Methods Guide for Comparative Effectiveness Reviews

Quantitative Synthesis—An Update

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.

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

This research was funded through contracts from the Agency for Healthcare Research and Quality to the following Evidence-based Practice Centers: Mayo Clinic (290-2015-00013-I); Kaiser Permanente (290-2015-00007-I); RAND Corporation (290-2015-00010-I); Alberta (290-2015-00001-I); Pacific Northwest (290-2015-00009-I); RTI (290-2015-00011-I); Brown (290-2015-00002-I); and the Scientific Resource Center (290-2012-00004-C).

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policy makers, 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).

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

Suggested citation: Morton SC, Murad MH, O'Connor E, Lee CS, Booth M, Vandermeer BW, Snowden JM, D'Anci KE, Fu R, Gartlehner G, Wang Z, Steele DW. Quantitative Synthesis—An Update. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Scientific Resource Center under Contract No. 290-2012-0004-C). AHRQ Publication No. 18-EHC007-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2018. Erratum October 2022. Posted final reports are located on the Effective Health Care Program search page. DOI: https://doi.org/10.23970/AHRQEPCMETHGUIDE3.

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

Contract No.: 290-2012-00004-C

Prepared by:
Scientific Resource Center
Portland, OR

Investigators:
Sally C. Morton, Ph.D., M.Sc.[1]
M. Hassan Murad, M.D., M.P.H.[2]
Elizabeth O'Connor, Ph.D.[3]
Christopher S. Lee, Ph.D., R.N.[1]
Marika Booth, M.S.[4]
Benjamin W. Vandermeer, M.Sc.[5]
Jonathan M. Snowden, Ph.D.[1]
Kristen E. D'Anci, Ph.D.[6]
Rongwei Fu, Ph.D.[7]
Gerald Gartlehner, M.D., M.P.H.[8]
Zhen Wang, Ph.D.[2]
Dale W. Steele M.D., M.S.[9]

[1] Scientific Resource Center for the AHRQ Effective Health Care Program, Portland VA Research Foundation, VA Portland Health Care Systems, Portland, OR
[2] Mayo Clinic Evidence-based Practice Center, Rochester MN
[3] Kaiser Permanente Research Affiliates Evidence-based Practice Center, Portland OR
[4] Southern California Evidence-based Practice Center – RAND Corporation, Santa Monica, CA
[5] University of Alberta Evidence-based Practice Center, Edmonton, AB
[6] ECRI Institute – Penn Medicine Evidence-based Practice Center, Plymouth Meeting, PA
[7] Pacific-Northwest Evidence-based Practice Center – Oregon Health & Sciences University, Portland, OR
[8] RTI International – University of North Carolina (UNC) Evidence-based Practice Center, Chapel Hill, NC
[9] Brown University Center for Evidence-based Medicine, Providence, RI

Erratum
In the original version of this report there was an error in Bucher's method equation on page 37.

Text in the original report: "When there are only two sets of trials, say, A vs. B and B vs. C, Bucher's method is sufficient to provide the indirect estimate of A vs. C as: $\log(OR_{AC})=\log(OR_{AB})-\log(OR_{BC})$ and $Var(Log(OR_{AC})) = Var(Log(OR_{AB})) + Var(Log(OR_{BC}))$, where OR is the odds ratio."

The corrected text: "When there are only two sets of trials, say, A vs. B and C vs. B, Bucher's method is sufficient to provide the indirect estimate of A vs. C as: $\log(OR_{AC})=\log(OR_{AB})-\log(OR_{CB})$ and $Var(Log(OR_{AC})) = Var(Log(OR_{AB})) + Var(Log(OR_{CB}))$, where OR is the odds ratio."

This has been corrected in the report.

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, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov.

Gopal Khanna, M.B.A.
Director
Agency for Healthcare Research and Quality

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

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

Elisabeth Kato, M.D., M.R.P.
Task Order Officer
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

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 investigators.

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:

Eric Bass, M.D., M.P.H
Director, Johns Hopkins University
Evidence-based Practice Center
Professor of Medicine, and Health Policy and Management
Johns Hopkins University
Baltimore, MD

Mary Butler, M.B.A., Ph.D.
Co-Director, Minnesota Evidence-based Practice Center
Assistant Professor, Health Policy & Management
University of Minnesota
Minneapolis, MN

Roger Chou, M.D., FACP
Director, Pacific Northwest Evidence-based Practice Center
Portland, OR

Lisa Hartling, M.S., Ph.D.
Director, University of Alberta Evidence-Practice Center
Edmonton, AB

Susanne Hempel, Ph.D.
Co-Director, Southern California Evidence-based Practice Center
Professor, Pardee RAND Graduate School
Senior Behavioral Scientist, RAND Corporation
Santa Monica, CA

Robert L. Kane, M.D.*
Co-Director, Minnesota Evidence-based Practice Center
School of Public Health
University of Minnesota
Minneapolis, MN

Jennifer Lin, M.D., M.C.R.
Director, Kaiser Permanente Research Affiliates Evidence-based Practice Center
Investigator, The Center for Health Research, Kaiser Permanente Northwest
Portland, OR

Christopher Schmid, Ph.D.
Co-Director, Center for Evidence Synthesis in Health
Professor of Biostatistics
School of Public Health
Brown University
Providence, RI

Karen Schoelles, M.D., S.M., FACP
Director, ECRI Evidence-based Practice Center
Plymouth Meeting, PA

Tibor Schuster, Ph.D.
Assistant Professor
Department of Family Medicine
McGill University
Montreal, QC

*Deceased March 6, 2017

Jonathan R. Treadwell, Ph.D.
Associate Director, ECRI Institute
Evidence-based Practice Center
Plymouth Meeting, PA

Tom Trikalinos, M.D.
Director, Brown Evidence-based Practice Center
Director, Center for Evidence-based Medicine
Associate Professor, Health Services, Policy & Practice
Brown University
Providence, RI

Meera Viswanathan, Ph.D.
Director, RTI-UNC Evidence-based Practice Center
Durham, NC
RTI International
Durham, NC

C. Michael White, Pharm. D., FCP, FCCP
Professor and Head, Pharmacy Practice
School of Pharmacy
University of Connecticut
Storrs, CT

Tim Wilt, M.D., M.P.H.
Co-Director, Minnesota Evidence-based Practice Center
Director, Minneapolis VA-Evidence Synthesis Program
Professor of Medicine, University of Minnesota
Staff Physician, Minneapolis VA Health Care System
Minneapolis, MN

Abstract

Quantitative synthesis, or meta-analysis, is often essential for Comparative Effective Reviews (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 it is intended to provide a consistent approach for situations and decisions that are commonly faced by AHRQ Evidence-based Practice Centers (EPCs). The goal is to describe choices as explicitly as possible, and in the context of EPC requirements, with an appropriate degree of confidence.

This guide addresses issues in the order that they are usually encountered 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 chapter addresses how to extract and utilize data from individual studies to construct effect sizes, followed by a chapter on statistical model choice. The fourth chapter considers quantifying and exploring heterogeneity. The fifth describes an indirect evidence technique that has not been included in previous guidance – network meta-analysis, also known as mixed treatment comparisons. The final section in the report lays out future research suggestions.

Contents

Introduction **1**
Background 1
Methods 1
Literature Search and Review 2
Consensus and Recommendations 2
Chapter 1. Decision to Combine Trials **3**
1.1. Goals of the Meta-Analysis 3
1.2. Clinical and Methodological Heterogeneity 3
1.3. Best Evidence Versus All Evidence 5
1.4. Assessing the Risk of Misleading Meta-analysis Results 6
Wide-Ranging Effect Sizes 6
Suspicion of Publication or Reporting Bias 6
Small Studies Effect 6
1.5. Special Considerations When Pooling a Small Number of Studies 7
Rare Outcomes 7
Small Sample Sizes 7
Wide-Ranging Effect Sizes 8
1.6. Statistical Heterogeneity 8
1.7. Conclusion 9
Recommendations 9
Chapter 2. Optimizing Use of Effect Size Data **10**
2.1. Introduction 10
2.2. Nuances of Binary Effect Sizes 10
Data Needed for Binary Effect Size Computation 10
Choosing Among Effect Size Options 11
Equation Set 2.1. Risk Difference 12
Equation Set 2.2. Risk Ratio 13
Equation Set 2.3. Odds Ratios 14
Equation Set 2.4. Peto Odds Ratios 15
2.3. Continuous Outcomes 16
Assembling Data Needed for Effect Size Computation 16
(Weighted) Mean Difference 16

Standardized Mean Difference 17
2.4. Special Topics 18
Crossover Trials 18
Cluster Randomized Trials 19
Mean Difference and Baseline Imbalance 19
Recommendations: 20
Chapter 3. Choice of Statistical Model for Combining Studies 21
3.1. Introduction 21
General Considerations for Model Choice 21
Choice of Random Effects Model and Estimator 22
Role of Generalized Linear Mixed Effects models 23
3.2. A Special Case: Combining Rare Binary Outcomes 23
A Note on an Exact Method for Sparse Binary Data 24
3.3. Bayesian Methods 25
A Note on using a Bayesian Approach for Sparse Binary Data 26
3.4. Recommendations 26
Chapter 4. Quantifying, Testing, and Exploring Statistical Heterogeneity 27
4.1. Statistical Heterogeneity in Meta-analysis 27
4.2. Visually Inspecting Heterogeneity 27
Forest Plots 27
Funnel Plots 27
4.3. Quantifying Heterogeneity 28
Graphical Options for Examining Contributions to *Q* 29
Between-Study Variance 29
Inconsistency Across Studies 30
4.4. Exploring Heterogeneity 31
Meta-Regression 31
Multiple Meta-regression 31
Subgroup Analysis 32
Detecting Outlying Studies 32
4.5. Special Topics 33
Baseline Risk (Control-Rate) Meta-regression 33
Multivariate Meta-analysis 33

Dose-Response Meta-analysis 34
Recommendations 35
Chapter 5. Network Meta-Analysis (Mixed treatment comparisons/indirect comparisons) 36
5.1. Rationale and Definition 36
5.2. Assumptions 36
5.3. Statistical Approaches 37
Overview 37
Simple Indirect Comparisons 37
Mixed Effects and Hierarchical Models 38
Frequentist Approach 38
Bayesian Approach 38
Arm-Based Versus Contrast-Based Models 39
Assessing Consistency 39
5.4. Considerations of Model Choice and Software 41
Consideration of Indirect Evidence 41
Choice of Method 41
Mixed Effects and Hierarchical Models 42
Commonly Used Software 43
5.5. Inference From Network Meta-analysis 43
Approaches for Rating the Strength of Evidence: 43
Interpreting Ranking Probabilities and Clinical Importance of Results 44
5.6. Presentation and Reporting 44
Recommendations 45
Future Research Suggestions 46
Chapter 1. Decision To Combine Trials 46
Chapter 2. Optimizing Use of Effect Size Data 46
Chapter 3. Choice of Statistical Model for Combining Studies 46
Chapter 4. Quantifying, Testing, and Exploring Statistical Heterogeneity 46
Chapter 5. Network Meta-analysis (Mixed Treatment Comparisons/Indirect Comparisons) 46
References 47

Tables

Table 2.1. Assembling binary data for effect size computation 10
Table 2.2. Organizing binary data for effect size computation 12

Table 2.3. Benefits and disadvantages of binary data effect sizes[41] 15
Table 5.1. Impact of network geometry on choice of analysis method 42

Figures

Figure 1.1. Pooling decision tree 4
Figure 5.1. Common network geometry (simple indirect comparison, star, network with at least one closed loop) 42

Introduction

Background

The purpose of this document is to consolidate and update quantitative synthesis guidance provided in three previous methods guides.[1-3] We focus primarily on comparative effectiveness reviews (CERs), which are systematic reviews that compare the 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 and do not address diagnostic studies, individual patient level analysis, or observational studies, which are addressed elsewhere.[4]

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 it is intended to provide a consistent approach for situations and decisions that are commonly faced by Evidence-based Practice Centers (EPCs). The goal is to describe choices as explicitly as possible and in the context of EPC requirements, with an appropriate degree of confidence.

