visual examination of forest, funnel, normal probability and Baujat plots (described in detail earlier in this chapter) alone may be helpful in identifying studies with inherent outlying characteristics. Additional procedures that may be helpful in interpreting the influence of single studies are quantifying the summary effect without each study (often called one study removed), and performing cumulative meta-analyses. The table below (**Table 4.1**) presents results from the one study removed procedure (performed in Comprehensive Meta Analysis v3).

Table 4.1 Summary Statistics with One Study Removed

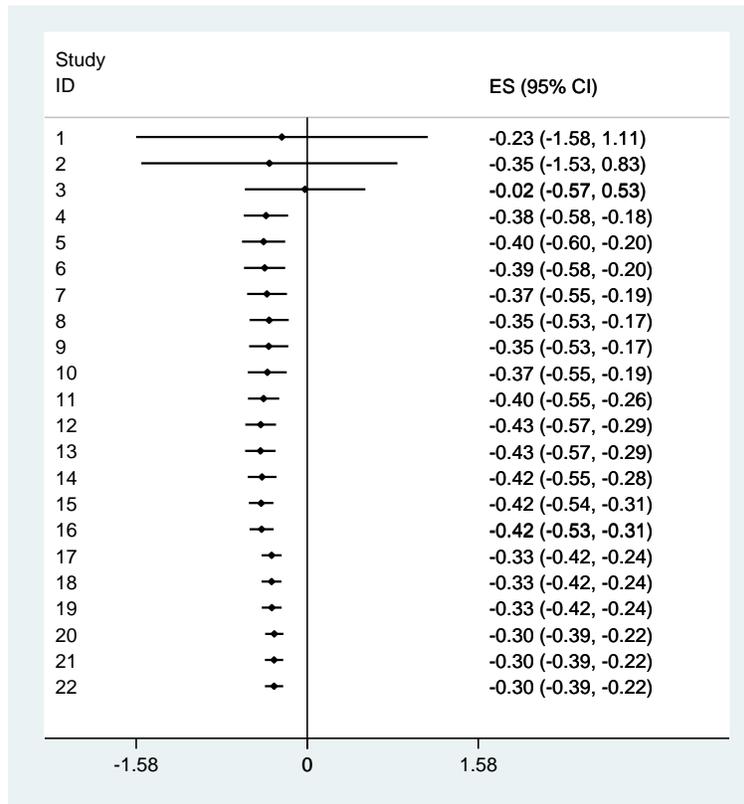
Study number	RR	LL95%CI	UL95%CI	p-Value
1	0.437	0.425	0.450	0.000
2	0.437	0.424	0.450	0.000
3	0.439	0.427	0.452	0.000
4	0.438	0.425	0.452	0.000
5	0.433	0.421	0.446	0.000
6	0.438	0.425	0.450	0.000
7	0.438	0.425	0.451	0.000

8	0.433	0.420	0.445	0.000
9	0.441	0.428	0.454	0.000
10	0.434	0.419	0.448	0.000
11	0.438	0.425	0.451	0.000
12	0.432	0.419	0.445	0.000
13	0.434	0.422	0.447	0.000
14	0.434	0.422	0.447	0.000
15	0.450	0.437	0.463	0.000
16	0.431	0.418	0.443	0.000
17	0.441	0.428	0.454	0.000
18	0.411	0.399	0.424	0.000
19	0.448	0.435	0.461	0.000
20	0.441	0.428	0.454	0.000
21	0.440	0.427	0.453	0.000
22	0.435	0.422	0.448	0.000
23	0.436	0.424	0.449	0.000
24	0.432	0.419	0.444	0.000
25	0.417	0.404	0.429	0.000
26	0.434	0.420	0.447	0.000
Summary Effect	0.435	0.423	0.448	0.000

In this example, it may be that study 18 has characteristics that may exert an influence on the overall effect estimate as without its inclusion the summary effect would be lower (risk ratio of 0.411 vs the summary effect of 0.435).

Using cumulative meta-analysis,¹³⁴ it is possible to graph the accumulation of evidence of trials reporting at treatment effect. Simply put, this approach integrates all information up to and including each trial into summary estimates. By looking at the related graphical output ((**Figure 4.15**) example below from Stata's **metacum** command), we can examine large shifts in the summary effect that may lead use to examine study-level factors that should be considered when considering such studies potential outliers. In the example depicted below, we would want to examine trials number 3 and 17 as their addition to the meta-analysis results in a shift in the summary estimate. Irrespective of how potential outlying studies are detected (graphically using forest, funnel, normal probability or Baujat, or by one study removed of cumulative analysis), sensitivity analyses should be performed to quantify what changed comparing study inclusion and removal and to what degree. Specifically, the influence of study inclusion and removal on both the summary effect and heterogeneity should be quantified and presented for transparency.¹⁵

Figure 4.15. Cumulative meta-analysis



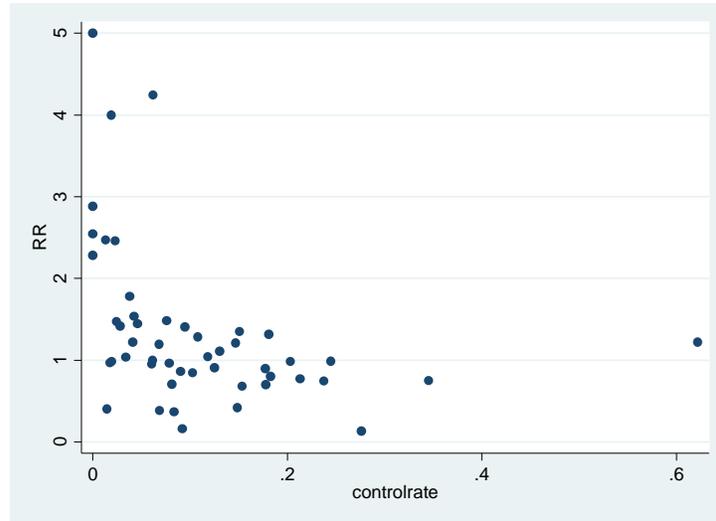
4.7 Special topics

Baseline risk (control-rate) meta-regression

For studies with binary outcomes, the “control rate” refers to the proportion of subjects in the control group who experienced the event. The control rate is viewed as a surrogate for covariate differences between studies because it is influenced by illness severity, concomitant treatment, duration of follow-up and/or other factors that differ across studies.^{29, 30} Patients with higher underlying risk for poor outcomes may experience different benefits and/or harms from treatment than patients with lower underlying risk.³¹ Hence, the control-rate can be used to test for interaction between underlying population risk at baseline and treatment benefit, particularly in the setting of significant heterogeneity or otherwise known differences in control rate across studies. To examine for an interaction between underlying population risk and treatment benefit, we recommend the following approach. First, generate a scatter plot of treatment effect against control rate as a useful preliminary approach to visually assess whether there may be a relation between the two. Since the RD is frequently highly correlated with the control rate,³⁰ we recommend using a RR or OR when examining a treatment effect against the control rate in all steps. In a simulated example involving 49 trials ($Q = 106.14$, $p < 0.001$, $I^2 = 54.8\%$), the scatter plot of treatment effects in each study against the control-rate in each study reveals that higher risk ratios (particularly those above 2.0) only entailed studies wherein the control group event rate was less than 0.10 (**Figure 4.16**). This gives us preliminary insight into how differences in

baseline risk (control rate) may influence the amount of heterogeneity observed in the meta-analysis.

Figure 4.16. Effect size against control rate



Second, generate a simple weighted regression of the effect size on the control rate. Simple weighted regressions tend to identify a significant relation between control rate and treatment effect twice as often compared with more suitable approaches (below).^{30,32} A negative finding based on a simple weighted regression (i.e. slope not significantly different from zero) would be most likely replicated by the more complicated methods, and a positive finding (i.e. slope significantly different from zero) would need to be verified by a more comprehensive method. In the working example, we observe that using simple weighted regression there is a significant relationship between the control rate and intervention effect (specifically favoring a reduction in the intervention effect in samples with higher baseline risk) (**Figure 4.17**).

Figure 4.17. Simple weighted linear regression output

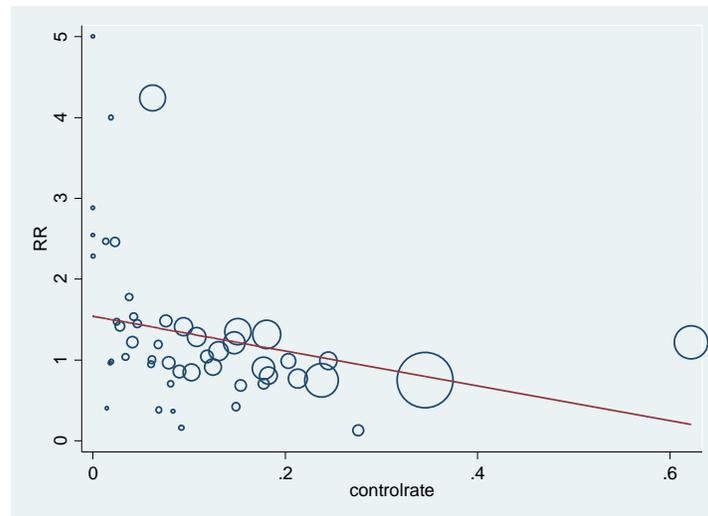
Source	SS	df	MS	Number of obs	=	49
Model	2.81809884	1	2.81809884	F(1, 47)	=	4.61
Residual	28.7124501	47	.610903193	Prob > F	=	0.0369
Total	31.5305489	48	.656886435	R-squared	=	0.0894
				Adj R-squared	=	0.0700
				Root MSE	=	.7816

_ES	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
controlrate	-1.85875	.8654244	-2.15	0.037	-3.59976 - .117741
_cons	1.525555	.1784839	8.55	0.000	1.166492 1.884619

Third, if there is a positive finding based on a simple weighted regression, consider using hierarchical meta-regression models³⁰ or Bayesian meta-regression³² models to validate and refine the presence of an interaction between underlying population risk and treatment benefit using formal control rate meta-regression. These approaches incorporate the covariate of control rate in explaining variance in the treatment effect under the hypothesis that the control rate is a

surrogate for differences in baseline risk among studies.³³ With the working simulated example, we continue with a meta-regression using control rate as study-level factor to explain heterogeneity in the meta-analysis. Indeed, in this simulated example control-rate was confirmed as an influential factor using meta-regression ($\beta = -2.16$, $SE = 1.00$, $z = -2.16$, $p = 0.036$) (graph below generated from Stata's **metareg** command (**Figure 4.18**)). It is also clear that not all studies fit nicely on this meta-regression line; there is considerable residual I^2 even after control-rate is incorporated into the model.

Figure 4.18. Effect size against control rate



Multivariate meta-regression

It may be desirable to examine the influence of more than one study-level factor on the heterogeneity observed in meta-analyses. Recalling general cautions and specific recommendations about the inherent low statistical power in analyses wherein there is fewer than 10 studies for each study-level factors modelled,^{15, 130, 131} multivariate meta-regression should only be considered when study-level characteristics to be considered are pre-specified, scientifically defensible and based on hypotheses, and when the number of studies meets or exceeds 10 studies for each study-level factor included in meta-regression.