EPC investigators are encouraged to follow these recommendations but may choose to use alternative methods if deemed necessary after discussion with their AHRQ project officer. If alternative methods are used, investigators are required to provide a rationale for their choices, and if appropriate, to state the strengths and limitations of the chosen methods in order to promote consistency, transparency, and learning. In addition, several steps in meta-analysis require subjective judgment, such as when combining studies or incorporating 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 encountered 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 chapter addresses how to extract and utilize data from individual studies to construct effect sizes, followed by a chapter on statistical model choice. The fourth chapter considers quantifying and exploring heterogeneity. The fifth describes an indirect evidence technique that has not been included in previous guidance – network meta-analysis, also known as mixed treatment comparisons. The final section in the report lays out future research suggestions.

Methods

This guide was developed by 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 among AHRQ, Directors of EPCs, the Scientific Resource Center, and the Methods Steering Committee, quantitative synthesis was identified as a high-priority methods topic and a need was identified to update the original guidance.[1, 5] Once confirmed as a Methods Workgroup, the SRC solicited EPC workgroup volunteers, particularly those with quantitative methods expertise, including statisticians, librarians, thought leaders, and methodologists. Charged by AHRQ to update current guidance, the workgroup consisted of members from eight of 13 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 as described below. The larger workgroup then was split into subgroups each taking responsibility for a different chapter. 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 worked with each subgroup to identify a relevant search strategy for each chapter, and then combined these strategies into one overall search conducted for all chapters combined. The librarian conducted the search on the ARHQ SRC Methods Library, a bibliographic database curated by the SRC currently containing more than 16,000 citations of methodological works for systematic reviews and comparative effectiveness reviews, using descriptor and keyword strategies to identify quantitative synthesis methods research publications (descriptor search=all quantitative synthesis descriptors, and the keyword search=quantitative synthesis, meta-anal*, metaanal*, meta-regression in [anywhere field]). Search results were limited to English language and 2009 and later to capture citations published since AHRQ's previous methods guidance on quantitative synthesis. Additional articles were identified from recent systematic reviews, reference lists of reviews and editorials, and through the expert review process.

The search yielded 1,358 titles and abstracts which were reviewed by all workgroup members using ABSTRACKR software (available at http://abstrackr.cebm.brown.edu). Each subgroup separately identified articles relevant to their own chapter. Abstract review was done by single review, investigators included anything that could be potentially relevant. Each subgroup decided separately on final inclusion/exclusion based on full text articles.

Consensus and Recommendations

Reaching consensus if possible is of great importance for AHRQ methods guidance. The workgroup recognized this importance in its first meeting and agreed on a process for informal consensus and conflict resolution. Disagreements were thoroughly discussed and if possible, consensus was reached. If consensus was not reached, analytic options are discussed in the text. We did not employ a formal voting procedure to assess consensus.

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

Chapter 1. Decision to Combine Trials

Elizabeth O'Connor, Ph.D., Kristen E. D'Anci, Ph.D., Jonathan M. Snowden Ph.D.

1.1. Goals of the Meta-analysis

Meta-analysis is a statistical method for synthesizing (also called combining or pooling) the benefits and/or harms of a treatment or intervention across multiple studies. The overarching goal of a meta-analysis 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]

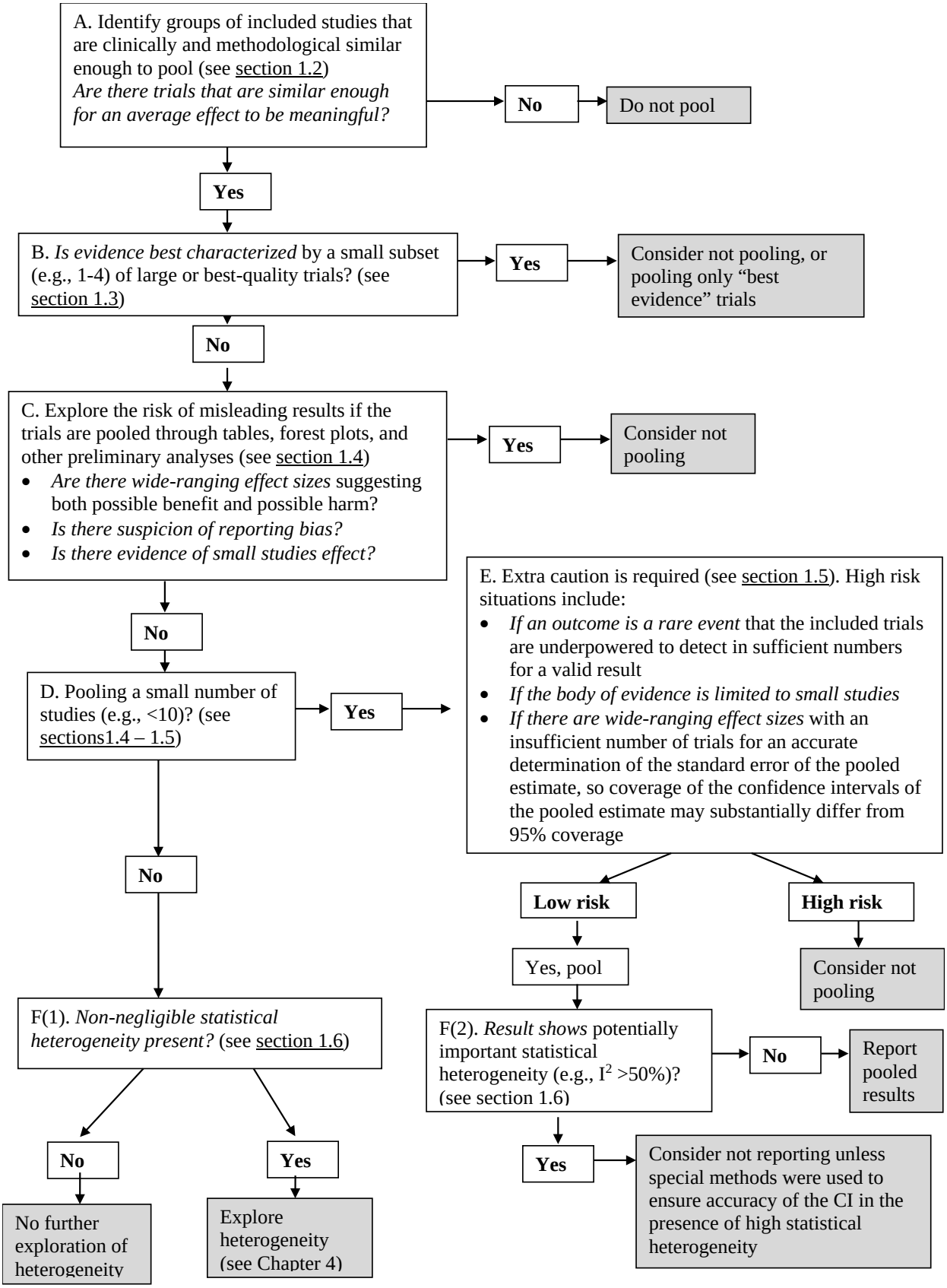
Clinical, methodological, and statistical factors must all be considered when deciding whether to combine studies in a meta-analysis. Figure 1.1 depicts a decision tree to help investigators think through these important considerations, which are discussed below.

1.2. Clinical and Methodological Heterogeneity

Studies must be reasonably similar to be pooled in a meta-analysis.[1] Even when the review protocol identifies a coherent and fairly narrow body of literature, the actual included studies may represent a wide range of 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.

The first step in the decision tree is to explore the clinical and methodological heterogeneity of the included studies (Step A, Figure 1.1). The goal is to identify groups of trials that are similar enough that an average effect would make a sensible summary. There is no objective measure or universally accepted standard for deciding whether studies are "similar enough" to pool; this decision is inherently a matter of judgment.[6] Verbeek and colleagues suggest working through key sources of variability in sequence, beginning with the clinical variables of intervention/exposure, control condition, and participants, before moving on to methodological areas such as study design, outcome, and follow-up time. When there is important variability in these areas, investigators should consider whether there are coherent subgroups of trials, rather than the full group, that can be pooled.[6]

Figure 1.1. Pooling decision tree



Clinical heterogeneity refers to characteristics related to the participants, interventions, types of outcomes, and study setting. 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] For example, in a review of a lipid-lowering medication, researchers might be comfortable combining studies that target younger and middle-aged adults, but expect different effects with older adults, who have high rates of comorbidities and other medication use. Others suggest that it may be acceptable to combine interventions with likely similar mechanisms of action.[6] For example, a researcher may combine studies of depression interventions that use a range of psychotherapeutic approaches, on the logic that they all aim to change a person's thinking and behavior in order to improve mood, but not want to combine them with trials of antidepressants, whose mechanism of action is presumed to be biochemical.

Methodological heterogeneity refers to variations in study methods (e.g., study design, measures, and study conduct). A common question regarding study design, is whether it is acceptable to combine studies that randomize individual participants with those that randomize clusters (e.g., when clinics, clinicians, or classrooms are randomized and individuals are nested within these units). We believe this is generally acceptable, with appropriate adjustment for cluster randomization as needed.[10] 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.

Outcome measures are a common source of methodological heterogeneity. First, trials may have a wide array of specific instruments and cut-points for a common outcome. For example, a review considering pooling the binary outcome of depression prevalence may find measures that range from a depression diagnosis based on a clinical interview to scores above a cut-point on a screening instrument. 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. In addition, investigators should take steps to harmonize outcomes to the extent possible.

Second, there is also typically substantial variability in the statistics reported across studies (e.g., odds ratios, relative risks, hazard ratios, 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 potentially less accurate results (due to assumptions required to estimate missing data) with the potential advantage of pooling a more complete set of studies. If a substantial proportion of the studies require calculations that involve assumptions or estimates (rather than straightforward calculations) in order to combine them, then it may be preferable to show results in a table or forest plot without a pooled estimate

1.3. Best Evidence Versus All Evidence

Sometimes the body of evidence comprises a single trial or small number of trials that clearly represent the best evidence, along with a number of additional trials that are much smaller or with other important limitations (Step B, Figure 1.1). The "best evidence" trials are generally very large trials with low risk of bias and with good generalizability to the population of interest. In this case, it may be appropriate to focus on the one or few "best" trials rather than combining them with the rest of the evidence, particularly when addressing rare events that small studies are underpowered to examine.[11, 12] For example, an evidence base of one large, multi-center trial of an intervention to prevent stroke in patients with heart disease could be preferable to a pooled

analysis of 4-5 small trials reporting few events, and combining the small trials with the large trial may introduce unnecessary uncertainty to the pooled estimate.

1.4. Assessing the Risk of Misleading Meta-analysis Results

Next, reviews should explore the risk that the meta-analysis will show results that do not accurately capture the true underlying effect (Step C, Figure 1.1). Tables, forest plots (without pooling), and some other preliminary statistical tests are useful tools for this stage. Several patterns can arise that should lead investigators to be cautious about combining studies.

Wide-Ranging Effect Sizes

Sometimes one study may show a large benefit and another study of the same intervention may show a small benefit. This may be due to random error, especially when the studies are small. However, this situation also raises the possibility that observed effects truly are widely variable in different subpopulations or situations. Another look at the population characteristics is warranted in this situation to see if the investigators can identify characteristics that are correlated with effect size and direction, potentially explaining clinical heterogeneity.

Even if no characteristic can be identified that explains why the intervention had such widely disparate effects, there could be unmeasured features that explain the difference. If the intervention really does have widely variable impact in different subpopulations, particularly if it is benefiting some patients and harming others, it would be misleading to report a single average effect.

Suspicion of Publication or Reporting Bias

Sometimes, due to lack of effect, trial results are never published (risking publication bias), or are only published in part (risking reporting bias). These missing results can introduce bias and reduce the precision of meta-analysis.[13] Investigators can explore the risk of reporting bias by comparing trials that do and do not report important outcomes to assess whether outcomes appear to be missing at random.[13] For example, investigators may have 30 trials of weight loss interventions with only 10 reporting blood pressure, which is considered an important outcome for the review. This pattern of results may indicate reporting bias as 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 investigators may decide to combine the studies reporting blood pressure that were conducted in high CVD risk populations. However, investigators should be clear about the applicable subpopulation. 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.

Small Studies Effect

If small studies show larger effects than large studies, the pooled results may overestimate the true effect size, possibly due to publication or reporting bias.[14] When investigators have at least 10 trials to combine they should examine small studies effects using standard statistical tests such as the Egger test.[15] If there appears to be a small studies effect, the investigators may decide not to report pooled results since they could be misleading. On the other hand, small studies effects could be happening for other reasons, such as differences in sample

characteristics, attrition, or assessment methods. These factors do not suggest bias, but should be explored to the degree possible. See Chapter 4 for more information about exploring heterogeneity.

1.5. Special Considerations When Pooling a Small Number of Studies

When pooling a small number of studies (e.g., <10 studies), a number of considerations arise (Step E, Figure 1.1):

Rare Outcomes

Meta-analyses of rare binary outcomes are frequently underpowered, and tend to overestimate the true effect size, so pooling should be undertaken with caution.[11] 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.[12] One example is all-cause mortality, which is frequently provided as part of the participant flow results, but may not be a primary 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. Investigators should consider calculating an optimal information size (OIS) when events are rare to see if the combined group of studies has sufficient power to detect group differences. This could be a concern even for a relatively large number of studies, if the total sample size is not very large.[16] See Chapter 3 for more detail on handling rare binary outcomes.

Small Sample Sizes

When pooling a relatively small number of studies, pooling should be undertaken with caution if the body of evidence is limited only to small studies. Results from small trials are less likely to be reliable than results of large trials, even when the risk of bias is low.[17] First, in small trials it is difficult to balance the proportion of patients in potentially important subgroups across interventions, and a difference between interventions of just a few patients in a subgroup can result in a large proportional difference between interventions. Characteristics that are rare are particularly at risk of being unbalanced in trials with small samples. In such situations there is no way to know if trial effects are due to the intervention or to differences in the intervention groups. In addition, patients are generally drawn from a narrower geographic range in small trials, making replication in other trials 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 lower risk of bias than are small trials. Therefore, when the trials have small sample sizes, pooled effects are less likely to reflect the true effects of the intervention. In this case, the required or optimal information size can help the investigators determine whether the sample size is sufficient to conclude that results are likely to be stable and not due to random heterogeneity (i.e., truly significant or truly null results; not a type I or type II error).[16, 18]An option in this case would be to pool the studies and acknowledge imprecision or other limitations when rating the strength of evidence.

What would be considered a “small” trial varies for different fields and outcomes. For addressing an outcome that only happens in 10% of the population, a small trial might be 100 to 200 per intervention arm, whereas a trial addressing a continuous quality of life measure may be

small with 20 to 30 per intervention. Looking carefully at what the studies were powered to detect and the credibility of the power calculations may help determine what constitutes a "small" trial. Investigators should also consider how variable the impact of an intervention may be over different settings and subpopulations when determining how to weigh the importance of small studies. For example, the effects of a counseling intervention that relies on patients to change their behavior in order to reap health benefits may be more strongly influenced by characteristics of the patients and setting than a mechanical or chemical agent.

Wide-Ranging Effect Sizes

When the number of trials to be pooled is small, there is a heightened risk that statistical heterogeneity will be substantially underestimated, resulting in 95% confidence intervals that are inappropriately narrow and do not have 95% coverage. This is especially concerning when the number of studies being pooled is fewer than five to seven.[19-21]

Accounting for these factors should guide an evaluation of whether it is advisable to pool the relatively small group of studies. As with many steps in the multi-stage decision to pool, the conclusion that a given investigator arrives at is subjective, although such evaluations should be guided by the criteria above. If consideration of these factors reassures investigators that the risk of bias associated with pooling is sufficiently low, then pooling can proceed. The next step of pooling, whether for a small, moderate, or large body of studies, is to consider statistical heterogeneity.

1.6. Statistical Heterogeneity

Once clinical and methodological heterogeneity and other factors described above have been deemed acceptable for pooling, investigators should next consider statistical heterogeneity (Step F, Figure 1.1). We discuss statistical heterogeneity in general in this chapter, and provide a deeper methodological discussion in Chapter 4. This initial consideration of statistical heterogeneity is accomplished by conducting a preliminary meta-analysis. Next the investigator must decide if the results of the meta-analysis are valid and should be presented, rather than simply showing tables or forest plots without pooled results. If statistical heterogeneity is very high, the investigators may question whether an "average" effect is really meaningful or useful. If there is a reasonably large number of trials, the investigators may shift to exploring effect modification with high heterogeneity, however this may not be possible if few trials are available. 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 judgment.

While there are a variety of methods for characterizing statistical heterogeneity, one common method is the I^2 statistic, the proportion of total variance in the pooled trials that is due to inter-study variance, as opposed to random variation.[22] The Cochrane manual proposes ranges for interpreting I^2:[10] 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. Ranges overlap to reflect that other factors—such as the number and size of the trials and the magnitude and direction of the effect—must be taken into consideration. Other measures of statistical heterogeneity include Cochrane's Q and τ^2, but these heterogeneity statistics do not have intrinsic standardized scales that allow specific values to be characterized as "small," "medium," or "large" in any meaningful way.[23]

However, τ^2 can be interpreted on the scale of the pooled effect, as the variance of the true effect. All these measures are discussed in more detail in Chapter 4.

Although widely used in quantitative synthesis, the I^2 statistic has come under criticism in recent years. One important issue with I^2 is that it can be an inaccurate reflection of statistical heterogeneity when there are few studies to pool and high statistical heterogeneity.[24, 25] For example, 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 I^2 is 50% or more.[26] In addition, I^2 is correlated with the sample size of the included studies, generally increasing with larger samples.[27] 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.[28, 29] 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.[29] 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. Thus, I^2 is one piece of information that should be considered, but generally should not be the primary deciding factor for whether to pool.

1.7. Conclusion

In the end, the decision to pool boils down to the question: will the results of a meta-analysis help you find a scientifically valid 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? Further, do the clinical, methodological, and statistical features of the body of studies permit them to be quantitatively combined and summarized in a valid fashion? Each of these decisions can be broken down into specific considerations (outlined in Figure 1.1) There is broad guidance to inform investigators in making each of these decisions, but generally the choices involved are subjective. The investigators' scientific goal might factor into the evaluation of these considerations: for example, if investigators seek a general summary of the combined effect (e.g., direction only) versus an estimated effect size, the consideration of whether to pool may be weighed differently. In the end, 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.

Recommendations

- Use Figure 1.1 when deciding whether to pool studies

Chapter 2. Optimizing Use of Effect Size Data

Christopher S. Lee, Ph.D., R.N., Benjamin W. Vandermeer, M.Sc.

2.1. Introduction

The employed methods for meta-analysis will depend upon the nature of the outcome data. The two most common data types encountered in trials are binary/dichotomous (e.g., dead or alive, patient admitted to hospital or not, treatment failure or success, etc.) and continuous (e.g., weight, systolic blood pressure, etc.). Some outcomes (e.g., heart rate, counts of common events) that are not strictly continuous, are often treated as continuous for the purposes of meta-analysis based on assumptions of normality and the belief that statistical methods that are applied to normal distributions can be applicable to other distributions (central limit theory). Continuous outcomes are also frequently analyzed as binary outcomes when there are clinically meaningful cut-points or thresholds (e.g., a patient's systolic blood pressure may be classified as low or high based on whether it is under or over 130mmHG). While this type of dichotomization may be more clinically meaningful it reduces statistical information, so investigators should provide their rationale for taking this approach.

Other less common data types that do not fit into either the binary or continuous categories include ordinal, categorical, rate, and time to event to data. Meta-analyzing these types of data will usually require reporting of the relevant statistics (e.g., hazard ratio, proportional odds ratio, incident rate ratio) by the study authors.

2.2. Nuances of Binary Effect Sizes

Data Needed for Binary Effect Size Computation

Under ideal circumstances, the minimal data necessary for the computation of effect sizes of binary data would be available in published trial documents or from original sources. 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. This approach will facilitate 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 in Treatment Group	Treatment n	Events in Control Group	Control n
Study X	5	25	6	25
Study Y	23	194	21	189

In many instances, a single study (or subset of studies) to be included in the meta-analysis provides only one measure of association (an odds ratio, for example), and the sample size and

event counts are not available. In that case, the meta-analytic effect size will be dictated by the available data. However, choosing the appropriate effect size is important for integrity and transparency, and every effort should be made to obtain all the data presented in Table 2.1. Note that CONSORT guidance requires that published trial data should include the number of events and sample sizes for both treatment and control groups.[30] And, PRISMA guidance supports describing any processes for obtaining and confirming data from investigators[31] – a frequently required step.

In the event that data are only available in an effect size from the original reports, it is important to extract both the mean effect sizes and the associated 95% confidence intervals. Having raw event data available as in Table 2.1 not only facilitates the computation of various effect sizes, but also allows for the application of either binomial (preferred) or normal likelihood approaches;[32] only normal likelihood can be applied to summary statistics (e.g., an odds ratio and confidence interval in the primary study report).

Choosing Among Effect Size Options

One absolute measure and two relative measures are commonly used in meta-analyses involving binary data. The RD (an absolute measure) is a simple metric that is easily understood by clinicians, patients, and other stakeholders. The relative measures, RR or OR, are also used frequently. All three metrics should be considered additive, just on different scales. That is, RD is additive on a raw scale, RR on a log scale, and OR on a logit scale.

Risk Difference

The RD is easily understood by clinicians and patients alike, and therefore most useful to aid decision making. However, the RD tends to be less consistent across studies compared with relative measures of effect size (RR and OR). Hence, the RD may be a preferred measure in meta-analyses when the proportions of events among control groups are relatively common and similar across studies. When events are rare and/or when event rates differ across studies, however, the RD is not the preferred effect size to be used in meta-analysis because combined estimates based on RD in such instances have more conservative confidence intervals and lower 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

Number Needed To Treat Related to Risk Difference

The number needed to treat (NNT) represents the number of patients that need to receive the treatment for one to benefit.[33] Because this is conceptually straightforward, this statistic resonates with clinicians and lay stakeholders. The NNT is the inverse of the risk difference, calculated as:

$$NNT = \frac{1}{|RD|}$$

Where,
NNT = number needed to treat
RD = risk difference

In case of a negative RD, the number needed to harm (NNH) or number needed to treat for one patient to be harmed is = − 1/RD.

The Wald method[34] is commonly used to calculate confidence intervals for NNT. It is reasonably adequate for large samples and probabilities not close to either 0 or 1, however it can be less reliable for small samples, probabilities close to either 0 or 1, or unbalanced trial designs.[35] An adjustment to the Wald method (i.e., adding pseudo-observations) helps mitigate concern about its application in small samples,[36] but it doesn't account for other sources of limitations to this method. The Wilson method of calculating confidence intervals for NNT, as described in detail by Newcome,[37] has better coverage properties irrespective of sample size, is

free of implausible results, and is argued to be easier to calculate compared with Wald confidence intervals.[35] *Therefore, the Wilson method is preferable to the Wald method for calculating confidence intervals for NNT. When considering using NNT as the effect size in meta-analysis, see commentary by Lesaffre and Pledger.38 When considering using NNT as the effect size in meta-analysis, see commentary on the superior performance of combined NNT on the RD scale as opposed to the NNT scale.*

Risk Ratio

It is important to note that the RR and OR are effectively equivalent for event rates below about 10%. In such cases, the RR is chosen over the OR simply for interpretability (an important consideration) and not substantive differences. 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 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 OR. If switching between death and survival, for example, is 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 possible. Moreover, investigators should be particularly attentive to the definition of an outcome event when using a RR.