As an example using simulated trial data, we may want to consider the influence of both the control-rate (as a proxy for differences in baseline risk among studies) and the mean age of the trial samples (based on published evidence of an age-related treatment effect). In this same simulated sample, we observed a difference in treatment effect by control-rate. Hence, we may want to disambiguate the influence of differences in baseline risk vs. differences in sample mean age on the heterogeneity we observed in the meta-analysis. The figure below (**Figure 4.19**) presents output from a random effects meta-regression¹³¹ involving these two study-level characteristics (control rate and age) as generated using Comprehensive Meta Analysis v3).

Figure 4.19. Multivariate meta-regression

Main results for Model 1, Random effects (REML), Knapp Hartung, Log risk ratio

Covariate	Coefficient	Standard Error	95% Lower	95% Upper	t-value df = 46	2-sided P-value	VIF
Intercept	-1.5194	0.3252	-2.1740	-0.8648	-4.67	0.0000	25.777
Control Rate	-0.0984	0.4753	-1.0551	0.8582	-0.21	0.8368	1.112
Age	0.0330	0.0058	0.0214	0.0446	5.73	0.0000	1.112

Statistics for Model 1

Test of the model: Simultaneous test that all coefficients (excluding intercept) are zero

F = 18.67, df = 2, 46, p = 0.0000

Goodness of fit: Test that unexplained variance is zero

Tau² = 0.0378, Tau = 0.1944, I² = 0.00%, Q = 43.31, df = 46, p = 0.5857

Comparison of Model 1 with the null model

Total between-study variance (intercept only)

Tau² = 0.1347, Tau = 0.3671, I² = 54.64%, Q = 105.82, df = 48, p = 0.0000

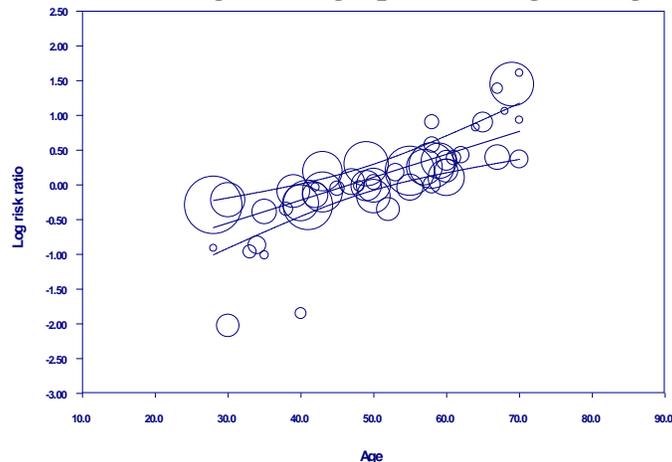
Proportion of total between-study variance explained by Model 1

R² analog = 0.72

Number of studies in the analysis 49

Based on the output above, we have evidence of sample mean age modifying the treatment effect observed across studies such that older age is associated with greater risk. We also have evidence that after adjusting for differences in study age, the influence of control-rate on the treatment effect was no longer significant. Further, we have evidence of a model that results in greater fit to these data compared with an empty (intercept only model). In this case of meta-regression, we also have evidence of reduced heterogeneity (*Q* from 105.82 to 43.31 and no longer significant) and reduced inconsistency across trials (*I*² from 54.64% to 0.00%). Hence, we have evidence of reduced heterogeneity and also a specific study-level factor associated with a gradient of risk across studies (mean study age adjust for control-rate vs. log risk ratio presented in the figure below) (**Figure 4.20**).

Figure 4.20. Multivariate meta-regression graph (showing one significant relationship)



Multivariate meta-analysis

There are both inherent benefits and disadvantages of using meta-analysis as a tool to examine multiple outcomes simultaneously (i.e. multivariate meta-analysis).¹³⁵⁻¹³⁷ One of the advantages of multivariate meta-analysis is being able to incorporate multiple outcomes into one model as opposed to the conduct of multiple univariate meta-analyses wherein the outcomes are handled statistically as being independent.¹³⁷ Another advantage of multivariate meta-analysis is being able to gain insight into relationships between study outcomes.^{137, 138} A final advantage of multivariate meta-analysis is that different clinical conclusions may be made compared with univariate meta-analysis.¹³⁷ In that case, it can be considered easier to present results from a single multivariate meta-analysis compares with the results from different analyses that may make different assumptions.

Some of the major potential issues involved with the joint modeling of multiple outcomes in meta-analysis (reviewed by Jackson and colleagues)¹³⁷ include the disconnect between how outcomes are handled within each trial (typically in a univariate vs. multivariate fashion) vs a multivariate meta-analysis, estimation difficulties particularly around correlations between outcomes (seldom reported; see Bland¹³⁹ for additional commentary), overcoming assumptions of normally- distributed random effects with joint outcomes (difficult to justify with joint distributions), marginal model improvement in the multivariate vs. univariate case (often not sufficient trade off in effort), and issue of amplification of publication bias (e.g. secondary outcomes are not published are frequently).¹³⁷ Another issue is the appropriate quantification of heterogeneity in multivariate meta-analysis; but, there are newer alternatives that seem to make this less of a potential limitation including but not limited to the multivariate H^2 statistic (the ratio of a generalization of Q and its degrees of freedom, with an accompanying generalization of I^2 (I_H^2)).¹⁴⁰ Finally, limitations to existing software for broad implementation and access to multivariate meta-analysis was a long-standing barrier to this approach. With currently available add-on or base statistical packages, multivariate meta-analysis is able to be performed more readily,¹³⁷ and emerging approaches to meta-analyses are available to be integrated into standard statistical output.¹⁴¹

Overall, multivariate meta-analysis approaches may not be accessible to stakeholders involved with systematic reviews,¹³⁹ and are dependent of sophisticated model selection and estimator selection procedures. Hence, our overall recommendation is the multivariate meta-regression only be performed by statisticians with particular experience and expertise in this approach as well as a strong grasp in how to communicate the findings of meta-regression to broad audiences including the lay public.

Dose-response meta-analysis

Considering different exposure or treatment levels has been a longstanding consideration in meta-analyses involving binary outcomes,^{142, 143} and new methods have been developed to extend this approach to differences in means.¹⁴⁴ Meta-regression is commonly employed to test the relationship between exposure or treatment level and the intervention effect (i.e. dose-response). The best case scenario for testing dose responses using meta-regression are when

there are several trials that used each dosing of the intervention vs. control. That way, subgroup analysis can be performed to provide evidence of effect similarity within groups of study by dose in addition to a gradient of treatment effects across groups.¹⁵ Although incorporating study-level average dose can be considered it should only be conducted in circumstances where there was limited-to-no variation in dosing within intervention arms of the studies included. Otherwise, the common problem of ecological bias in meta-regression may lead to biased results.¹³³ In the case of trials involving differences in means, dose-response models are estimated within each study in a first stage and an overall curve is obtained by pooling study-specific dose-response coefficients in a second stage.¹⁴⁴ A key benefit to this emerging approach to differences in means is modeling non-linear dose-response curves in unspecified shapes (including the cubic spline described in the derivation study).¹⁴⁴ Considering the inherent low statistical power associated with meta-regression in general, results of dose-response meta-regression should generally not be used to indicate that a dose response does not exist.¹⁵

4.8 Major Recommendations Regarding Heterogeneity

- Statistical heterogeneity should be expected, quantified and sufficiently addressed in all meta-analyses.
- Multiple metrics of heterogeneity and inconsistency should be generated to make substantive and cumulative interpretations (i.e. Q , Q - df , τ^2 , τ and I^2 along with their respective confidence intervals when possible).
- A non-significant Q should not be interpreted as the absence of heterogeneity.
- In addition to forest and funnel plots, normal probability and Baujat plots may help identify studies contributing most to heterogeneity.
- Random-effects meta-regressive techniques should be used only under full consideration of low power associated with limited studies (i.e. <10 studies per characteristic modelled), the potential for ecological bias, and should only be applied to specified, scientifically defensible hypotheses.
- Exploring well-established subgroup using meta-regression (with across study effect quantification) is favored over analyses within subgroups.
- Multivariate meta-regression should only be performed by statisticians with particular experience and expertise in this approach as well as a strong grasp in how to communicate the findings of meta-regression to broad audiences including the lay public.

Chapter V: Network Meta-Analysis (Mixed treatment comparisons/indirect comparisons)

5.1 Rationale and Definition

The comparative effectiveness agenda and focus on patient-important outcome have driven researchers to provide stakeholders with head to head comparative estimates. However; head to head trials are uncommon. The majority of trials compare active agents to placebo. Industry has minimal incentive to compare active agents; which has left patients and clinicians unable to compare the available treatment options with sufficient certainty.

Therefore, a rationale has emerged to compare agents indirectly. If we know that intervention A is better than B by a certain amount, and we know how B compares to C; we can indirectly infer the magnitude of effect comparing A to C. Occasionally, a very limited number of head to head trials would be available (i.e., there may be a small number of trials directly comparing A to C). Such trials will likely produce imprecise estimates due to the small sample size and number of events. In this case, the indirect comparisons of A to C can be pooled with the direct comparisons, to produce what is commonly called a network meta-analysis estimate. The rationale for producing such an aggregate estimate is to increase precision, and to utilize all the available evidence for decision making.

Frequently, more than two active interventions are available and stakeholders want to compare (rank) numerous interventions, creating a network of interventions with comparisons accounting for all the permutations of pairings within the network.

5.2 Assumptions

There are three key assumptions required for network meta-analysis to be valid:

I. Homogeneity of direct evidence

When important heterogeneity (unexplained differences in treatment effect) across trials is noted, confidence in a pooled estimate decreases.¹⁴⁵ This is true for any meta-analysis. In a NMA, direct evidence (within each pairwise comparison) should be sufficiently homogeneous. This can be evaluated using the standard methods for evaluating heterogeneity (I^2 statistic, Cochran Q test, and visual inspection of forest plots for consistency of point estimates from individual trials and overlap of confidence intervals).

II. Transitivity, similarity or exchangeability

Patients enrolled in trials of different comparisons in a network need to be sufficiently similar in terms of the distribution of effect modifiers. In other words, patients should be similar to the extent that it is plausible that they were equally likely to have received any of the treatments in the network.

Transitivity cannot be assessed quantitatively. However, this can be evaluated conceptually. Researchers need to identify important effect modifiers in the network and assess whether differences reported by studies are large enough to affect the validity of the transitivity assumption.

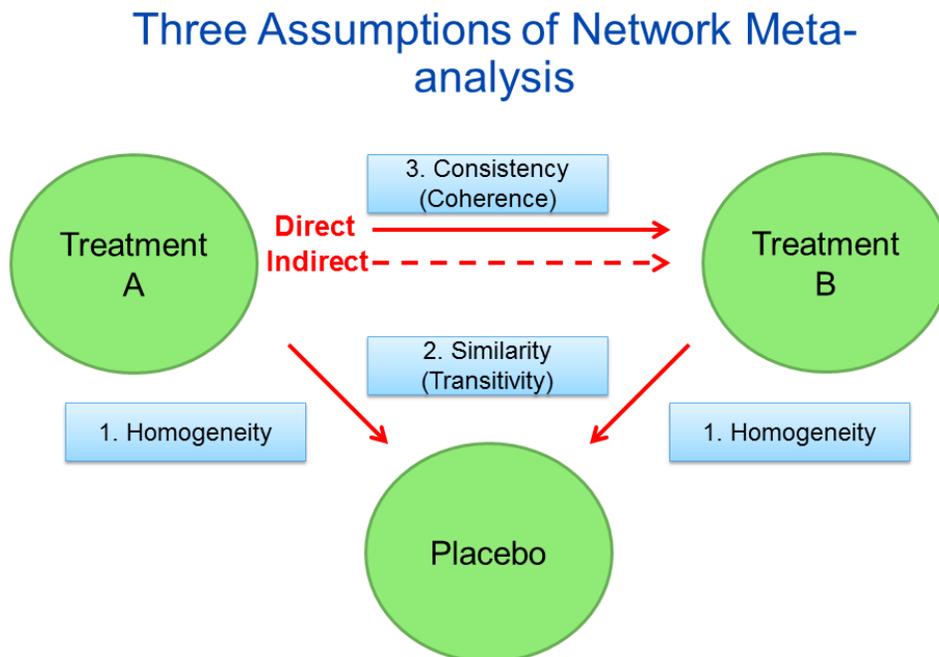
III. Consistency or Coherence (between Direct and Indirect Evidence)

Comparing direct and indirect estimates in closed loops in a network demonstrates whether the network is consistent (also called coherent). Important differences between direct and indirect evidence may invalidate combining them in a pooled NMA estimate.