The calculation of RR using binary data can be performed considering the labeling listed in Table 2.2. Of particular 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

$$RR = \frac{A/n_1}{C/n_2}$$
$$ln_{RR} = \ln(RR)$$
$$V_{ln_{RR}} = \frac{1}{A} + \frac{1}{C} - \frac{1}{n_1} - \frac{1}{n_2}$$
$$SE_{ln_{RR}} = \sqrt{V_{ln_{RR}}}$$
$$LL_{ln_{RR}} = ln_{RR} - 1.96 * SE_{ln_{RR}}$$
$$UL_{ln_{RR}} = ln_{RR} + 1.96 * SE_{ln_{RR}}$$
$$RR = \exp(ln_{RR})$$
$$\text{LL of the 95\%CI} = \exp(LL_{ln_{RR}})$$
$$\text{UL of the 95\%CI} = \exp(UL_{ln_{RR}})$$

Where,
RR = risk ratio
ln_{RR} = natural log of the risk ratio
V_{lnRR} = variance of the natural log of the risk ratio
SE_{lnRR} = standard error of the natural log of the risk ratio
LL_{lnRR} = lower limit of the 95% confidence interval of the natural log of the risk ratio
UL_{lnRR} = 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 an RR.

Odds Ratios

An alternative relative metric for use with binary data is the OR. Given that ORs are frequently presented in models with covariates, it is important to note that the OR is 'non-collapsible,' meaning that effect modification varies depending on the covariates for which control has been made; this favors the reporting of RR over OR, particularly when outcomes are common and covariates are included.[39] The calculation of OR using binary data 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

$$OR = \frac{AD}{BC}$$

$$\ln_{OR} = \ln(OR)$$
$$V_{\ln_{OR}} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$
$$SE_{\ln_{OR}} = \sqrt{V_{\ln_{OR}}}$$
$$LL_{\ln_{OR}} = \ln_{OR} - 1.96 * SE_{\ln_{OR}}$$
$$UL_{Log_{OR}} = \ln_{OR} + 1.96 * SE_{\ln_{OR}}$$
$$OR = \exp(\ln_{OR})$$
$$\text{LL of the 95\%CI} = \exp(LL_{\ln_{OR}})$$
$$\text{UL of the 95\%CI} = \exp(UL_{\ln_{OR}})$$

Where,
OR = odds ratio
Ln_{OR} = natural log of the odds ratio
V_{lnOR} = variance of the natural log of the odds ratio
SE_{lnoR} = standard error of the natural log of the odds ratio
LL_{lnOR} = lower limit of the 95% confidence interval of the natural log of the odds ratio
UL_{lnOR} = 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 event rates are less

than 1%. But in contrast, the Peto OR is biased when treatment effects are large, due to centering around the null hypothesis, and in the instance of imbalanced treatment and control groups.[40]

Equation Set 2.4. Peto odds ratios

$$\mathrm{ORpeto} = \exp[\{A - E(A)\}/v]$$

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**. In the table, the term "baseline risk" is the proportion of subjects in the control group who experienced the event. The term "control rate" is sometimes used for this measure as well.

Table 2.3. Benefits and disadvantages of binary data effect sizes[41]

Effects Size	Benefits	Disadvantages
Risk Difference	- may be more easily interpretable among lay audiences - on the familiar percentage scale - can be converted to NNT or NNH for clinical interpretability - can address zero-event studies[42]	- not consistent between studies with differing baseline risks. - not commonly reported in individual trials. - not preferred when there is heterogeneity between studies in duration and incident rates[43]
Relative Risk	- easily interpretable - commonly reported in individual trials considered in meta-analyses - more likely to be consistent even with differing baseline risks	- values of "death" and "survival" are not reciprocals of each other as would be intuitively expected. - dependent on arbitrary definition of event versus no event.
Odds Ratio	- more likely to be 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 rather than an assessment of appropriateness

Time-to-Event and Count Outcomes

For time to event data, the effect size measure is a hazard ratio (HR), which is commonly estimated from the Cox proportional hazards model. In the best-case scenario, HR and associated 95% confidence intervals are available from all studies, the time horizon is 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, an HR and associated dispersion can still be extracted and meta-analyzed. However, this approach raises concerns about reproducibility due to observer variation.[44]

Incident rate ratio (IRR) is used for count data and can be estimated from a Poisson or negative binomial regression model. The IRR is a relative metric based on counts of events (e.g., number of hospitalizations, or days of length of stay) over time (i.e., per person-year) compared between trial arms. 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 as these modeling decisions can be sources of between-study heterogeneity. Moreover, studies that include count data may have zero counts in both groups, which may require less common and more nuanced approaches to meta-analysis like Poisson regression with random intervention effects.[45]

2.3. Continuous Outcomes

Assembling Data Needed for Effect Size Computation

Meta-analysis of studies presenting continuous data requires both estimated differences between the two groups being compared and estimated standard errors of those differences. Estimating the between-group difference is easiest when the study provides the mean difference. While both a standardized mean difference and ratio of means could be given by the study authors, studies more often report means for each group. Thus, a mean difference or ratio of means often must be computed.

If estimates of the standard errors of the mean are not provided studies commonly provide confidence intervals, standard deviations, p-values, z-statistics, and/or t-statistics, which make it possible to compute the standard error of the mean difference. In the absence of any of these statistics, other methods are available to estimate standard error.[45]

(Weighted) Mean Difference

The mean difference (formerly known as weighted mean difference) 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 measures the outcome on the same scale or on scales that can be easily converted. For example, total weight can be pooled using mean difference even if different studies reported weights in kilograms and pounds; however it is not 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 the mean difference is straightforward and 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 as mentioned above, other pieces of data may also be used.

Some studies report values as change from baseline, or alternatively present both baseline and final values. In these cases, it is possible to pool differences in final values in some studies with differences in change from baseline values in other studies, since they will be estimating the same value in a randomized control trial. 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 the most common response is to compute a standardized mean difference (SMD) for each study and then pool these across all studies in the meta-analysis. By dividing the mean difference by a pooled estimate of the standard deviation, we theoretically put all scales in the same unit (standard deviation), and are then able to statistically combine all the studies. While the standardized mean difference could be used even when studies use the same metric, it is generally preferred to use mean difference. Interpretation of results is easier when the final pooled estimate is given in the same units as the original studies.

Several methods can compute SMDs. The most frequently used are Cohen's *d* and Hedges' g.

Cohen's d

Cohen's *d* is the simplest S. computation; it is defined as the mean difference divided by the pooled standard deviation of the treatment and control groups.[5] For a given study, Cohen's *d* can be computed as:

$$d = \frac{m_T - m_C}{s_{pooled}}$$

Where m_T and m_C are the treatment and control means and s_{pooled} is essentially the square root of the weighted average of the treatment and control variances.

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).[46] For this reason, Hedges g is more often used.

Hedges' g

Hedges' g is a transformation of Cohen's d that attempts to adjust for small sample bias. 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 with Cohen's d, but the reduction lessens as the total sample size increases. The formula is:

$$d\left(1 - \frac{3}{4N - 9}\right)$$

Where *N* is the total trial sample size.

For very large sample sizes the two estimates will be very similar.

Back Transformation of Pooled SMD

One disadvantages of reporting standardized mean difference is that units of standard deviation are difficult to interpret clinically. Guidelines do exist but are often thought to be arbitrary and not applicable to all situations.47 An alternative is to back transform the pooled SMD into a scale 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. Alternatives for estimation include using the standard deviation from the largest study or using a pooled estimate of the standard deviations across studies.5 One should include a sensitivity analysis and be transparent about the approach used.

Ratio of Means

Ratio of Means (RoM), also known as response ratio, 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.

2.4. 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—a result of the two measurements on different within-person treatments.[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. If correlation is available, the pooled standard error can be computed using the following formula:

$$SE_P = \sqrt{SE_T^2 + SE_C^2 + 2rSE_TSE_C}$$

Where r is the within-patient correlation and SE_P, SE_T, and SE_C are the pooled, treatment, and control standard errors respectively

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] An imputation of 0.5 has

been suggested as a good conservative estimate of correlation in the absence of any other information.[48]

If a cross-over study reports its data by period, investigators have sometimes used first period data only when including cross-over trials in their meta-analyses—essentially treating the study as if it were a parallel design. This eliminates correlation issues, but has the disadvantage of omitting half the data from the trial.

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 of the same size. The design effect (DE) of a cluster randomized trial is the multiplicative multiplier needed to adjust the standard error computed as if the trial were a standard parallel design. Reported results from cluster trials may not reflect the design effect, and thus it will need to be computed by the investigator. The formula for computing the design effect is:

$$DE = 1 + (M - 1)ICC$$

Where M is the average cluster size and ICC is the intra-class correlation coefficient (see below).

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 followup data.
2. Use change from baseline data.
3. Use an ANCOVA model that adjusts for the effects of baseline imbalance.[49]

As long as trials are balanced at baseline, all three methods will give similar unbiased estimates of mean difference.[5] When baseline imbalance is present, it can be shown that using 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.[50] 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 the correlation is less than or equal to 0.5, then using the follow up values will be less biased (with respect to the estimate in the ANCOVA model) than using the change from baseline values. If the correlation is greater than 0.5, then change from baseline values will be less biased than using the follow up values.[51] 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.[52] A recent study[51] showed that all approaches were unbiased when there were both few trials and small sample sizes within the trials.

Recommendations

- For binary outcomes:
  - The analyst should consider carefully which binary measure to analyze.
  - If conversion to NNT or NNH is sought, then the risk difference is the preferred measure.
  - The risk ratio and odds ratio are likely to be more consistent than the risk difference when the studies differ in baseline risk.
  - The risk difference is not the preferred measure when the event is rare.
  - The risk ratio is not the preferred measure if switching between occurrence and non-occurrence of the event is important to the meta-analysis.
  - The odds ratio can be misleading.
- For continuous outcomes:
  - The mean difference is the preferred measure when studies use the same metric.
  - When calculating standardized mean difference, Hedges' g is preferred over Cohen's d due to the reduction in bias.
- General:
  - If baseline values are unbalanced, one should perform an ANCOVA analysis. If ANCOVA cannot be performed and the correlation is greater than 0.5, change from baseline values should be used to compute the mean difference. If the correlation less than or equal to 0.5, follow-up values should be used.
  - Data from clustered randomized trials should be adjusted for the design effect.

Chapter 3. Choice of Statistical Model for Combining Studies

Rongwei Fu, Ph.D., Marika Booth, M.S., Dale W. Steele, M.D., M.S.

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 and study and follow a random distribution. The differences between observed effect sizes are not only due to sampling error, but also to variation in the 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 distributions[53] or distribution free models[54, 55] have also been proposed.

Recent advances in meta-analysis include the development of alternative models to the fixed or random effects models. 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.[56] Stanley and Doucouliagosb proposed an unrestricted weighted least squares (WLS) estimator with multiplicative error for meta-analysis and claimed superiority to both conventional fixed and random effects,[57] though Mawdsley et al.[58] found modest differences when compared with the random effects model. These methods have not been fully compared with the many estimators developed within the framework of the fixed and random effects models and are not readily available in most statistical packages; thus they will not be further considered here.

General Considerations for Model Choice

Considerations for model choice include but are not limited to heterogeneity across treatment effects, the number and size of included studies, the type of outcomes, and potential bias. We recommend against choosing a statistical model based on the significance level of a heterogeneity test, for example, picking a fixed effects model when the p-value for the test of heterogeneity is more than 0.10 and a random effects model when $P < 0.10$, since such an approach does not take the many factors for model choice into full consideration.

In practice, clinical and methodological heterogeneity 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 special considerations for rare binary outcomes (discussed below in the section on combining rare binary outcomes). For a binary outcome, when the estimate of between-study heterogeneity is zero, a fixed effects model (e.g., the Mantel-Haenszel method, inverse variance method, Peto method (for OR), or fixed effects logistic regression) provides an effect estimate similar to that produced by a random effects model. The 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 systematic review includes 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 of effect sizes, in particular, a normal distribution, is

not justified. In this case, neither the random effects model nor the fixed effects model provides an appropriate estimate and investigators may choose not to combine all studies.[10] Investigators can choose to combine only the large studies if they are well conducted with good quality and are expected to provide unbiased effect estimates. Other potential differences between small and large studies should also be examined.

Choice of Random Effects Model and Estimator

The most commonly used random effects model for combined effect estimates is based on an estimator developed by DerSimonian and Laird (DL) due to its simplicity and ease of implementation.[59] 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. [19, 60, 61] 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 the t-distribution and an adjusted confidence interval for the overall effect. [62-64] We refer to this as the HKSJ method. Biggerstaff–Tweedie (BT) proposed another variant of the DL method by incorporating error in the point estimate of between-study heterogeneity into the estimation of the overall effect.[65] There are also many other likelihood based estimators such as maximum likelihood estimate, restricted maximum likelihood estimate and profile likelihood (PL) methods, which better account for the uncertainty in the estimate of between-study variance.[19]

Several simulation studies have been conducted to compare the performance of different estimators for combined effect size.[19-21, 66, 67] For example, Brockwell et al. showed the PL method provides an estimate with better coverage probability than the DL method.[19] Jackson et al. showed that with a small number of studies, the DL method did not provide adequate coverage probability, in particular, when there was moderate to large heterogeneity.[20] However, these results supported the usefulness of the DL method for larger samples. In contrast, the PL estimates resulted in coverage probability closer to nominal values. IntHout et al. compared the performance of the DL and HKSJ methods and showed that the HKSJ method consistently resulted in more adequate error rates than did the DL method, especially when the number of studies was small, though they did not evaluate coverage probability and power.[67] Kontopantelis and Reeves 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). [21] When the goal is to obtain an accurate estimate of overall effect size and the associated error interval, they recommended using the DL method when heterogeneity is low and using the PL method when heterogeneity is high, where the definition of high heterogeneity varies by the number of studies. The PL method overestimated coverage probability in the absence of between-study heterogeneity. Methods like BT and HKSJ, despite being developed to address the limitations of the DL method, were frequently outperformed by the DL method. 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.[68] 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.[21] It is appropriate to use the DL method when the heterogeneity is low. Another disadvantage of the PL method is that it does not always converge. In those situations, investigators may choose the DL method with sensitivity analyses using other methods, such as the HKSJ method. If non-convergence is due to high heterogeneity, investigators should also reevaluate the appropriateness of combining 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 events are common. Note that the confidence interval produced by the PL method may not be symmetric. It is also worth noting that OR, RR, HR, and incidence rate ratio statistics should be analyzed on the logarithmic scale when the PL, DL, or HKSJ method is used. Finally, a Bayesian approach can also be used since this approach takes the variations in all parameters into account (see the section on Bayesian methods, below).

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 recommend the use of the generalized linear mixed effects model assuming random treatment effects. For aggregated binary data, a combined OR can 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. For aggregated count data, a combined rate ratio can be generated by assuming the Poisson distribution with a log link. Results from using the generalized linear models and directly combining effect measures are similar when the number of studies and/or the sample sizes are large.

3.2. A Special Case: Combining Rare Binary Outcomes

When combining rare binary outcomes (such as adverse event data), few or zero events often occur in one or both arms in some of the studies. In this case, the binomial distribution is not well-approximated by the normal approximation and choosing an appropriate model becomes complicated. The DL method does not perform well with low-event rate binary data.[43, 69] A fixed effects model often out performs the DL method even in the presence of heterogeneity.[70] When event rates are less than 1 percent, the Peto OR method has been shown to provide the least biased, most powerful combined estimates with the best confidence interval coverage,[43] 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).[71, 72] Otherwise, when treatment and control group sizes are very different, effect sizes are large, or when events become more frequent (5 percent to 10 percent), the Mantel-Haenszel method (without a correction factor) or a fixed effects logistic regression provide better combined estimates.

Within the past few years, many methods have been proposed to analyze sparse data from simple averaging,[73] exact methods,[74, 75] Bayesian approaches[76, 77] 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 not use continuity corrections, and to include studies with zero events in both arms in the meta-analysis. Great efforts have been made to develop methods that can include such studies.

Bhaumik et al. proposed the simple (unweighted) average (SA) treatment affect with the 0.5 continuity correction, and found that the bias of the SA estimate in the presence of even significant heterogeneity is minimal compared with the bias of MH estimates (with 0.5 correction).[73] A simple average was also advocated by Shuster.[78] However, potential confounding remains an issue for an unweighted estimator. Spittal et al. showed that Poisson regression works better than the inverse variance method for rare events.[79] Kuss et al. conducted a comprehensive simulation of eleven methods, and recommended the use of the beta-binomial model for the three common effect measures (OR, RR, and RD) as the preferred meta-analysis methods for rare binary events with studies of zero events in one or both arms.[80] The beta-binomial model assumes that the observed events follow a binomial distribution and the binomial probabilities follow a beta distribution. In Kuss's simulation, using a generalized linear model framework to model the treatment effect, an OR was estimated using a logit link, and an RR, using a log link. Instead of using an identity link, RD was estimated based on the estimated event probabilities from the logit model. This comprehensive simulation 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.

Given the development of new methods that can handle studies with zero events in both arms, we advise that older methods that use continuity corrections 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 packages. When the between-study heterogeneity is large and/or 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. Statistical methods can never completely solve the issue of sparse data. Investigators should always conduct sensitivity analyses[81] using alternative methods to check the robustness of results to different methods, and acknowledge the inadequacy of data sources when presenting the meta-analysis results, in particular, when the proportion of studies with zero events in both arms are high. 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 an Exact Method for Sparse Binary Data

For rare binary events, the normal approximation and asymptotic theory for large sample size does not work satisfactorily and exact inference has been developed to overcome these limitations. Exact methods do not need continuity corrections. However, simulation analyses do not identify a clear advantage of early developed exact methods[75, 82] over a logistic regression or the Mantel-Haenszel method even in situations where these exact methods would theoretically be advantageous.[43] Recent developments of exact methods include Tian et al.'s method of combining confidence intervals[83] and Liu et al.'s method of combining p-value functions.[84] Yang et al.[85] 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,[84] though the comparative performance

of these methods has not been thoroughly evaluated. Investigators may choose to use exact methods with considerations for the interpretation of effect measures, but we do not specifically recommend exact methods over other models discussed above.

3.3. Bayesian Methods

A Bayesian framework provides a unified and comprehensive approach to meta-analysis that accommodates a wide variety of outcomes, often, using generalized linear model (GLM) with normal, binomial, Poisson and multinomial likelihoods and various link functions.[86]

It should be noted that while these GLM models are routinely implemented in the frequentist framework, and are not specific to the Bayesian framework, extensions to more complex situations 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.[87]

There are several specific advantages inherent to the Bayesian framework. First, the Bayesian posterior parameter distributions fully incorporate the uncertainty of all parameters. These posterior distributions need not be assumed to be normal.[88] In random-effects meta-analysis, standard methods use only the most likely value of the between-study variance,[59] rather than incorporating the full uncertainty of each parameter. Thus, Bayesian credible intervals will tend to be wider than confidence intervals produced by some classical random-effects analysis such as the DL method.[89] However, when the number of studies is small, the between-study variance will be poorly estimated by both frequentist and Bayesian methods, and the use of vague priors can lead to a marked variation in results,[90] particularly when the model is used to predict the treatment effect in a future study.[91] A natural alternative is to use an informative prior distribution, based on observed heterogeneity variances in other, similar meta-analyses.[92-94]

Full posterior distributions can provide a more informative summary of the likely value of parameters than the frequentist approach. When communicating results of meta-analysis to clinicians, the Bayesian framework allows direct probability statements to be made and provides the rank probability that a given treatment is best, second best, or worst (see the section on interpreting ranking probabilities and clinically important results in Chapter 5 below). Another advantage is that posterior distributions of functions of model parameters can be easily obtained such as the NNT.[86] Finally, the Bayesian approach allows full incorporation of parameter uncertainty from meta-analysis into decision analyses.[95]

Until recently, Bayesian meta-analysis required specialized software such as WinBUGS,[96] OpenBUGS,[97] and JAGS.[98, 99] Newer open source software platforms such as Stan[100] and Nimble[101, 102] provide additional functionality and use BUGS-like modeling languages. In addition, 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.[103] For example, in R, the package *bmeta* currently generates JAGS code to implement 22 models.[104] The R package *gemtc* similarly automates generation of JAGS code and facilitates assessment of model convergence and inconsistency.[105, 106] On the other hand, Bayesian meta-analysis can be implemented in commonly used statistical packages. For example, SAS PROC MCMC can now implement at least some Bayesian hierarchical models[107] directly, as can Stata, version 14, via the *bayesmh* command.[108]

When vague prior distributions are used, Bayesian estimates are usually similar to estimates obtained from the above frequentist methods.[90] Use of informative priors requires considerations to avoid undue influence on the posterior estimates. Investigators should provide adequate

justifications for the choice of priors and conduct sensitivity analyses. Bayesian methods currently require more work in programming, MCMC simulation and convergence diagnostics.

A Note on Using a Bayesian Approach for Sparse Binary Data

It has been suggested that using a Bayesian approach might be a valuable alternative for sparse event data since Bayesian inference does not depend on asymptotic theory and takes into account all uncertainty in the model parameters.[109] The Bayesian fixed effects model provides good estimates when events are rare for binary data.[70] However, the choice of prior distribution, even when non-informative, may impact results, in particular, when a large proportion of studies have zero events in one or two arms.[80, 90, 110] Nevertheless, other simulation studies found that when the overall baseline rate is very small and there is moderate or large heterogeneity, Bayesian hierarchical random effect models can provide less biased estimates for the effect measures and the heterogeneity parameters.[77] To reduce the impact of the prior distributions, objective Bayesian methods have been developed[76, 111] with special attention paid to the coherence between the prior distributions of the study model parameters and the meta-parameter,[76] though the Bayesian model was developed outside the usual hierarchical normal random effects framework. Further evaluations of these methods are required before recommendations of these objective Bayesian methods might be made.

3.4 Recommendations

- The PL method appears to generally perform best. The DL method is also appropriate when the between-study heterogeneity is low.
- For study-level aggregated binary data and count data, the use of a generalized linear mixed effects model assuming random treatment effects is also recommended.
- For rare binary events,
  - Methods that use continuity corrections should be avoided.
  - 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 low.
  - When the between-study heterogeneity is high, and/or there are studies with zero events in both arms, more recently developed methods such as a beta-binomial model could be explored and used.
  - Sensitivity analyses should be conducted with acknowledgement of the inadequacy of data.
- If investigators choose Bayesian methods, use of vague priors is supported.

Chapter 4. Quantifying, Testing, and Exploring Statistical Heterogeneity

Christopher S. Lee, Ph.D., R.N.

4.1. Statistical Heterogeneity in Meta-analysis

Statistical heterogeneity was explained in general in Chapter 1. In this chapter, we provide a deeper discussion from a methodological perspective. Statistical heterogeneity must be expected, quantified and sufficiently addressed in meta-analyses.[112] We recommend performing graphic and quantitative exploration of heterogeneity in combination.[113] In this chapter, it is assumed that a well-specified research question has been posed, the relevant literature has been reviewed, and a set of trials meeting selection criteria have been identified. Even when trial selection criteria are aimed toward identifying studies that are adequately homogenous, it is common for trials included in a meta-analysis to differ considerably as a function of (clinical and/or methodological) heterogeneity that was reviewed in Chapter 1. Even when these sources of heterogeneity have been accounted for, statistical heterogeneity often remains. Statistical heterogeneity refers to the situation where estimates across studies have greater variability than expected from chance variation alone.[113, 114]

4.2. Visually Inspecting Heterogeneity

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. However, forest and funnel plots can be helpful in the interpretation of heterogeneity particularly when examined in combination with quantitative results.[113, 115]

Forest Plots

Forest plots can help identify potential sources and the extent of statistical heterogeneity. Meta-analyses with limited heterogeneity will produce forest plots with grossly visual overlap of study confidence intervals and the summary estimate. In contrast, a crude sign of statistical heterogeneity would be poor overlap.[115] An important recommendation is to graphically present between-study variance on forest plots of random effects meta-analyses using prediction intervals, which are on the same scale as the outcome.[93] The 95% prediction interval estimates where true effects would be expected for 95% of future studies.[93] When between-study variance is greater than zero, the prediction interval will cover a wider range than the confidence interval of the summary effect.[116] As proposed by Guddat et al.[117] and endorsed by IntHout et al.,[116] including the prediction interval as a rectangle at the bottom of forest plots helps differentiate between-study variation from the confidence interval of the summary effect that is commonly depicted as a diamond.