Consistency refers to the agreement between indirect and direct comparison for the same treatment comparison. If a pooled effect size for a direct comparison equals the pooled effect size from indirect comparison, we say the network is consistent; otherwise, the network is inconsistent or incoherent.^{146, 147} Multiple causes have been proposed for inconsistency, such as differences in patients, treatments, settings, timing, and other factors between direct comparison studies and to studies from which the indirect comparison was imputed; and differences in the risk of bias across the network.

Statistical models have been developed to assume consistency in the network (consistency models) or account for inconsistency between direct and indirect comparison (inconsistency models). Consistency is a key assumption/prerequisite for a valid network meta-analysis and should be always evaluated. If there is a substantial inconsistency between direct and indirect evidence, we should not conduct network meta-analysis. Fortunately, inconsistency can be evaluated statistically. **Figure 5.1** depicts the three assumptions required for a NMA:

Figure 5.1. The three assumptions of network meta-analysis



5.3 Statistical Approaches

Overview

The simplest approach of an indirect comparison is to qualitatively compare the point estimates and the overlap of confidence intervals from two direct comparisons that use a common comparator. Two treatments are likely to have comparable effectiveness if their direct effects relative to a common comparator (e.g. placebo) have the same direction and magnitude, and if there is considerable overlap in their confidence intervals. Under this situation, the qualitative indirect comparison is useful by saving the resources of going through formal testing, and more informative than simply stating that there is no available direct evidence. Such qualitative comparisons have to be done cautiously because the degree of overlap of confidence intervals is not a reliable substitute for formal testing.

In most cases, however, a more explicit and formal approach of indirect comparisons might be preferable. Formal testing methods are more reliable than qualitative assessments because they adjust the comparison of the interventions by the results of their direct comparison with a common control group at least partially using the strength of the original RCTs.¹⁴⁸ This preserves the advantages of randomization of the component trials.

Many statistical models for network meta-analysis have been developed and applied in the literature. These models range from simple indirect comparisons to more complex mixed effects and hierarchical models, developed in both Bayesian and Frequentist frameworks using both contrast level and arm level data. We also distinguish between Arm-based vs contrast-based models.

Simple Indirect Comparisons

Simple indirect comparisons apply when there is no closed loop in the evidence network. At least three statistical methods are available to conduct simple indirect comparisons: 1) logistic regression, 2) random effects meta-regression, and 3) an adjusted indirect comparison method proposed by Bucher et al.¹⁴⁹ When there are only two sets of trials, say, A vs. C and B vs. C, Bucher's method should be enough to get the indirect estimate of A vs. B.

Under ideal circumstances (i.e. no differences in prognostic factors existed among included studies), all three methods result in unbiased estimates of direct effects.¹⁵⁰ However, logistic regression uses arm level dichotomous outcomes data and is limited to odds ratios as the measure of effect. By contrast, meta-regression and adjusted indirect comparisons typically use contrast level data and can be extended to risk ratios, risk differences, mean difference and any other effect measures. Meta-regression (as implemented in Stata, *metareg*) and adjusted indirect comparisons are most convenient to compare trials with two treatment arms. A simulation study supports the use of random effects for either of these approaches.¹⁵⁰

Mixed effects and hierarchical models

More complex statistical models are required for more complex networks with closed loops where a treatment effect could be informed by both direct and indirect evidence. These models typically assume random treatment effects and take the complex data structure into account, and may be broadly categorized as mixed effects, or hierarchical models.

Frequentist approach

Lumley proposed the term “network meta-analysis” and the first network meta-analysis model in the frequentist framework, a random-effects inconsistency model by incorporating sampling variability, heterogeneity, and inconsistency.¹⁵¹ The inconsistency follows a common random-effects distribution with mean of 0. It can use arm-level and contrast-level data and can be easily implemented in statistical software, including R’s lme package. However, studies included in the meta-analysis cannot have multiple arms.

Further development of network meta-analysis models in the frequentist framework addressed how to handle multi-armed trials as well as was new methods of assessing inconsistency (e.g., Salanti et al., 2008; White et al. 2012, Higgins et al. 2012, Greco et al. 2015).¹⁵²⁻¹⁵⁵ Salanti et al. provided a general formulation of network meta-analysis model with either contrast-based data or arm-based data, and defined the inconsistency in a standard way as the difference between ‘direct’ evidence and ‘indirect’ evidence.¹⁵² In contrast, White et al. (2012) and Higgins et al. (2012) proposed to use a treatment-by design interaction to evaluate inconsistency of evidence and developed consistency and inconsistency models based on contrast-based multivariate random-effects meta-regression.^{153, 154} These models could be implemented using *network*, a suite of commands in Stata with input data being either arm-level or contrast level.

Bayesian approach

Lu and Ades (2004, 2012) proposed the first Bayesian network meta-analysis model for multi-arm studies that included both direct and indirect evidence.^{156 157} The treatment effects were represented by basic parameters and functional parameters, and the evidence inconsistency was defined as a function between a functional parameter and at least two basic parameters. The Bayesian model has been extended to incorporate study-level covariates in an attempt to explain between-study heterogeneity and reduce inconsistency,¹⁵⁸ to allow for repeated measurements of a continuous endpoint that varies over time,⁸² or to appraise novelty effects.¹⁵⁹ Additionally, Dias et al. (2013) set out a generalized linear model framework for the synthesis of data from randomized controlled trials, which could be applied to binary outcomes, continuous outcomes, rate models, competing risks, or ordered category outcomes.⁸⁰ Very commonly, a vague (flat) prior is chosen for the treatment effect and heterogeneity parameters in Bayesian network meta-analysis. An alternative would be to derive the prior from the predictive distributions for the degree of heterogeneity as expected in various settings depending on the outcomes assessed and comparisons made.¹⁶⁰

In the network meta-analysis framework, frequentist and Bayesian approaches often provide similar results because of the common practice to use non-informative priors in the Bayesian

analysis.^{156, 161, 162} Frequentist approaches, when implemented in a statistical package, are easily applied in real-life data analysis. Bayesian approaches are highly adaptable to complex evidence structures and provide a very flexible modeling framework, but would need a better understanding of the model specification and specialized programming skills.

Arm-based vs contrast-based models

It is important to differentiate arm-based/contrast-based models from arm-level/contrast-level data. Arm-level and contrast-level data describe how outcomes are reported in the original studies. Arm-level data list absolute effect size per study arm (e.g. number of withdrawals from a trial per group); while contrast-level data show the difference of outcomes between arms, aka, relative effect size or average treatment effect (e.g. odds ratio of withdrawals).

Contrast-based models resemble the traditional approaches used in direct meta-analyses that relative effect sizes and associated variance are first estimated and then pooled to get estimates for treatment comparison. Contrast-based models preserve randomization and, largely, alleviate risk of observed and unobserved imbalance between arms within a study. They use relative effect sizes and reduce variability of outcomes across studies. Contrast-based models are the dominate approach used in direct meta-analyses and network meta-analyses in current practice.

Arm-based models, instead of pooling relative effect sizes in contrast-based models, directly combine observed absolute effect size in individual arms across studies. Although arm-based models break randomization and suffer higher risk of bias, multiple models have been proposed, especially under Bayesian framework.¹⁶³⁻¹⁶⁷ Arm-based models allow 1) estimation of absolute measures (e.g. treatment-specific event rate) which may be important to patients and clinicians; 2) modelling outcome data in each arm directly (e.g. proportion of events using binominal distribution) and avoiding approximation of normality of relative effect size (e.g. log transformation of odds ratio); and 3) bypassing explicit modelling of correlations among multiple arms within a study. However, the validity of arm-based methods is under debate. Many experts argue arm-based models should be avoided.^{153, 168, 169}

Assessing consistency

Network meta-analysis generates results for all pairwise comparisons; however, consistency can only be evaluated when at least one closed loop exists in the network. In other words, the network must have at least one treatment comparison with direct evidence. Many statistical methods are available to assess consistency.^{149, 157, 170-176}

These methods can generally be categorized into two types: an overall consistency measure for the whole network, and loop based approach in which direct and indirect estimates are compared. In the following section, we will focus a few widely used methods in the literature.

- i. Single Measure Network Coherency: Approaches using a single measure that represents coherence for the whole network. Lumley assumes that, for each treatment

- comparison (with or without direct evidence), there is a different inconsistency factor; and the inconsistency factor varies for all treatment comparisons and follows a common random-effects distribution. The variance of the differences, ω , also called incoherence, measures the overall inconsistency of the network.¹⁷⁰ ω above 0.25 suggests substantial inconsistency and network meta-analysis is inappropriate.¹⁷⁷
- ii. Global Wald Test: Another approach is to use global Wald test, where inconsistency factor follows a chi2 distribution under consistency assumption.¹⁵³ p value <0.10 should be used to show statistical significance, indicating that the model is inconsistent.
 - iii. Loop based approach: This approach involves comparing direct and indirect estimates for each comparison. Although a single inconsistency measure is easy to calculate and interpret, it conceals important sources of inconsistency (if multiple loops exist) in the network. Comparing direct and indirect estimates can be done in various ways:
 - a. *Z-test*: A simple z-test can be used to compare the difference of pooled effect sizes between direct and indirect comparison.¹⁴⁹ This test can be easily applied in any statistical software or Microsoft Excel. Benefits include that it is a simple and easy to apply method and can identify specific loops with large inconsistency. Limitations include the need for multiple correlated tests.
 - b. *Node-splitting (side-splitting)*: A treatment comparison is also known as “node” or “side” in the network. Dias et al. suggested that each node can be assessed by comparing difference of the pooled estimate from direct evidence to the pooled estimate without direct evidence.¹⁷¹ Node-splitting can be implemented using Stata “network sidesplit” command or R “gemtc” package.
 - c. *Inconsistency plot*: Several graphical tools have been developed. One is inconsistency plot developed by Chaimani et al.¹⁷² Similar to forest plot, inconsistency plot graphically presents inconsistency factor (absolute difference between direct and indirect estimates) and related confidence interval for all triangular and quadratic loops in the network. Stata “ifplot” command can be used.

It is important to understand the limitations of these methods. A statistical insignificance does not prove consistency in the network. Similar to Cochran's Q test used for heterogeneity testing, statistical methods are under powered due to limited number of studies in direct comparisons. Random effect models, by incorporating heterogeneity, increase confidence intervals of the pooled effect size. This can hide important differences between direct and indirect evidence and further reduce the possibility of detecting inconsistency.