Funnel Plots

Funnel plots are often thought of as representing bias, but they also can aid in detecting sources of heterogeneity. Funnel plots are essentially the plotting of effect sizes observed in each study (x-axis) around the summary effect size versus 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.[115] 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 the bounds of the 95% confidence limits.[115] Asymmetry in funnel plots can be difficult to detect visually,[118] and can be misleading due to multiple contributing factors.[113, 119, 120] Formal tests for funnel plot asymmetry (such as Egger's test[15] for continuous outcomes, or the arcsine test proposed by Rucker et al.,[27] for binary data) are available but should not be used with a meta-analysis involving fewer than 10 studies because of limited power.[113] Given the above cautions and considerations, funnel plots should only be used to complement other approaches in the preliminary analysis of heterogeneity.

4.3. Quantifying Heterogeneity

The null hypothesis of homogeneity in meta-analysis is that all studies are evaluating the same effect,[22] (i.e., all studies have the same true effect parameter that may or may not be equivalent to zero) and the alternative hypothesis is that at least one study has an effect that is different from the summary effect.

A commonly-used heterogeneity test statistic is *Q*,[59] which is computed as the sum of squared deviations of each study's estimate from the summary estimate, with each study's contribution weighted in the same manner as in the meta-analysis (commonly via the inverse variance method):[121]

$$Q = \sum_{i=1}^{k} w_i (x_i - \hat{x}_w)^2$$

Where *Q* is the heterogeneity statistic,
w is the study weight based on inverse variance weighting,
x is the observed effect size in each trial, and
$\hat{x}_W$ is the summary estimate in a fixed-effect meta-analysis.

The *Q* statistic is assumed to have an approximate χ^2 distribution with $k - 1$ degrees of freedom. When *Q* is in excess over $k - 1$ and the associated p-value is low (typically, a p-value of <0.10 is used as a cut-off), the null hypothesis of homogeneity can be rejected.[22, 122] 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[123, 124] and may detect unimportant heterogeneity when the number of studies included in a meta-analysis is large. Importantly, since heterogeneity is expected in meta-analyses even without statistical tests to support that claim, non-significant *Q* statistics must not be interpreted as the absence of heterogeneity. Moreover, the interpretation of *Q* in meta-analyses is more complicated than typically represented, because the actual distribution of *Q* is dependent on the measure of effect[125] and only approximately χ^2 in large samples.[122] Even if the null distribution of *Q* were χ^2, universally interpreting all values of *Q* greater than the mean of $k - 1$ as indicating heterogeneity would be an oversimplification.[122] There are expansions to approximate *Q* for meta-analyses of standardized mean difference,[125] risk difference,[125] and odds ratios[126] that should be used as alternatives to *Q*, particularly when sample sizes of studies included in a meta-analysis are small.[122] The *Q* statistic and expansions

thereof must be interpreted along with other heterogeneity statistics and with full consideration of their limitations.

Graphical Options for Examining Contributions to Q

Hardy and Thompson proposed using probability plots to investigate the contribution that each study makes to *Q*.[127] When each study is labeled, those deviating from the normal distribution in a probability plot have the greatest influence on *Q*.[127] Baujat and colleagues proposed another graphical method to identify studies that have the greatest impact on *Q*.[128] Baujat proposed plotting the contribution to the heterogeneity statistic for each study 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 the Baujat plot presentation, studies that have the greatest influence on *Q* are located in the upper right corner for easy visual identification. Smaller studies have been shown to contribute more to heterogeneity than larger studies,[129] which would be visually apparent in Baujat plots. We recommend using these graphical approaches only when there is significant heterogeneity, and only when it is important to identify specific studies that are contributing to heterogeneity.

Between-Study Variance

DerSimonian and Laird proposed a non-iterative method-of-moments parameter of between-study variance (τ^2)[60] (described by Higgins et al. as "among-study variance")[22] that remains widely used in meta-analyses:

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

Where τ^2 is the parameter of between-study variance of the true effects,
DL is the DerSimonian and Laird approach to τ^2,
Q is the heterogeneity statistic (as above),
k -1 is the degrees of freedom, and
w is the weight applied to each study based on inverse variance weighting.

Since variance cannot be less than zero, a τ^2 less than zero is set to zero. The value of τ^2 is integrated into the weights of random-effects meta-analysis as presented in Chapter 3. Since the DerSimonian and Laird approach to τ^2 is derived in part from *Q*, the problems with *Q* described above apply to the τ^2 parameter.[122] There are many alternatives to DerSimonian and Laird when estimating between-study variance. In a recent simulation, Veroniki and colleagues[121] compared 16 estimators of between-study variance; they argued that the Paule and Mandel[130] method of estimating between-study variance is a better alternative to the DerSimonian and Laird parameter for continuous and binary data because it less biased (i.e., yields larger estimates) when between-study variance is moderate-to-large.[121] At the time of this guidance, the Paule and Mandel method of estimating between-study variance is only provisionally recommended as an alternative to DerSimonian and Laird.[129, 131] Moreover, Veroniki and colleagues provided evidence that the restrictive maximum likelihood estimator[132] is a better alternative to the DerSimonian and Laird parameter of between-study variance for continuous data because it yields similar values for low-to-moderate between-study variance and larger estimates in conditions of high between-study variance.[121]

Inconsistency Across Studies

Another statistic that should be generated and interpreted even when *Q* is not statistically significant is the proportion of variability in effect sizes across studies that is explained by heterogeneity vs. random error or I^2 that is related to *Q*.[22, 133]

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

Where *Q* is the estimate of between-study variance, and
k -1 is the degrees of freedom.

For random-effects models, Higgins and Thompson[25] proposed estimating I^2 as:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Where τ^2 is the parameter of between-study variance, and
σ^2 is the within-study variance.

I^2 is a 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 attributable to heterogeneity), the I^2 statistic has several advantages over other heterogeneity statistics including its relative simplicity as a signal-to-noise ratio, and focus on how heterogeneity may be influencing interpretation of the meta-analysis.[59] It is important to note that I^2 increases with increasing study precision and hence is dependent on sample size.[27] By various means, confidence/uncertainty intervals can be estimated for I^2 including Higgins' test-based method.[22, 23] the assumptions involved in the construction of 95% confidence intervals cannot be justified in all cases, but I^2 confidence intervals based on frequentist assumptions generally provide sufficient coverage of uncertainty in meta-analyses.[133] In small meta-analyses, it has even been proposed that confidence intervals supplement or replace biased point estimates of I^2.[26] It is important to note that since I^2 is based on *Q* or τ^2, any problems that influence *Q* or τ^2 (most notably the number of trials included in the meta-analysis) will also indirectly interfere with the computation of I^2. It is also important to consider that I^2 also is dependent on which between-study variance estimator is used. For example, there is a high level of agreement comparing I^2 derived from DerSimonian and Laird vs. Paul and Mandel methods of estimating between-study variance.[131] In contrast, I^2 derived from other methods of estimating between-study variance have low levels of agreement.[131]

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.[59] 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).[10] 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),[24-26] and inherent differences in I^2 comparing binary and continuous effect sizes.[28, 29] 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.[134] 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 a signal-to-noise metric, not a statistic about the magnitude of heterogeneity.

4.4. Exploring Heterogeneity

Meta-regression

Meta-regression is a common approach employed to examine the degree to which study-level factors explain statistical heterogeneity.[135] Random effects meta-regression, as compared with fixed effect meta-regression, allows for residual heterogeneity (i.e., between-study variance that is not explained by study-level factors) to incorporated into the model.[136]Because of this feature, among other benefits described below and in Chapter 3, random effects meta-regression is recommended over fixed effect meta-regression.[137] It is the default of several statistical packages to use a modified estimator of variance in random effects meta-regression that employs a *t* distribution in lieu of a standard normal distribution when calculating p-values and confidence intervals (i.e., the Knapp-Hartung modification).[138] This approach is recommended to help mitigate false-positive rates that are common in meta-regression.[137] 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.[136]Currently, the Cochrane manual recommends that there be at least 10 studies per characteristic modelled in meta-regression[10] over the enduring concern about inflated false-positive rates with too few studies.[137] 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, 10] 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. The focus of meta-regression should be on factors that are plausible. Second, meta-regression should be carried out under full consideration of ecological bias (i.e., the inherent problems associated with aggregating individual-level data).[139] As classic examples, the mean study age or the proportion of study participants who were female may result in different conclusions in meta-regression as opposed to how these modifying relationships functioned in each trial.[135]

Multiple 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 are fewer than 10 studies for each study-level factors modelled,[10, 136, 137] multiple meta-regression (that is, a meta-regression model with more than one study-level factor included) should only be considered when study-level characteristics are pre-specified, scientifically defensible, and based

on hypotheses, and when there are 10 or more studies for each study-level factor included in meta-regression.

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, 10] Like other forms of meta-regression, subgroup analyses have a high false-positive rate.[137] and may be misleading when few studies are included. There are two general approaches to handling subgroups in meta-analysis. First, a common use is to perform meta-analyses within subgroups without any statistical between-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 overall heterogeneity.[10] Second, it is recommended to incorporate the subgrouping factor into a meta-regression framework.[140] Doing so allows for quantification of both within and among subgroup heterogeneity as well as well as formal statistical testing that informs whether 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 to meta-regression using a continuous study-level factor.

Detecting Outlying Studies

Under consideration that removal of one or more studies from a meta-analysis may interject bias in the results,[10] identification of outlier studies may help build the evidence necessary to justify 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. One study removed procedures simply involve sequentially estimating the summary effect without each study to determine if single studies are having a large influence on model results. Using cumulative meta-analysis, [141] 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 graphical output (from Stata's *metacum* command or the R *metafor cumul()* function), one can examine large shifts in the summary effect that may serve as evidence for study removal. Another benefit of cumulative meta-analysis is detecting shifts in practice (e.g., guideline changes, new treatment approval or discontinuation) that would foster subgroup analysis.

Viechtbauer and Chung proposed other methods that should be considered to help identify outliers. One option is to examine extensions of linear regression residual diagnostics by using studentized deleted residuals.[142] Other options are to examine the difference between the predicted average effect with and without each study (indicating by how many standard deviations the average effect changes) or to examine what effect the deletion of each study has

on the fitted values of all studies simultaneously (in a metric similar to Cook's distance).[142] Particularly in combination, there methods serve as diagnostics that are more formal than visual inspection and both one study removed and cumulative meta-analysis procedures.

4.5. 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 can be 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 may differ across studies.[143, 144] Groups of patients with higher underlying risk for poor outcomes may experience different benefits and/or harms from treatment compared with groups of patients who have lower underlying risk.[145] Hence, the control-rate can be used to test for interactions between underlying population risk at baseline and treatment benefit.

To examine for an interaction between underlying population risk and treatment benefit, we recommend a simplified approach. First, generate a scatter plot of treatment effect against control rate to visually assess whether there may be a relation between the two. Since the RD tends to be highly correlated with the control rate,[144] we recommend using an RR or OR when examining a treatment effect against the control rate in all steps. The purpose of generating a scatterplot is simply to give preliminary insight into how differences in baseline risk (control rate) may influence the amount of observed variability in effect sizes across studies. Second, use hierarchical meta-regression[144] or Bayesian meta-regression[146] models to formally test the interaction between underlying population risk and treatment benefit. Although a weighted regression has been proposed as an intermediary step between developing a scatter plot and meta-regression, this approach identifies a significant relation between control rate and treatment effect twice as often compared with more suitable approaches (above),[144, 146] and a negative finding would likely need to be replicated using meta-regression. Hence, the simplified two-step approach may help streamline the process.