5.4 Considerations of model choice and recommendations

Consideration of Indirect Evidence

Empirical explorations suggest that direct and indirect comparisons often agree,^{151, 178-182} but with notable exceptions.¹⁸³ In principle, the validity of combining the direct and indirect evidence relies on the transitivity assumption, that is, the invariance of treatment effects across study populations. However, in practice, trials can vary in numerous ways including population characteristics, interventions and cointerventions, length of followup, loss to followup, study quality, etc. Given the limited information in many publications and the inclusion of multiple treatments, the validity of combining the direct and indirect evidence is often unverifiable. The statistical methods to evaluate inconsistency are generally have low power, and confounded by the presence of statistical heterogeneity. They often failed to detect inconsistency in the evidence network.

Moreover, network meta-analysis by combining the direct and indirect evidence, like all other meta-analyses, essentially constitutes an observational study, and residual confounding can always be present. Systematic differences in characteristics among trials in a network can bias network meta-analysis results. In addition, all other considerations for meta-analyses, such as choice of effect measures or heterogeneity, also apply to network meta-analysis. Therefore, in general, investigators should compare competing interventions based on direct evidence from head-to-head RCTs whenever possible. When head-to-head RCT data are sparse or unavailable but indirect evidence is sufficient, investigators could consider incorporate indirect evidence and network meta-analysis as an additional analytical tool. If the investigators choose to ignore indirect evidence, they should explain why.

Choice of models

Although the development of network meta-analysis models exploded in the last ten years, there has been no systematic evaluation of their comparative performance, and the validity of the model assumptions in practice is generally hard to verify.

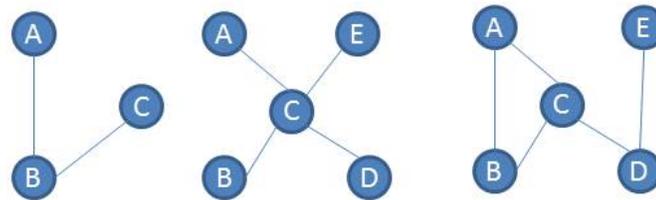
Investigators may choose a frequentist or Bayesian model based on the research team expertise, the complexity of the evidence network and research question. However, whichever method the investigators choose, they should assess the consistency of direct and indirect evidence, and invariance of treatment effects across studies and appropriateness of the chosen method on a case-by-case basis, paying special attention to comparability across different sets of trials. Investigators should explicitly state assumptions underlying indirect comparisons and conduct sensitivity analysis to check those assumptions. If the results are not robust, findings from indirect comparisons should be considered inconclusive. Interpretation of findings should explicitly address these limitations. Investigators should also note that simple adjusted indirect comparisons are generally underpowered, needing four times as many equally sized studies to achieve the same power as direct comparisons, and frequently lead to indeterminate results with wide confidence intervals.^{179, 181}

When the evidence of a network of interventions is consistent, investigators could combine direct and indirect evidence using network meta-analysis models. Conversely, they should refrain from

combining multiple sources of evidence from an incoherent network where there are substantial differences between direct and indirect evidence. Investigators should make efforts to explain the differences between direct and indirect evidence based upon study characteristics, though little guidance and consensus exists on how to interpret the results.

Lastly, the network geometry can also affect the choice of analysis method as demonstrated in **Figure 5.2**.

Figure 5.2. Impact of network geometry on analysis method.



Statistical methods	Simple indirect comparison	Star network	Network with at least one closed loop
Qualitative assessment	Yes		
Adjusted indirect comparison, random-effects meta regression, logistic regression	Yes	Yes	
Mixed-effects linear regression			Yes
Multivariate random-effects meta-regression		Yes	Yes

5.5 Inference from network meta-analysis

Stakeholders (users of evidence) require a rating of the strength of evidence assigned to estimates of comparative effectiveness. The strength of evidence demonstrates how much certainty we should have in the estimates.

The general framework for assessing the strength of evidence used by the EPC program is described elsewhere. However; for network meta-analysis, guidance is evolving and may require some additional computations; therefore, we briefly discuss the possible approaches. We also discuss inference from rankings and probabilities commonly presented with a network meta-analysis.

Approaches for Rating the Strength of Evidence:

The original GRADE guidance was simple and involved rating down all evidence derived from indirect comparisons (or NMA with mostly indirect evidence) for indirectness. Therefore, following this original GRADE guidance, evidence derived from most NMAs would be rated to have moderate strength at best.¹⁸⁴

Salanti et al. evaluated the transitivity assumption and network inconsistency under the indirectness and inconsistency domains of GRADE; respectively. They judged the risk of bias based on a ‘contribution matrix’ which gives the percentage contribution of each direct estimate to each network meta-analysis estimate.¹⁸⁵ A final global judgment of the strength of evidence is made for the overall rankings in a network.

More recently, GRADE published a new approach that is based on evaluating the strength of evidence for each comparison separately rather than making a judgment on the whole network.¹⁸⁶ The rationale for not making such an overarching judgment is that the strength of evidence (certainty in the estimates) is expected to be different for different comparisons. The approach requires presenting the three estimates for each comparison (direct, indirect and network estimates), then rating the strength of evidence separately for each one. In summary, researchers conducting NMA should present their best judgment on the strength of evidence to facilitate decision-making. Innovations and newer methodology are constantly evolving in this area.

Interpreting Ranking Probabilities and Clinical Importance of Results

Network meta-analysis results are commonly presented as probabilities of being most effective and as rankings. Such presentations should be interpreted with caution since they can be quite misleading.

Whether results were presented as probabilities, rankings or surface under the cumulative ranking curve (SUCRA), three pitfalls should be recognized:

- I. Such estimates are usually very imprecise. An empirical evaluation of 58 NMAs showed that the median width of the 95% CIs of the SUCRA was 65% (first to third quartile, 38% to 80%). In 28% of networks, there was a 50% or greater probability that the best-ranked treatment was actually not the best. No evidence showed a difference between the best-ranked intervention and the second and third best-ranked interventions in 90% and 71% of comparisons, respectively.
- II. When rankings suggest superiority of an agent over others, the absolute difference between this intervention and other active agents could be trivial. Converting the relative effect to an absolute effect is often needed to present results that are meaningful to clinical practice and relevant to decision making.¹⁸⁷ Such results can be presented for patient groups with varying baseline risks. The source of baseline risk can be obtained from observational studies judged to be most representative of the population of interest, from the average baseline risk of the control arms of the randomized trials included in meta-analysis, or from a risk stratification tool if one is known and commonly used in practice.¹⁸⁸

- III. Rankings hide the fact that each comparison may have its own risk of bias, limitations and strength of evidence.

5.6 Presentation and Reporting

Methodological evaluation of published network meta-analyses demonstrated great heterogeneity in reporting and numerous deficiencies. Commonly, network meta-analyses demonstrated unclear understanding of underlying assumptions, inappropriate search and selection of relevant trials, use of inappropriate or flawed methods, lack of objective and validated methods to assess or improve trial similarity, and inadequate comparison or inappropriate combination of direct and indirect evidence.¹⁸⁹⁻¹⁹¹

Such deficiencies necessitated the extension of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) statement that attempted to improve the reporting of systematic reviews incorporating network meta-analyses.¹⁹² The extension suggests reporting several items that are categorized as belonging to the title, abstract, introduction, methods, results, conclusions and funding sections of a manuscript. However, these items can also be viewed as those addressing the systematic review process itself to make it explicit and reproducible (search details, study selection methods, etc.); or items that relate to analysis (describe the model used, effect measure selected, sensitivity analysis, etc.); or items that relate to inference (strength of evidence, limitations and conclusions). All these recommendations of reporting are essential and should be considered minimal criteria. Many of these elements were unfortunately often neglected in published network meta-analyses. Nevertheless, additional information is clearly needed to assist stakeholders and decision makers appraise the evidence and judge its strengths, determine the applicability of evidence, and act on it.

Summary

In summation, we suggest that network meta-analyses present the following information:

- I. Description of the process of the systematic review: Similar to any other systematic review, the report should include the rationale for the review, research question, eligibility criteria (patient, interventions, comparisons, outcomes, included studies design and duration), search strategies, and details of study selection process. EPC guidance is available to help authors appropriately report such elements (cite).
- II. Description of analysis plan: the report should include the general approach used (e.g., Bayesian, Frequentist), the model (random vs fixed; consistency vs inconsistency), rationale for model choice, preplanned and post hoc subgroup and sensitivity analyses, software and syntax/command used, choice of priors for Bayesian analyses and description of how rankings were generated. A graphical presentation of network structure and geometry is also strongly encouraged to show the amount of direct evidence and help evidence users understand the nature of available evidence.

III.

Description of the results:

- a. *Data from individual studies:* Description of the characteristics of each study is needed to facilitate application of the results and judging indirectness of evidence. Description of risk of bias of each study and overall across studies is needed to facilitate judging the strength of evidence. Description of the effect size of each study is needed to permit evaluation of heterogeneity and to allow reproducibility of analysis.
- b. *Pooled results:* a relative association measure or other type of effect size (WMD, SMD, etc.) should be displayed for each pairwise comparison to allow comparative effectiveness inferences. Usually a tabular format (example) with interventions listed as columns and rows facilitate such presentation although other graphical presentations are also possible and innovations in presenting these complicated results are in progress. Rankings can be presented as probabilities or as area under the curve.
- c. *Robustness of results:* It is highly important to be explicit in describing how the choice of model or the choice of prior distributions affects conclusions. Assumptions made during the process should also be verified to determine how it impacts conclusions (e.g., including borderline eligible studies, using different correlation coefficients to impute measures of variability, etc.). A key methodological issue in network meta-analysis is to demonstrate the extent of consistency between direct and indirect comparisons to allow judgment of network coherence (consistency).

IV.

Inferences:

- a. *Reporting of the strength of evidence:* a mandatory step without which stakeholders cannot consume and apply evidence. Evidence profiles (i.e., tables that show the effect size in relative and absolute terms and the judgments made regarding the strength of evidence) can be constructed for key comparisons (as identified by stakeholders) or for all comparisons (if relevant).
- b. *Clinical importance/magnitude of effect:* network meta-analysis should not solely depend on ranking probabilities but rather present effect sizes for all comparisons. Presenting an absolute effect will be most helpful for decision-making and clinical practice.

Chapter VI: Stability and Sensitivity Analyses in Evidence Synthesis

Abstract

We present an overview of analyses undertaken to evaluate the *robustness* (or fragility) of systematic review and meta-analysis findings to changes in assumptions or methodological decision about the data and methods used for evidence synthesis. We discuss *stability analyses*, which evaluate the robustness of conclusions to different methodological choices. Stability analyses pertain to the selection of data to be synthesized (e.g., leave-one-out analyses); the specification of models for evidence synthesis (e.g., the use of a discrete within-study likelihood or a normal approximation); and the estimators and techniques for estimating model parameters (e.g., the use different estimators for quantifying between-study heterogeneity under the same statistical model). We also discuss *sensitivity analyses*, which evaluate the impact of different assumptions about the parameters of a particular statistical model on conclusions. In the context of systematic reviews, sensitivity analyses are used for handling missing study-level data, the selective dissemination of study results, or the estimation of unidentifiable or weakly identifiable model parameters. We identify challenges in the interpretation of stability and sensitivity analyses in evidence synthesis and offer recommendations for their use in applied work.

6.1 Background

Synthesizing evidence to answer well-specified research questions requires investigators to make innumerable choices among different sources of evidence, approaches for extracting data, and methods for synthesis (including any statistical modeling). In any particular project, the best choices are not always obvious and investigators often use multiple alternative approaches to probe the same body of evidence (e.g., repeating analyses after excluding a subset of studies, using different estimators to estimate the parameters of the same model, etc.). In addition, synthesis of information from multiple sources often requires the handling of missing data and the specification of models with parameters that cannot be identified from the available data; this is particularly true of evidence synthesis efforts that attempt to account for missing data (e.g., lack of data needed to estimate sampling variances), selective dissemination of study results (e.g., selective reporting or publication bias), or address biases in individual studies (e.g., confounding or selection bias).

Systematic reviewers often undertake a number of activities which aim to assess whether their conclusions are *robust* (i.e., not fragile) to different methodological decisions or assumptions about missing data and biases. Attempts to define robustness, to develop robust methods, and to understand the epistemological status of robustness have a long history in statistics, econometrics, mathematical modeling, and philosophy of science.^{193, 194} Importantly, robustness is determined as a joint property of the research question and target of inference, data, assumptions, and models used in evidence synthesis.

In this paper, we provide an overview of analyses that are useful for assessing robustness in applied evidence synthesis projects. Though we find that such analyses are often useful, we discuss challenges that arise in their interpretation and identify some caveats in popular approaches. We adopt a pragmatic perspective and provide recommendations for practitioners without delving into epistemological issues (see for e.g., Wimsatt, (1981) Woodward, (2006) and Orzack).^{194, 195} Many of our examples use as the target of inference the construction of a response surface for the mean effect size conditional on a rich set of covariates describing the interventions, study designs, and included populations using a large collection of relevant studies.¹⁹⁶ We often assume that the investigators wish to estimate this target by combining the data and likelihood with the investigators prior beliefs in a Bayesian analysis; we view this as an important approach to evidence synthesis (and the most tenable form of meta-analysis when data are obtained through published reports of completed studies). Nevertheless, our points are general and apply to alternative approaches for synthesis, ranging from non-quantitative approaches to generalized evidence synthesis.

6.2. Assessing the Robustness of Evidence Synthesis

We find it useful to organize the various activities that aim to assess the robustness of evidence synthesis results into stability and sensitivity analyses.

Stability analyses evaluate the robustness of conclusions to distinct methodological choices. These analyses pertain to the selection of data to be synthesized (e.g., repeating the analyses after excluding studies with outlying effect estimates); the specification of models for evidence synthesis (e.g., the decision to include certain covariates in a meta-regression model); the choice of estimators for estimating model parameters (e.g., the use of alternative estimators for the heterogeneity parameter in a particular statistical model); or aspects of estimation and computing techniques (e.g., the choice of starting values for iterative procedures or the practical aspects of using MCMC techniques to fit Bayesian models).

Sensitivity analyses evaluate the implications of assumptions about unidentifiable model parameters on results, conditional on some class of statistical models. In the context of evidence synthesis, sensitivity analyses can be used for handling missing study-level data (e.g., in before-after studies or studies of net-change the correlation between the pre-treatment and post-treatment means are almost never reported and need to be imputed), selective dissemination of study results (e.g., selective reporting or publication bias), or the estimation of model parameters when studies are deemed susceptible to biases (e.g., due to confounding or selection bias).

In **Table 6.1** we list examples for sensitivity and stability analyses at various steps of the systematic review process. In the following section we discuss some of these examples in more detail to give a more complete account of stability and sensitivity analysis in evidence synthesis. In Sections 3 and 4 we discuss the conduct of stability and sensitivity analyses in more detail and in Section 5 we address issues of interpretation.

Recommendations for stability and sensitivity analyses in systematic reviews:

We propose the following recommendations for stability and sensitivity analyses in systematic reviews:

1. When planning the review, identify stability analyses to be performed by considering the major methodologic decisions for which reasonable analysts might disagree on the preferred approach.
2. When planning the review, consider the use of sensitivity analyses for addressing the impact of missing data, publication bias, and the risk of bias of individual studies on study results. Select sensitivity analyses to be performed by considering the research question and target of inference, the available data, and background knowledge about the substantive area and the relevant methods.
3. Describe all planned stability and sensitivity analyses, their rationale, and the approach for interpreting their results in the study protocol. For sensitivity analyses that are driven by unidentifiable assumptions justify assumptions based on background knowledge.
4. Describe any additional stability and sensitivity analyses undertaken during the review process that were not pre-planned. Explain why these analyses were undertaken and acknowledge the extent to which they were driven by the examination of the data and expert input.
5. Report the methods and findings of stability and sensitivity analyses in full to allow users of the systematic review to form their own conclusions regarding the interpretation of results.
6. Use substantive and methodological knowledge to interpret the findings of stability and sensitivity analyses, including both cases where results were robust and cases where results depended on the choice of methods or assumptions.

6.3 Conducting Stability Analyses

“A complex analysis involves numerous implementation or analytical decisions. The audience for such an analysis typically wishes to be assured that conclusions are not artifacts of such decisions, but rather are stable over analyses that differ in apparently innocuous ways.”¹⁹⁷ It is useful to consider stability analyses as they pertain to different stages of the systematic review process: the search and selection of studies; the extraction of data from eligible studies; choice of models for evidence synthesis; modes of statistical inference; estimation approaches for the parameters given a model and mode of inference.

Search and selection of studies: Stability analyses aim to identify whether the provenance of the studies or other characteristics that ostensibly should not influence study results have any impact on results. For example, reviewers are often interested in seeing whether indexing of studies in particular databases (e.g., PubMed vs. non-PubMed),¹⁹⁸ the language of the report from which data are extracted (e.g., English vs. non-English language), study sample size (e.g., studies above vs. below the median sample size) correlates with the estimated effect size. Additional examples include dropping one study at a time to check whether a study is particularly influential, excluding studies with extreme (“aberrant”) effects,¹⁹⁹ or all subsets meta-analyses [cite Olkin RSM]. Plots are usefully for summarizing such analyses.¹²⁵ For example, the exclusion sensitivity plot or similar graphs,^{200, 201} can reveal studies that have a particularly large influence on the result of a meta-analysis. Graphs are also useful for studying the distribution of all subsets meta-analysis results.

Extraction of data from eligible studies: It is not uncommon that a single study will report multiple estimates for the same parameter; when multiple estimates for the same parameter are available reviewers might wish to examine whether the choice of a particular type of estimate has an impact on results. For example, crude (unadjusted) mean difference estimates for the effect of a binary treatment may be reported along with multivariably adjusted results (including adjustments with different sets of covariates). For example, reviewers might examine a meta-analysis using unadjusted study-level results, as well as maximally adjusted results, or results adjusted for a minimum (predetermined) set of covariates. Of note, such stability analyses are fairly straightforward for meta-analyses of collapsible effect measures (e.g., mean differences) from randomized trials, but are more complicated for meta-analyses of observational studies (because different adjustment sets imply different assumptions about confounding) and non-collapsible effect measures (e.g., odds ratios or hazard ratios).

Choice of statistical models for evidence synthesis: Often, more than one statistical model can be used to obtain the same estimand. For example, when analyzing discrete outcomes, it is possible to use a discrete (e.g., binomial for binary data) or normal (approximate) within-study likelihood. Performing the analyses both ways might be a useful way to assess whether the normal approximation is adequate (admittedly, an assessment of limited and primarily methodological interest). Of course, in many such cases the preferred model might be chosen on the grounds of theory or simulation studies (e.g., several simulation studies provide strong evidence in favor of using the discrete within-study likelihood when the maximization algorithm converges). As another example of stability analysis for the choice of statistical model, when adopting a Bayesian mode of inference, different priors can be used to represent different beliefs about parameters (in Bayesian analyses the likelihood and prior constitute the model). These stability analyses might produce different results to the extent that the posterior distribution is influenced by the prior. In this case, the beliefs of the user(s) of the review determine which analyses should be considered as better addressing the research question.

Choice of mode of inference: As noted in the Background section, we find Bayesian methods to be a natural choice for evidence synthesis, particularly when studies are identified retrospectively, data are obtained from secondary sources (journal articles, gray literature, etc.), and reviewers cannot influence the design of the studies being synthesized. With the increasing availability of software to perform Bayesian analyses, meta-analyses conducted using both Bayesian and non-Bayesian methods have become common. This sort of stability analysis provides an indirect assessment of the relative amount of information in the prior and data (likelihood). The interpretation of numerical comparisons between Bayesian and non-Bayesian estimates is not straightforward (because parameters are treated differently in the two modes of inference). In addition, an assessment of the relative amount of information in the prior and data can be obtained entirely within the Bayesian analysis (e.g., by prior-posterior comparisons).

Estimation approaches: Given a particular model and mode of inference, it is often the case that multiple estimation approaches are available for the same estimand. Agreement in the estimates obtained with different approaches can provide some reassurance that the data is adequate for drawing conclusions, without the estimators having undue influence. For example, if one adopts a normal-normal random effects model and a frequentist mode of inference, there exist multiple estimators of the between-study variance. Many reasonable candidate estimators

are (asymptotically) equivalent, yet empirical analyses and simulation studies (reviewed in Chapter III) demonstrate that modest discrepancies are not uncommon with the (finite) sample sizes seen in applied reviews (by sample size we mean both the number of studies and the number of participants in each study). It is often considered good practice to repeat the analysis using different estimators; in view of well-known theoretical results, lack of stability across analyses conveys information about the (inadequacy) of the data.

6.4 Conducting Sensitivity Analysis

Within a given analytic model, a **sensitivity analysis** typically examines a continuous family of departures from a critical assumption(s), i.e. changes in parameters, where attention focuses on the relationship between the magnitude of the departure and the magnitude of the change in conclusions.¹⁹⁷ Sensitivity analysis can be conceptualized as an examination of the dependence of results on unidentifiable model parameters. As will be seen below, using appropriately parameterized models, general qualitative statements that might apply to all meta-analyses can be replaced by quantitative statements that are specific to a particular analysis.

In contrast to opportunities for stability analyses (which arise throughout the systematic review process), sensitivity analyses are conditional on particular (classes of) models. We might claim that in all applied evidence synthesis projects, many model parameters are non-identifiable: <DEFINE> unless the investigators are willing to make extremely strong assumptions. For example, most meta-analyses are affected by publication bias and selective reporting. Unless strong assumptions are made about the number of missing studies and the relationship of their results to those of the included studies, the mean effect or the regression of the mean effect on covariates cannot be identified. Thus, even though rarely pursued formally, sensitivity analyses could allow a more realistic interpretation of the results of all applied systematic reviews. We now consider some examples of where the need for sensitivity analysis can arise in meta-analyses using published data. Of note, there are multiple approaches for conducting sensitivity analysis for a given problem, and each approach requires systematic reviewers to make many choices, each of which can be subjected to stability analysis.

Missing data required for quantitative synthesis: Missing data are ubiquitous in systematic reviews and different values for the missing quantities can change meta-analysis conclusions. For example, in meta-analyses of net changes (comparative studies with before-after mean measurements in each of the groups being compared) the within-group means are correlated but an estimate of the correlation is almost never reported and different values for that quantity would result in a different standard error for the net change. A common approach in this case is to perform a sensitivity analysis by “plugging in” different values for the missing correlation estimate to assess impact on results (more formal treatments are possible, e.g., by specifying the likelihood of the complete data, possibly with an informative, background-knowledge-based, prior for the within-arm correlation). Similar issues arise in the analysis of data from cluster randomized trials where the intra-class correlation coefficient is almost never reported.