Multivariate Meta-analysis

There are both inherent benefits and disadvantages of using meta-analysis to examine multiple outcomes simultaneously (that is, "multivariate meta-analysis"), and much methodological work has been done in both frequentist and Bayesian frameworks in recent years.[147-156]. Some of these methods are readily available in statistical packages (for example, Stata *mvmeta*).

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 as being independent.[150] Another advantage of multivariate meta-analysis is being able to gain insight into relationships among study outcomes.[150, 157] An additional advantage of multivariate meta-analysis is that different clinical conclusions may be made;[150] it may be considered easier to present results from a single multivariate meta-analysis than from several univariate analyses that may make different assumptions. Further, multivariate methods may have the potential to reduce the impact of outcome reporting bias.[150, 158, 159]

Some of the major potential issues involved with the joint modeling of multiple outcomes in meta-analysis (reviewed by Jackson and colleagues)[150] include:

i. the disconnect between how outcomes are handled within each trial (typically in a univariate fashion) compared with a multivariate meta-analysis;
ii. estimation difficulties particularly around correlations between outcomes (seldom reported; see Bland[160] for additional commentary);
iii. overcoming assumptions of normally-distributed random effects with joint outcomes (difficult to justify with joint distributions);
iv. marginal model improvement in the multivariate vs. univariate case (often not sufficient trade off in effort); and
v. amplification of publication bias (e.g., secondary outcomes are not published as frequently).[150]

New methods not requiring within-study correlations are being developed to overcome the second limitation.[161, 162]

Another potential challenge is the appropriate quantification of heterogeneity in multivariate meta-analysis; but, there are newer alternatives that seem to make this less of a concern. These methods include but are 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)).[163] Finally, limitations to existing software for broad implementation and access to multivariate meta-analysis has been a long-standing barrier to this approach. With currently available add-on or base statistical packages, however, multivariate meta-analysis can be more readily performed,[150] and emerging approaches to multivariate meta-analyses are available to be integrated into standard statistical output.[153] However, the gain in precision of parameter estimates is often modest, and the conclusions from the multivariate meta-analysis are often the same as those from the univariate meta-analysis for individual outcomes,[164] which may not justify the increased complexity and difficulty.

With the exception of diagnostic testing meta-analysis (which provides a natural situation to meta-analyze sensitivity and specificity simultaneously, but which is out of scope for this report) and network meta-analysis (a special case of multivariate meta-analysis with unique challenges, see Chapter 5), multivariate meta-analysis has not been widely used in practice. However, we are likely to see multivariate meta-analysis approaches become more accessible to stakeholders involved with systematic reviews.[160] In the interim, however, we do not recommend this approach be used routinely.

Dose-Response Meta-analysis

Considering different exposure or treatment levels has been a longstanding consideration in meta-analyses involving binary outcomes.[165, 166] and new methods have been developed to extend this approach to differences in means.[167] 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-response using meta-regression is when there are several trials that compared the dose level versus control for each dosing level. 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.[10] 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. In many instances, exposure needs to be grouped for effective comparison (e.g., ever vs. never exposed), but doing so raises the issues of non-independence and covariance between estimates.[168] Hamling et al., developed a method of deriving relative effect and

precision estimates for such alternative comparisons in meta-analysis that are more reasonable compared with methods that ignore interdependence of estimates by level.[168] 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.[167] 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).[167] 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.[10]

Recommendations

- Statistical heterogeneity should be expected, visually inspected and quantified, and sufficiently addressed in all meta-analyses.
- Prediction intervals should be included in all forest plots.
- Investigators should be consider evaluating multiple metrics of heterogeneity, between-study variance, and inconsistency (i.e., Q, τ^2 and I^2 along with their respective confidence intervals when possible).
- A non-significant Q should not be interpreted as the absence of heterogeneity, and there are nuances to the interpretation of Q that carry over to the interpretation of τ^2 and I^2.
- Random effects is the preferred method for meta-regression that should be used under consideration of low power associated with limited studies (i.e., <10 studies per study-level factor) and the potential for ecological bias.
- We recommend a simplified two-step approach to control-rate meta-regression that involves scatter plotting and then hierarchical or Bayesian meta-regression.
- Routine use of multivariate meta-analysis is not recommended.

Chapter 5. Network Meta-Analysis (Mixed Treatment Comparisons/Indirect Comparisons)

M. Hassan Murad, M.D., M.P.H., Gerald Gartlehner, M.D., M.P.H., Rongwei Fu, Ph.D., Zhen Wang, Ph.D.

5.1. Rationale and Definition

Decision makers, whether patients, providers or policymakers generally want head-to-head estimates of the comparative effectiveness of the different interventions from which they have to choose. However, head-to-head trials are relatively uncommon. The majority of trials compare active agents with placebo, which has left patients and clinicians unable to compare across treatment options with sufficient certainty.

Therefore, an approach 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 with C; we can indirectly infer the magnitude of effect comparing A with C. Occasionally, a very limited number of head-to-head trials are available (i.e., there may be a small number of trials directly comparing A with 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 with C can be pooled with the direct comparisons, to produce what is commonly called a network meta-analysis estimate (NMA). 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) many interventions, creating a network of interventions with comparisons accounting for all the permutations of pairings within the network. The following guidance focuses on NMA of randomized controlled trials. NMA of nonrandomized studies is statistically possible; however, without randomization, NMA assumptions would likely not be satisfied and the results would not be reliable.

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.[169] This is true for any meta-analysis. In an 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, τ^2, 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.[170] Similarly, active and placebo controlled interventions across trials need to be sufficiently similar in order to attribute the observed change in effect size to the change in interventions.

Transitivity cannot be assessed quantitatively. However, it 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 (Between Direct and Indirect Evidence)

Comparing direct and indirect estimates in closed loops in a network demonstrates whether the network is consistent (previously 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 is similar to the pooled effect size from indirect comparison, we say the network is consistent; otherwise, the network is inconsistent or incoherent.[171, 172] Multiple causes have been proposed for inconsistency, such as differences in patients, treatments, settings, timing, and other factors.

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 always be evaluated. If there is substantial inconsistency between direct and indirect evidence, a network meta-analysis should not be performed. Fortunately, inconsistency can be evaluated statistically.

5.3. Statistical Approaches

Overview

The simplest indirect comparison approach 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. However, such qualitative comparisons have to be interpreted cautiously because the degree to which confidence intervals overlap is not a reliable substitute for formal hypothesis testing. Formal testing methods adjust the comparison of the interventions by the results of their direct comparison with a common control group and at least partially preserve the advantages of randomization of the component trials.[173]

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, and using both contrast level and arm level data.

Simple Indirect Comparisons

Simple indirect comparisons apply when there is no closed loop in the evidence network. A closed loop means that each comparison in a particular loop has both direct and indirect evidence. At least three statistical methods are available to conduct simple indirect comparisons: (1) the adjusted indirect comparison method proposed by Bucher et al,[174] (2) logistic regression, and (3) random effects meta-regression.

When there are only two sets of trials, say, A vs. B and C vs. B, Bucher's method is sufficient to provide the indirect estimate of A vs. C as: $\log(OR_{AC})=\log(OR_{AB})-\log(OR_{CB})$ and $Var(Log(OR_{AC})) = Var(Log(OR_{AB})) + Var(Log(OR_{CB}))$, where OR is the odds ratio. Bucher's method is valid only under a normality assumption on the log scale.

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. Under ideal circumstances (i.e., no differences in prognostic factors exist among included studies), all three methods result in unbiased estimates of direct effects.[175] Meta-regression (as implemented in Stata, *metareg*) and adjusted indirect comparisons are the most convenient approaches for comparing trials with two treatment arms. A simulation study supports the use of random effects for either of these approaches.[175]

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, and constructed a random-effects inconsistency model by incorporating sampling variability, heterogeneity, and inconsistency.[176] 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 more than two arms.

Further development of network meta-analysis models in the frequentist framework addressed how to handle multi-armed trials as well as new methods of assessing inconsistency.[171, 177-179] Salanti et al. provided a general network meta-analysis formulation 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.[177] In contrast, White et al. and Higgins et al. 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.[171, 178] These models can 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 proposed the first Bayesian network meta-analysis model for multi-arm studies that included both direct and indirect evidence.[180] The treatment effects are represented by basic parameters and functional parameters. Basic parameters are effect parameters that are directly compared to the baseline treatment, and functional parameters are represented as functions of basic parameters. Evidence inconsistency is defined as a function of 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,[181] to allow for repeated measurements of a continuous endpoint that varies over time,[87] or to appraise novelty effects.[182] A Bayesian multinomial network meta-analysis model was also developed for unordered (nominal) categorical outcomes allowing for partially observed data in which exact event counts may not be known for each category.[183] Additionally, Dias et al. 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.[86]

Commonly, a vague (flat) prior is chosen for the treatment effect and heterogeneity parameters in Bayesian network meta-analysis. A vague prior distribution for heterogeneity however may not be appropriate when the number of studies is small.[184] An informative prior for heterogeneity can be obtained from the empirically derived predictive distributions for the degree of heterogeneity as expected in various settings (depending on the outcomes assessed and comparisons made).[185] In the NMA framework, frequentist and Bayesian approaches often provide similar results; particularly because of the common practice to use non-informative priors in the Bayesian analysis.[186-188] 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 need a better understanding of the model specification and specialized programing skills.

Arm-Based Versus 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 represent raw data per study arm (e.g., the number of events from a trial per group); while contrast-level data show the difference in outcomes between arms in the form of absolute or relative effect size (e.g., mean difference or the odds ratio of events).

Contrast-based models resemble the traditional approaches used in meta-analysis of direct comparisons. Absolute or relative effect sizes and associated variances are first estimated (per study) and then pooled to produce an estimate of the treatment comparison. Contrast-based models preserve randomization and, largely, alleviate risk of observed and unobserved imbalance between arms within a study. They use effect sizes relative to the comparison group and reduce the variability of outcomes across studies. Contrast-based models are the dominant approach used in direct meta-analysis and network meta-analysis in current practice.

Arm-based models depend on directly combining the observed absolute effect size in individual arms across studies; thereby producing a pooled rate or mean of the outcome per arm. Estimates can be compared among arms to produce a comparative effect size. Arm-based models break randomization; therefore, the comparative estimate will likely be at an increased risk of bias. Following this approach, nonrandomized studies or even noncomparative studies can be included in the analysis. Multiple models have been proposed for the arm-based approach, especially in the Bayesian framework.[177, 189-192] However, the validity of arm-based methods is under debate. [178, 193, 194]

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.[173, 174, 176, 195-200]

These methods can generally be categorized into two types: (1) an overall consistency measure for the whole network; and (2) a loop-based approach in which direct and indirect estimates are compared. In the following section, we will focus on a few widely used methods in the literature.

i. Single Measure for Network Consistency: These approaches use a single measure that represents consistency 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.[176] A ω value above 0.25 suggests substantial inconsistency and in this case, network meta-analysis may be considered inappropriate.[201]

ii. Global Wald Test: Another approach is to use global Wald test, which tests an inconsistency factor that follows a X^2 distribution under the null consistency assumption.[178] A p-value less than 0.10 can be used to determine statistical significance. Rejection of the null is evidence that the model is not consistent.

iii. Loop-based approach: This approach involves comparing direct and indirect estimates for each comparison. This approach is preferred over a global test (i.e., a single measure for the whole network). 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 the pooled effect sizes between direct and indirect comparisons.[174] Benefits of this approach include simplicity, ease of application, and the ability to identify specific loops with large inconsistency. Limitations include the need for multiple correlated tests.

   b. *Side-splitting:* A "node" is a treatment and a "side" (or edge) is a comparison. Dias et al. suggests that each comparison can be assessed by comparing the difference of the pooled estimate from direct evidence to the pooled estimate without direct evidence.[196] Side-splitting (sometimes referred to as node-splitting) can be implemented using the Stata *network sidesplit* command or R *gemtc* package.