Handling publication bias and other dissemination biases: Publication bias – the preferential publication of statistically significant or otherwise remarkable results (e.g.,

contradictory) – and other dissemination biases (e.g., selective outcome and analysis reporting) are special types of missing data problems that are believed to be common and cannot be addressed (or even detected) simply by considering the information uncovered by systematic reviewers. For example, it is impossible to know the number of conducted and otherwise eligible but unpublished studies simply by examining the studies actually uncovered. Inference in the presence of publication bias is inherently risky because all methods for handling publication bias, ranging from qualitative summaries to full scale modeling of the publication selection process, rely on unverifiable assumptions. Examples of formal sensitivity analyses for suspected publication and other dissemination biases include the many selection model-based methods (e.g., the likelihood based approach of Copas.²⁰² More commonly used approaches, such as trim-and-fill and failsafe-N, can also be viewed as sensitivity analysis methods for publication bias, albeit ones where the underlying missing data mechanism is less explicitly represented.

Adjusting for possible bias due to deficiencies in study design, planning, or conduct: Most studies, both randomized and non-randomized, have identifiable deficiencies in their design and conduct (e.g., poor allocation concealment in trials; baseline confounding in observational studies; dropout and loss-to followup in all studies, etc.). The different sources of bias can be represented in the statistical model used for evidence synthesis as “bias parameters,” but it is well-understood that these parameters are not identifiable. It follows that sensitivity analysis is the only viable approach for assessing the impact of the bias factors on the results of the evidence synthesis.

6.5 Interpreting Stability and Sensitivity Analyses

Because different types of stability and sensitivity analyses have different goals and require different approaches, it is hard to provide strict rules for their interpretation. The examples we have provided in previous sections should cover many of the most common situations in applied systematic reviews. In this section we try to identify some important general themes; in the next section we provide some actionable recommendations for applied systematic reviews.

Robustness is a joint property of the target of inference, the data, and methods: As in statistical analyses of primary data, the robustness of evidence synthesis depends on the target of inference, the data, and the statistical methods (model and estimation). For example, the robustness of results might vary depending on whether the target of inference is the average effect size, the average effect for studies with a set of particular characteristics, or that the probability that either effect is more extreme than some pre-specified value. Furthermore, when estimating the average effect using an iterative method-of-moments estimation approach, the robustness of the analysis will depend on the number of available studies. Finally, the choice of statistical methods can impact robustness by weakening reliance on particular assumptions. For example, method-of-moments estimators for the normal-normal meta-analysis model are robust to distribution of the random effects as they do not require its specification for consistency. In contrast maximum likelihood estimators might be more sensitive to this assumption.

Robustness and its absence are both interesting: Contrary common belief, both robustness and its lack are informative in systematic reviews. For example, if results appear unstable in leave-one-out analyses can be useful for identifying studies with unique (substantive or methodological) characteristics. Similarly, if results are sensitive to assumptions about missing data or publication bias reviewers' can avoid drawing inappropriately strong conclusions from a body of evidence. Of note, in some cases we are interested in demonstrating that stability analyses have no appreciable impact on results. For example, when choosing different starting values for an iterative algorithm that is known to converge to some limiting value we typically wish to illustrate that diverse starting values lead to the same conclusions. Discovering that different values lead to different results indicates problems (in programming, model specification, etc.) that need to be examined further. In other cases, it is generally expected that different choices will almost always have some impact on the results. For example, the inclusion of different variables in meta-regression models or the use of different prior distributions are generally expected to have some impact on results. Such stability analyses are informative because the patterns of stability or instability of results can aid the examination of the body of evidence.

The problem of forking paths: Even in simple evidence synthesis projects investigators need to make many methodological choices. The number of possible combinations is very large and there is a risk of producing an overwhelming amount of results or engendering selective reporting. The need for stability analysis will be reduced by detailed and unambiguous pre-specification of research methods in the review protocol (informed by substantive and methodological considerations), the reporting and justification of any unanticipated analyses, and the complete presentation of all results

Some stability analyses are more easily interpretable than others: Stability analyses comparing different methodological choices but keeping the target of inference and statistical model the same are fairly straightforward to interpret, while analyses relying on different models or estimating different parameters can be more challenging. The results of leave-one-out analyses or comparisons of different estimators for the same parameter can be interpreted in view of substantive knowledge and statistical theory. Yet, discrepancies between Bayesian vs. non-Bayesian analyses or comparisons of fixed vs. random effects models are more challenging because they entail very different modeling assumptions. In a sense, examining the stability of results across different targets of inference or models and is harder because different targets of inference often represent a fundamental change in the research question and different models can have a large influence on the research questions that can be examined. In contrast, stability analyses holding the model and target of inference constant pertain to the choice of data and the relationship of the data with the statistical model (e.g., the adequacy of the data to estimate the model parameters), but not about the model itself.

Sensitivity analysis is done in view of a particular class of models: Sensitivity analysis examines the impact of different assumptions about unidentifiable parameters of (classes of) models on results. Because, by definition, sensitivity analysis relates to parameters for which the data do not provide adequate information, the approach to sensitivity analysis is to a large extent a matter of style and statistical philosophy. For example, some experts prefer to use sensitivity analysis to obtain bounds for the target of inference under different assumptions about

unidentified parameters without quantifying (eliciting) the relative probability of different values for the unidentified quantities. Others prefer the specification of a distribution for the possible values of the unidentified parameters followed by a propagation of the uncertainty implied by that distribution to the estimation of the target of inference. We see value in both approaches and in letting systematic reviewers make the choice they find most appropriate on a case-by-case basis.

Table 6.1. Opportunities for stability and sensitivity analysis in the systematic review process

<i>Stability analyses</i>	
Data collection and pre-processing, including exploratory data analysis	Perform analysis after excluding studies not published in full
	Methods to detect sensitivity to influential studies or subsets of studies (leave out one, exclusion sensitivity plot)
	Exclusion of edge cases with respect to inclusion/exclusion criteria
	Alternative methods for approximating summary statistics (e.g., approximating the mean using the median in analyses of mean differences)
Specification of the probability model for evidence synthesis	Choice of likelihood function: for example, decide whether to assume asymptotic normality of log odds ratio vs. using a discrete (binomial) within-study likelihood.
	Comparison of alternative effect size indices (odds-ratio, risk difference)
Estimation and testing	Compare alternative variance estimators (DerSimonian-Laird vs. hierarchical Bayes)
	Estimation approach (maximum likelihood estimated, restricted maximum likelihood estimation, MCMC simulation)
<i>Sensitivity analyses</i>	
Impact of assumptions about unidentifiable model parameters on results	In meta-analyses of net change the correlation between the baseline and post treatment mean in
	In meta-analysis of cluster randomized trials the correlation
	When publication bias is

Future Research Suggestions

The following are suggestions for directions in future research generated from each chapter of the manuscript:

Chapter I: Decision to Combine Trials of Treatment Efficacy or Harm

- Key need on the decision to pool: more clear guidance about the minimum number of trials it is likely valid to pool at given levels of statistical heterogeneity

Chapter II: Optimizing Use of Effect Size Data

- Better reporting from RCTs (e.g. reporting of SDs, correlations from crossover trials, ICCs from cluster randomized trials, raw numbers for binary outcomes) would result in higher quality meta-analyses.
- More research needed on ratio of means—both clinical interpretability and mathematical consistency across studies compared to standardized mean difference.
- More research and transparency on ANCOVA models in adjusting for baseline imbalance.
- Software packages that more easily enable use of different information.
- More research on methods to handle zeros in the computation of binary outcomes
- Need for empirical vs. anecdotal evidence on which metrics are most helpful in conveying meta-analysis results to multiple stakeholders.

Chapter III: Choice of Statistical Model for Combining Studies

- Evaluation of the newly developed statistical models for combining the typical effect measures (e.g., mean difference, OR, RR, RD for common binary data) and the relative performance of the new and currently used methods, which may lead to improved estimates for meta-analysis.
- Evaluation of the relative performance of the recently developed statistical models for combining binary outcome to generate more evidence for model choice.

Chapter IV: Quantifying, Testing and Exploring Statistical Heterogeneity

- Future insights into heterogeneity statistics for meta-analyses involving a limited number of studies.
- Greater emphasis placed on specified and scientifically defensible hypotheses in meta-regression.
- Better reporting of relationships among study outcomes to facilitate multivariate meta-regression.

Chapter V: Network Meta-Analysis (Mixed treatment comparisons/indirect comparisons)

There are multiple challenges in conducting network meta-analysis, including:

- Need for methods for combining individual patient data with aggregated data;
- Integrating evidence from RCTs and observational studies;
- Modeling time-to-event data
- Development of user friendly software similar to that available for traditional pairwise meta-analysis
- Empirical or theoretical evidence to support model choice.

Chapter VI: Stability and Sensitivity Analyses in Evidence Synthesis

- Educational materials and case studies to demonstrate the conduct, interpretation and reporting of stability and sensitivity analyses.
- Simulation studies to determine the potential impact of various stability and sensitivity analysis choices

References

1. Fu R, Gartlehner G, Grant M, et al. *Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program* Agency for Healthcare Research and Quality. Rockville, MD: 2010.
2. Lau J, Chang S, Berkman N, et al. *EPC Response to IOM Standards for Systematic Reviews* Agency for Healthcare Research and Quality. Rockville, MD: 2013.
3. Lau J, Terrin N, Fu R. *Expanded Guidance on Selected Quantitative Synthesis Topics* Agency for Healthcare Research and Quality. Rockville, MD: 2013.
4. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. *Journal of clinical epidemiology*. 2010;63(5):502-12.
5. Fu R, Vandermeer BW, Shamliyan TA, et al. *Handling Continuous Outcomes in Quantitative Synthesis* Agency for Healthcare Research and Quality. Rockville, MD: 2013.
6. Verbeek J, Ruotsalainen J, Hoving JL. Synthesizing study results in a systematic review. *Scandinavian Journal of Work, Environment & Health*. 2012;38(3):282-90.
7. Berlin JA, Crowe BJ, Whalen E, et al. Meta-analysis of clinical trial safety data in a drug development program: Answers to frequently asked questions. *Clinical Trials*. 2013;10(1):20-31.
8. Gagnier JJ, Morgenstern H, Altman DG, et al. Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. *BMC Medical Research Methodology*. 2013;13(1):106.
9. Sun X, Guyatt G. Meta-analysis of randomized trials for health care interventions: one for all? *Journal of Evidence-Based Medicine*. 2009;2(1):53-6.
10. Turner RM, Bird SM, Higgins JP. The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. *PloS One*. 2013;8(3):e59202.
11. Rosén M. The aprotinin saga and the risks of conducting meta-analyses on small randomised controlled trials - a critique of a Cochrane review. *BMC Health Services Research*. 2009;19(9):34.
12. Bowater RJ, Escarela G. Heterogeneity and study size in random-effects meta-analysis. *Journal of Applied Statistics*. 2013;40(1):2-16.
13. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. *Journal of clinical epidemiology*. 2011;64(12):1283-93.
14. Wetterslev J, Thorlund K, Brok J, et al. Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC medical research methodology*. 2009;9(1):1.
15. Higgins JP, Green S. *Cochrane handbook for systematic reviews of interventions*: Wiley Online Library; 2008.
16. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-60.
17. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to meta-analysis*. Chichester, West Sussex, UK: Wiley; 2009.
18. Melsen WG, Bootsma MC, Rovers MM, et al. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. *Clinical Microbiology and Infection*. 2014;20(2):123-9.
19. Higgins J, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*. 2002;21(11):1539-58.
20. von Hippel PT. The heterogeneity statistic I² can be biased in small meta-

- analyses. *BMC Medical Research Methodology*. 2015;15(1):35.
21. Alba AC, Alexander PE, Chang J, et al. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. *Journal of clinical epidemiology*. 2016;70:129-35.
 22. Rhodes KM, Turner RM, Higgins JP. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Research synthesis methods*. 2015.
 23. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-34.
 24. Moher D, Schulz KF, Altman DG, et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*. 2001;357(9263):1191-4.
 25. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*. 2001;357(9263):1191-4.
 26. Brockhaus AC, Bender R, Skipka G. The Peto odds ratio viewed as a new effect measure. *Statistics in Medicine*. 2014;33(28):4861-74.
 27. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*. 2002;21(11):1575-600.
 28. Liu Z, Rich B, Hanley JA. Recovering the raw data behind a non-parametric survival curve. *Systematic Reviews*. 2014;3(1):1.
 29. M.W. M. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med*. 1996;15(16):1713-28.
 30. Schmid CH, Lau J, McIntosh MW, et al. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med*. 1998;17(17):1923-42.
 31. Glasziou PILM. An evidence based approach to individualising treatment. *BMJ*. 1995;311:1356.
 32. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med*. 1997;16(23):2741-58.
 33. Morton SC, Adams JL, Suttrop MJ, et al. Meta-regression Approaches: What, Why, When, and How?. *Technical Reviews*. 2004;8.
 34. J Sweeting M, J Sutton A, C Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*. 2004;23(9):1351-75.
 35. Durlak JA. How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology*. 2009;jsp004.
 36. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, New Jersey: L. Erlbaum; 1988.
 37. Senn S. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*. 1989;8(4):467-75.
 38. Senn S. Change from baseline and analysis of covariance revisited. *Statistics in Medicine*. 2006;25(24):4334-44.
 39. Balk EM, Earley A, Patel K, et al. *Empirical Assessment of Within-Arm Correlation Imputation in Trials of Continuous Outcomes Agency for Healthcare Research and Quality*. Rockville, MD: 2012.
 40. McKenzie JE, Herbison GP, Deeks JJ. Impact of analysing continuous outcomes using final values, change scores and analysis of covariance on the performance of meta-analytic methods: a simulation study. *Research Synthesis Methods*. 2015.
 41. McKenzie JE, Herbison GP, deeks JJ. Impact of analysing continuous outcomes using final values, change scores and analysis of covariance on the performance of

- meta-analytic methods: a simulation study. *Research Synthesis Methods*. 2015.
42. Camilli G, de la Torre J, Chiu C. A Noncentral t Regression Model for Meta-Analysis *Journal of Educational and Behavioral Statistics*. 2010;35(2):125-53.
43. Claggett B, Xie M, Tian L. Meta-Analysis With Fixed, Unknown, Study-Specific Parameters. *Journal of the American Statistical Association*. 2014;109(508):1660-71.
44. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*. 2014;15(2):pp. 222-33.
45. Doi SA, Barendregt JJ, Khan S, et al. Advances in the Meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemporary Clinical Trials*. 2015;45(Pt. A):130-8.
46. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine*. 2015.
47. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials*. 1986;7(3):177-88.
48. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*. 2001;20(6):825-40.
49. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*. 2007;28(2):105-14.
50. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*. 2015;45(Pt. A):139-45.
51. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine*. 2002;21(21):3153-9.
52. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*. 2001;20(12):1771-82.
53. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*. 2001;20(24):3875-89.
54. Biggerstaff B, Tweedie R. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*. 1997;16(7):753-68.
55. Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*. 2010;140(4):961-70.
56. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research*. 2012;21(4):409-26.
57. Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. *Statistical methods in medical research*. 2015.
58. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between DerSimonian–Laird and restricted maximum likelihood. *Statistical Methods in Medical Research*. 2012;21(6):657-9.
59. Int'Hout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*. 2014;14(1):25.
60. Cornell JE, Mulrow CD, Localio R, et al. Random-Effects Meta-analysis of Inconsistent Effects: A Time for Change. *Annals of Internal Medicine*. 2014;160(4):267-70.
61. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*. 2007;26(1):53-77.

62. Shuster JJ, Walker MA. Low-event-rate meta-analyses of clinical trials: implementing good practices. *Statistics in medicine*. 2016.
63. Bhaumik DK, Amatya A, Normand SL, et al. Meta-Analysis of Rare Binary Adverse Event Data. *Journal of the American Statistical Association*. 2012;107(498):555-67.
64. Vázquez F, Moreno E, Negrín M, et al. Bayesian robustness in meta-analysis for studies with zero responses. *Pharmaceutical statistics*. 2016.
65. Bai O, Chen M, Wang X. Bayesian Estimation and Testing in Random Effects Meta-Analysis of Rare Binary Adverse Events. *Statistics in biopharmaceutical research*. 2016;8(1):49-59.
66. Fleiss J. Review papers: The statistical basis of meta-analysis. *Statistical methods in medical research*. 1993;2(2):121-45.
67. Vandermeer B, Bialy L, Hooton N, et al. Meta-analyses of safety data: a comparison of exact versus asymptotic methods. *Statistical Methods in Medical Research*. 2009;18(4):421-32.
68. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine*. 2010;29(12):1259-65.
69. Spittal MJ, Pirkis J, Gurrin LC. Meta-analysis of incidence rate data in the presence of zero events. *BMC medical research methodology*. 2015;15(1):1.
70. Kuss O. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. *Statistics in medicine*. 2015;34(7):1097-116.
71. Ma L, Soriano J. Analysis of distributional variation through multi-scale Beta-Binomial modeling. *arXiv preprint arXiv:1604.01443*. 2016.
72. Huang HY, Andrews E, Jones J, et al. Pitfalls in meta-analyses on adverse events reported from clinical trials. *Pharmacoepidemiology and Drug Safety*. 2011;20(10):1014-20.
73. Warren FC. *An Exploration of Evidence Synthesis Methods for Adverse Events*. Leicester, UK U7 - <http://hdl.handle.net/2381/10232> U8 - http://www.worldcat.org/title/exploration-of-evidence-synthesis-methods-for-adverse-events/oclc/806195349&referer=brief_results U13 - Sent #1 2015: University of Leicester; 2010.
74. Mehta CR. The exact analysis of contingency tables in medical research. *Recent Advances in Clinical Trial Design and Analysis*. Springer; 1995:177-202.
75. Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Statistics in Medicine*. 1995;14(19):2143-60.
76. Tian L, Cai T, Pfeffer MA, et al. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 x 2 tables with all available data but without artificial continuity correction. *Biostatistics*. 2009;10(2):275-81.
77. Liu D, Liu RY, Xie M. Exact Meta-Analysis Approach for Discrete Data and its Application to 2 x 2 Tables With Rare Events. *Journal of the American Statistical Association*. 2014;109(508):1450-65.
78. Yang G, Liu D, Wang J, et al. Meta-analysis framework for exact inferences with application to the analysis of rare events. *Biometrics*. 2016.
79. Bickel R. *Multilevel analysis for applied research: It's just regression!*: Guilford Press; 2007.
80. Dias S, Sutton AJ, Ades AE, et al. *Evidence Synthesis for Decision Making 2: A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials*. *Medical Decision Making*. 2013;33(5):607-17.
81. Sutton AJ, Welton NJ, Ades A, et al. *Evidence synthesis for decision making in healthcare*: John Wiley & Sons; 2012.

82. Dakin HA, Welton NJ, Ades AE, et al. Mixed treatment comparison of repeated measurements of a continuous endpoint: An example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Statistics in Medicine*. 2011;30(20):2511-35.
83. Schmid CH. Using Bayesian inference to perform meta-analysis. *Evaluation & the health professions*. 2001;24(2):165-89.
84. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*: John Wiley & Sons; 2004.
85. Lambert PC, Sutton AJ, Burton PR, et al. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*. 2005;24(15):2401-28.
86. GajicVeljanoski O, Cheung AM, Bayoumi AM, et al. The choice of a noninformative prior on between-study variance strongly affects predictions of future treatment effect. *Medical Decision Making*. 2013;33(3):356-68.
87. HIGGINS J, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*. 1996;15(24):2733-49.
88. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. 2009;172(1):137-59.
89. Turner RM, Davey J, Clarke MJ, et al. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology*. 2012;41(3):818-27.
90. Ades A, Lu G, Higgins J. The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making*. 2005;25(6):646-54.
91. Unit WMB. WinBUGS | MRC Biostatistics Unit. 2016. <http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/2016>.
92. OpenBUGS. OpenBUGS. 2016. <http://www.openbugs.net/w/FrontPage2016>.
93. . JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*; 2003. Vienna; 124.
94. JAGS - Just Another Gibbs Sampler. *JAGS - Just Another Gibbs Sampler.*; 2016. <http://mcmc-jags.sourceforge.net/2016>.
95. Carpenter B, Gelman A, Hoffman M, et al. Stan: A probabilistic programming language. *J Stat Softw*. 2016.
96. de Valpine P, Turek D, Paciorek CJ, et al. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*. 2016(just-accepted):1-28.
97. nimble-admin. NIMBLE | An R package for programming with BUGS models and compiling parts of R. 2016. <http://r-nimble.org/2016>.
98. Thompson J, Palmer T, Moreno S. Bayesian analysis in Stata using WinBUGS. *The Stata Journal*. 2006;6(4):530-49.
99. bmeta. bmeta - Bayesian meta-analysis & meta-regression in R - Gianluca Baio. . 2016. <https://sites.google.com/a/statistica.it/gianluca/bmeta2016>.
100. van Valkenhoef G, Lu G, de Brock B, et al. Automating network meta-analysis. *Research Synthesis Methods*. 2012;3(4):285-99.
101. van Valkenhoef G, Dias S, Ades AE, et al. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. *Research Synthesis Methods*. 2016;7(1):80-93.
102. SAS/STAT. SAS/STAT Software Examples: Bayesian Hierarchical Modeling for Meta-Analysis. . 2016.

http://support.sas.com/rnd/app/examples/stat/BayesMeta/new_example/index.html2016.

103. Bayesian “random-effects” models. Bayesian “random-effects” models | Stata News. . 2016. <http://www.stata.com/stata-news/news30-2/bayesian-random-effects/2016>.
104. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*. 2010;29(29):3046-67.
105. Senn S. Trying to be precise about vagueness. *Statistics in medicine*. 2007;26(7):1417.
106. Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Statistics in Medicine*. 2015;34(7):1097-116.
107. Moreno E, Vázquez-Polo FJ, Negrin MA. Objective Bayesian meta-analysis for sparse discrete data. *Statistics in medicine*. 2014;33(21):3676-92.
108. Vazquez FJ, Moreno E, Negrin MA, et al. Bayesian robustness in meta-analysis for studies with zero responses. *Pharmaceutical statistics*. 2016.
109. Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Statistics in Medicine*. 2003;22(14):2309-33.
110. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*. 2011;30(20):2481-98.
111. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*. 2010;29(12):1282-97.
112. Jackson D, White IR, Riley RD. A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression.

- Biometrical Journal. Biometrische Zeitschrift*. 2013;55(2):231-45.
113. Jackson D, Rollins K, Coughlin P. A multivariate model for the meta-analysis of study level survival data at multiple times. *Research Synthesis Methods*. 2014;5(3):264-72.
 114. Chen H, Manning AK, Dupuis J. A method of moments estimator for random effect multivariate meta-analysis. *Biometrics*. 2012;68(4):1278-84.
 115. Van den Noortgate W, López-López JA, Marín-Martínez F, et al. Meta-analysis of multiple outcomes: a multilevel approach. *Behavior Research Methods*. 2015;47(4):1274-94.
 116. Hurtado Rua SM, Mazumdar M, Strawderman RL. The choice of prior distribution for a covariance matrix in multivariate meta-analysis: a simulation study. *Statistics in Medicine*. 2015;34(30):4083-104.
 117. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine*. 2012;31(20):2179-95.
 118. Frosi G, Riley RD, Williamson PR, et al. Multivariate meta-analysis helps examine the impact of outcome reporting bias in Cochrane rheumatoid arthritis reviews. *Journal of Clinical Epidemiology*. 2014;68(5):542-50.
 119. Kent DM, Rothwell PM, Ioannidis JP, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*. 2010 Aug 12;11:85. PMID: 20704705.
 120. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol*. 2008 Oct;37(5):1158-60. PMID: 18832388.
 121. Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, et al. Assessing heterogeneity in meta-analysis: Q statistic or

- I2 index? *Psychol Methods*. 2006 Jun;11(2):193-206. PMID: 16784338.
122. Mittlbock M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med*. 2006 Dec 30;25(24):4321-33. PMID: 16991104.
123. Thorlund K, Imberger G, Johnston BC, et al. Evolution of heterogeneity (I2) estimates and their 95% confidence intervals in large meta-analyses. *PLoS One*. 2012;7(7):e39471. PMID: 22848355.
124. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007 Nov 3;335(7626):914-6. PMID: 17974687.
125. Anzures-Cabrera J, Higgins JP. Graphical displays for meta-analysis: An overview with suggestions for practice. *Res Synth Methods*. 2010 Jan;1(1):66-80. PMID: 26056093.
126. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med*. 1998 Apr 30;17(8):841-56. PMID: 9595615.
127. Baujat B, Mahe C, Pignon JP, et al. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat Med*. 2002 Sep 30;21(18):2641-52. PMID: 12228882.
128. Bowden J, Tierney JF, Copas AJ, et al. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Med Res Methodol*. 2011 Apr 07;11:41. PMID: 21473747.
129. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002 Jun 15;21(11):1559-73. PMID: 12111920.
130. Berkey CS, Hoaglin DC, Mosteller F, et al. A random-effects regression model for meta-analysis. *Stat Med*. 1995 Feb 28;14(4):395-411. PMID: 7746979.
131. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med*. 2004 Jun 15;23(11):1663-82. PMID: 15160401.
132. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003 Sep 15;22(17):2693-710. PMID: 12939780.
133. Berlin JA, Santanna J, Schmid CH, et al. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med*. 2002 Feb 15;21(3):371-87. PMID: 11813224.
134. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med*. 1992 Jul 23;327(4):248-54. PMID: 1614465.
135. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med*. 2002 Feb 28;21(4):589-624. PMID: 11836738.
136. Riley RD, Abrams KR, Lambert PC, et al. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat Med*. 2007 Jan 15;26(1):78-97. PMID: 16526010.
137. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med*. 2011 Sep 10;30(20):2481-98. PMID: 21268052.
138. Hedges LV. Comment on 'Multivariate meta-analysis: potential and promise'. *Stat Med*. 2011 Sep 10;30(20):2499; discussion 509-10. PMID: 25522447.
139. Bland JM. Comments on 'Multivariate meta-analysis: potential and promise' by Jackson et al, *Statistics in Medicine*. *Stat Med*. 2011 Sep 10;30(20):2502-3; discussion 9-10. PMID: 25522449.
140. Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Stat Med*. 2012 Dec 20;31(29):3805-20. PMID: 22763950.

141. Jackson D, Riley RD. A refined method for multivariate meta-analysis and meta-regression. *Stat Med*. 2014 Feb 20;33(4):541-54. PMID: 23996351.
142. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol*. 1992 Jun 1;135(11):1301-9. PMID: 1626547.
143. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology*. 1993 May;4(3):218-28. PMID: 8512986.
144. Crippa A, Orsini N. Dose-response meta-analysis of differences in means. *BMC Med Res Methodol*. 2016 Aug 02;16:91. PMID: 27485429.
145. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *Journal of clinical epidemiology*. 2011;64(12):1294-302.
146. Higgins JP, Jackson D, Barrett JK, et al. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research synthesis methods*. 2012 Jun;3(2):98-110. PMID: 26062084.
147. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*. 2012;3(2):80-97.
148. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003 Mar 1;326(7387):472. PMID: 12609941.
149. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology*. 1997 Jun;50(6):683-91. PMID: 9250266.
150. Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005;9(26):1-148.
151. Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in medicine*. 2002;21(16):2313-24.
152. Salanti G, Higgins JP, Ades A, et al. Evaluation of networks of randomized trials. *Statistical methods in medical research*. 2008;17(3):279-301.
153. White IR, Barrett JK, Jackson D, et al. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods*. 2012;3(2):111-25.
154. Higgins JPT, Jackson D, Barrett JK, et al. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods*. 2012;3(2):98-110.
155. Greco T, Edefonti V, Biondi-Zoccai G, et al. A multilevel approach to network meta-analysis within a frequentist framework. *Contemporary Clinical Trials*. 2015;42:51-9.
156. Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*. 2004;23(20):3105-24.
157. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*. 2012.
158. Cooper NJ, Sutton AJ, Morris D, et al. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in medicine*. 2009;28(14):1861-81.
159. Salanti G, Dias S, Welton NJ, et al. Evaluating novel agent effects in multiple-treatments meta-regression. *Statistics in Medicine*. 2010;29(23):2369-83.

160. Turner RM, Jackson D, Wei Y, et al. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine*. 2014.
161. Greco T, Landoni G, Biondi-Zoccai G, et al. A Bayesian network meta-analysis for binary outcome: how to do it. *Statistical Methods in Medical Research*. 2013.
162. Hong H, Carlin BP, Shamliyan TA, et al. Comparing Bayesian and Frequentist Approaches for Multiple Outcome Mixed Treatment Comparisons. *Medical Decision Making*. 2013;33(5):702-14.
163. Salanti G, Higgins JP, Ades AE, et al. Evaluation of networks of randomized trials. *Statistical methods in medical research*. 2008 Jun;17(3):279-301. PMID: 17925316.
164. Hawkins N, Scott DA, Woods B. 'Arm-based' parameterization for network meta-analysis. *Research synthesis methods*. 2015.
165. Hong H, Chu H, Zhang J, et al. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research synthesis methods*. 2016 Mar;7(1):6-22. PMID: 26536149.
166. Zhang J, Chu H, Hong H, et al. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Statistical methods in medical research*. 2015.
167. Zhang J, Carlin BP, Neaton JD, et al. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. *Clinical Trials*. 2014;11(2):246-62.
168. Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Research synthesis methods*. 2016 Mar;7(1):23-8. PMID: 26461457.
169. White IR. Network meta-analysis. *The Stata Journal*. 2015.
170. Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in medicine*. 2002 Aug 30;21(16):2313-24. PMID: 12210616.
171. Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*. 2010;29(7-8):932-44.
172. Chaimani A, Higgins JP, Mavridis D, et al. Graphical Tools for Network Meta-Analysis in STATA. *PloS One*. 2013;8(10):e76654.
173. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Medical Research Methodology*. 2013;13:35.
174. Donegan S, Williamson P, D'Alessandro U, et al. Assessing key assumptions of network meta-analysis: a review of methods. *Research Synthesis Methods*. 2013;4(4):291-323.
175. Piepho HP. Network-meta analysis made easy: detection of inconsistency using factorial analysis-of-variance models. *BMC medical research methodology*. 2014;14:61. PMID: 24885590.
176. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *Journal of clinical epidemiology*. 2008 May;61(5):455-63. PMID: 18394538.
177. van der Valk R, Webers CA, Lumley T, et al. A network meta-analysis combined direct and indirect comparisons between glaucoma drugs to rank effectiveness in lowering intraocular pressure. *Journal of clinical epidemiology*. 2009 Dec;62(12):1279-83. PMID: 19716679.
178. Baker SG, Kramer BS. The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C? *BMC Medical Research Methodology*. 2002;2(1):13.
179. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of

- randomized controlled trials. *Journal of clinical epidemiology*. 1997;50(6):683-91.
180. Caldwell DM, Ades A, Higgins J. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *British Medical Journal*. 2005;7521:897.
181. Glenny A, Altman D, Song F, et al. Indirect comparisons of competing interventions: NCCHTA; 2005.
182. Song F, Glenny A-M, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. *Controlled clinical trials*. 2000;21(5):488-97.
183. Chou R, Fu R, Huffman LH, et al. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *The Lancet*. 2006;368(9546):1503-15.
184. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of Clinical Epidemiology*. 2011;64(12):1277-82.
185. Salanti G, Del Giovane C, Chaimani A, et al. Evaluating the quality of evidence from a network meta-analysis. *PloS one*. 2014;9(7):e99682.
186. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ*. 2014;349:g5630.
187. Murad MH, Montori VM, Ioannidis JP, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA*. 2014;312(2):171-9.
188. Guyatt GH, Eikelboom JW, Gould MK, et al. Approach to outcome measurement in the prevention of thrombosis in surgical and medical patients: Antithrombotic Therapy and Prevention of Thrombosis: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *CHEST Journal*. 2012;141(2_suppl):e185S-e94S.
189. Bafeta A, Trinquart L, Seror R, et al. Reporting of results from network meta-analyses: methodological systematic review. *BMJ*. 2014;348.
190. Song F, Loke YK, Walsh T, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ*. 2009;338:b1147.
191. Hutton B, Salanti G, Chaimani A, et al. The quality of reporting methods and results in network meta-analyses: an overview of reviews and suggestions for improvement. *PloS one*. 2014;9(3):e92508.
192. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. *Annals of Internal Medicine*. 2015;162(11):777-84.
193. Stigler SM. The changing history of robustness. *The American Statistician*. 2012.
194. Wimsatt WC. Robustness, reliability, and overdetermination. *Scientific inquiry and the social sciences*. 1981:124-63.
195. Woodward J. Some varieties of robustness. *Journal of Economic Methodology*. 2006;13(2):219-40.
196. Rubin DB. Meta-analysis: literature synthesis or effect-size surface estimation? *Journal of Educational and Behavioral Statistics*. 1992;17(4):363-74.
197. Rosenbaum PR. *Observational studies*. 2nd ed. New York: Springer; 2002.
198. Halladay CW, Trikalinos TA, Schmid IT, et al. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. *Journal of clinical epidemiology*. 2015;68(9):1076-84.

199. Olkin I. Diagnostic statistical procedures in medical meta-analyses. *Stat. Med.* 1999;18(17-18):2331-41.
200. Bax L, Ikeda N, Fukui N, et al. More than numbers: the power of graphs in meta-analysis. *Am. J. Epidemiol.* 2009 15~January;169(2):249-55.
201. Elvik R. Evaluating the statistical conclusion validity of weighted mean results in meta-analysis by analysing funnel graph diagrams. *Accid. Anal. Prev.* 1998 March;30(2):255-66.
202. Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics.* 2000 September;1(3):247-62.