Several graphical tools have been developed to describe inconsistency. One is the inconsistency plot developed by Chaimani et al.[197] Similar to a forest plot, the inconsistency plot graphically presents an inconsistency factor (the absolute difference between the direct and indirect estimates) and related confidence interval for each of the triangular and quadratic loops in the network. The Stata *ifplot* command can be used for this purpose.

It is important to understand the limitations of these methods. Lack of statistical significance of an inconsistency test does not prove consistency in the network. Similar to Cochran's Q test of heterogeneity testing in traditional meta-analysis (which is often underpowered), statistical tests for inconsistency in NMA are also commonly underpowered due to the limited number of studies in direct comparisons.

When a network exhibits important inconsistency, the options are:

- Abandon NMA and only perform traditional meta-analysis;

- Present the results from inconsistency models (that incorporate inconsistency) and acknowledge the limited trustworthiness of the NMA estimates;
- Split the network to eliminate the inconsistent nodes;
- Attempt to explain the causes of inconsistency by conducting network meta-regression to test for possible covariates causing the inconsistency: and
- Use only direct estimates for the pairwise NMA comparisons that show inconsistency (i.e., use direct estimates for inconsistent comparisons and use NMA estimates for consistent comparisons).

There is no preferred strategy, and investigators need to choose the approach that fits the situation best.

5.4. Considerations of Model Choice and Software

Consideration of Indirect Evidence

Empirical explorations suggest that direct and indirect comparisons often agree,[174-176, 202-204] but with notable exceptions.[205] In principle, the validity of combining direct and indirect evidence relies on the transitivity assumption. However, in practice, trials can vary in numerous ways including population characteristics, interventions, and cointerventions, length of follow-up, loss to follow-up, study quality, etc. Given the limited information in many publications and the inclusion of multiple treatments, the validity of combining direct and indirect evidence is often unverifiable. The statistical methods to evaluate inconsistency generally have low power, and are confounded by the presence of statistical heterogeneity. They often fail to detect inconsistency in the evidence network.

Moreover, network meta-analysis, like all other meta-analytic approaches, 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 the 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 may consider incorporating 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 Method

Although the development of network meta-analysis models has exploded in the last 10 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 mode of inference based on the research team expertise, the complexity of the evidence network, and/or the research question. If investigators believe that the use of prior information is needed and that the data are insufficient to capture all the information available, then they should use a Bayesian model. On the other hand, a frequentist model is appropriate if one wants inferences to be based only on the data that can be incorporated into a likelihood.

Whichever method the investigators choose, they should assess the consistency of the direct and indirect evidence, and the invariance of treatment effects across studies and the

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.[174, 175]

When the evidence of a network of interventions is consistent, investigators can combine direct and indirect evidence using network meta-analysis models. Conversely, they should refrain from combining multiple sources of evidence from an inconsistent (i.e., incoherent) network where there are substantial differences between direct and indirect evidence that cannot be resolved by conditioning on the known covariates. 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 (**Figure 5.1)** can also affect the choice of analysis method as demonstrated in Table 5.1.

Figure 5.1. Common network geometry (simple indirect comparison, star, network with at least one closed loop)

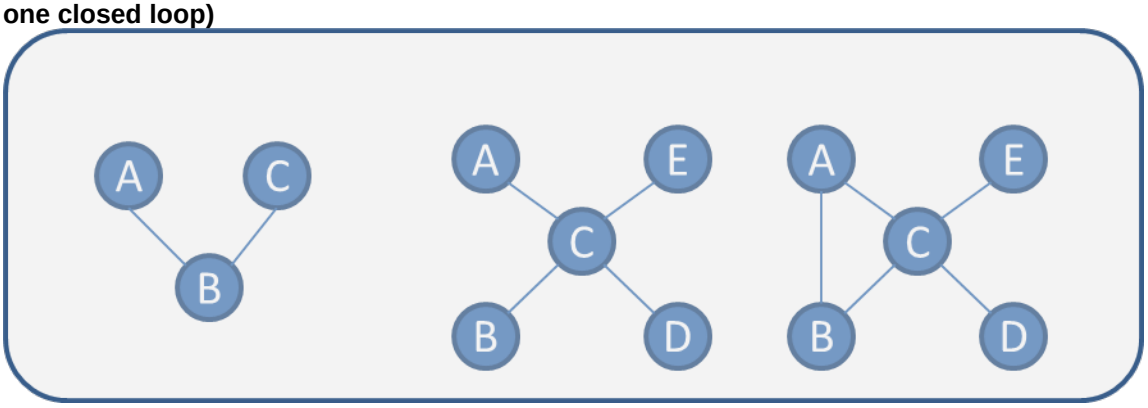


Table 5.1. Impact of network geometry on choice of analysis method

Methods	Simple indirect comparison	Star network*	Network with at least one closed loop†
Qualitative assessment	*X*		
Adjusted Indirect comparison, random-effects meta regression, logistic regression	*X*	*X*	
Lumley's mixed-effects linear regression[176]			*X*
Mixed effects and hierarchical models [171, 178]		*X*	*X*

X = Appropriate method for the network geometry

*In a star network, all interventions (A, B, D, E) connect to a common comparator (C). There is no other direct comparison.

†Network with at least one closed loop is an extension of a star network. Besides a common comparator (C) in the network, there is at least one direct comparison between other interventions (closed loop). In this example, A, B, and C form a closed loop.

Commonly Used Software

Many statistical packages are available to implement NMA. BUGS software (Bayesian inference Using Gibbs Sampling, WINBUGS, OPENBUGS) is a popular choice for conducting Bayesian NMA[206] that offers flexible model specification including NMA meta-regression. JAGS and STAN are alternative choices for Bayesian NMA. Stata provides user-written routines (http://www.mtm.uoi.gr/index.php/stata-routines-for-network-meta-analysis) that can be used to conduct frequentist NMA. In particular, the Stata command *network* is a suite of programs for importing data for network meta-analysis, running a contrast-based network meta-analysis, assessing inconsistency, and graphing the data and results. Further, in the R environment, three packages, *gemtc* (http://cran.r-project.org/web/packages/gemtc/index.html), *pcnetmeta* (http://cran.r-project.org/web/packages/pcnetmeta/index.html), and netmeta (http://cran.r-project.org/web/packages/netmeta/index.html), have been developed for Bayesian (*gemtc*, *pcnetmeta*) or frequestist (*netmeta*) NMA. The packages also include methods to assess heterogeneity and inconsistency, and data visualizations, and allow users to perform NMA with minimal programming.[207]

5.5. Inference From Network Meta-analysis

Stakeholders (users of evidence) require a rating of the strength of a body of evidence. 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 NMA, guidance is evolving and may require some additional computations; therefore, we briefly discuss the possible approaches to rating the strength of evidence. We also discuss inference from rankings and probabilities commonly presented with a network meta-analysis.

Approaches for Rating the Strength of Evidence

The original EPC and 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.[208] Subsequently, 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.[209] 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.[210] 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 of treatments. Results are also presented as the surface under the cumulative ranking curve (SUCRA). SUCRA is a simple transformation of the mean rank that is used to provide a hierarchy of the treatments accounting both for the location and the variance of all relative treatment effects. SUCRA would be 1 when a treatment is certain to be the best and 0 when a treatment is certain to be the worst.[211] Such presentations should be interpreted with caution since they can be quite misleading.

Whether results were presented as probabilities, rankings or SUCRA, three concerns 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 SUCRA estimates was 65% (the first quartile was 38%; the third quartile was 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 or 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.[212] 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.[213]
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 demonstrate great heterogeneity in reporting and numerous deficiencies. Commonly, network meta-analyses demonstrate an 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.[214-216] 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.[217] We advise the following information be presented so that the adequacy of the NMA can be assessed:

- Rationale for conducting an NMA, the mode of inference (e.g., Bayesian, Frequentist), and the model choice (random effects vs. fixed effects; consistency vs inconsistency model, common heterogeneity assumption, etc.);
- Software and syntax/commands used;
- Choice of priors for any Bayesian analyses;

- Graphical presentation of the network structure and geometry;
- Pairwise effect sizes to allow comparative effectiveness inference; and
- Assessment of the extent of consistency between the direct and indirect estimates.

Recommendations

- A network meta-analysis should always be based on a rigorous a rigorous systematic review.
- Conducting network meta-analysis requires that three assumptions be met:
  - Homogeneity of direct evidence
  - Transitivity, similarity, or exchangeability
  - Consistency (between direct and indirect evidence)
- Investigators may choose a frequentist or Bayesian mode of inference based on the research team's expertise, the complexity of the evidence network, and the research question.
- Evaluating inconsistency is a major and mandatory component of network meta-analysis.
- Evaluating inconsistency should not be only based on a conducting a global test. A loop-based approach can identify the comparisons that cause inconsistency.
- Inference based on the rankings and probabilities of treatments being most effective should be used cautiously. Rankings and probabilities can be misleading and should be interpreted based on the magnitude of pairwise effect sizes. Differences across interventions may not be clinically important despite such rankings.

Future Research Suggestions

The following are suggestions for directions in future research for each of the topics by chapter.

Chapter 1. Decision To Combine Trials

- Guidance regarding the minimum number of trials one can validly pool at given levels of statistical heterogeneity

Chapter 2. Optimizing Use of Effect Size Data

- Research on ratio of means—both clinical interpretability and mathematical consistency across studies compared with standardized mean difference
- Research on use of ANCOVA models for adjusting baseline imbalance
- Software packages that more easily enable use of different information
- Methods to handle zeros in the computation of binary outcomes
- Evidence on which metrics, and language used to describe these metrics, are most helpful in conveying meta-analysis results to multiple stakeholders

Chapter 3. Choice of Statistical Model for Combining Studies

- Evaluate newly developed statistical models for combining typical effect measures (e.g., mean difference, OR, RR, and/or RD) and compare with current methods

Chapter 4. Quantifying, Testing, and Exploring Statistical Heterogeneity

- Heterogeneity statistics for meta-analyses involving a small number of studies
- Guidance on specification of hypotheses in meta-regression
- Guidance on reporting of relationships among study outcomes to facilitate multivariate meta-analysis

Chapter 5. Network Meta-analysis (Mixed Treatment Comparisons/Indirect Comparisons)

- Methods for combining individual patient data with aggregated data
- Methods for integrating evidence from RCTs and observational studies
- Models for time-to-event data
- User friendly software similar to that available for traditional meta-analysis
- Evidence to support model choice

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, Terrin N, Fu R. Expanded Guidance on Selected Quantitative Synthesis Topics Agency for Healthcare Research and Quality. Rockville, MD: 2013.
3. 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.
4. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol. 2010;63(5):502-12. PMID: 18823754 http://dx.doi.org/10.1016/j.jclinepi.2008.06.007
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. Scand J Work Environ Health. 2012;38(3):282-90. http://dx.doi.org/10.5271/sjweh.3201
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. Clin Trials. 2013;10(1):20-31. http://dx.doi.org/10.1177/1740774512465495
8. Gagnier JJ, Morgenstern H, Altman DG, et al. Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. BMC Med Res Methodol. 2013;13(1):106. http://dx.doi.org/10.1186/1471-2288-13-106
9. Sun X, Guyatt G. Meta-analysis of randomized trials for health care interventions: one for all? J Evid Based Med. 2009;2(1):53-6. http://dx.doi.org/10.1111/j.1756-5391.2009.01006.x
10. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions: Wiley Online Library; 2008.
11. 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. http://dx.doi.org/10.1371/journal.pone.0059202
12. 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 Serv Res. 2009;19(9):34. http://dx.doi.org/10.1186/1472-6963-9-34
13. Schmid CH. Outcome reporting bias: a pervasive problem in published meta-analyses. American Journal of Kidney Diseases. 2016 2016;69(2):172-4.
14. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000;53(11):1119-29.
15. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34. PMID: 9310563.
16. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol. 2011;64(12):1283-93. PMID: 21839614 http://dx.doi.org/10.1016/j.jclinepi.2011.01.012
17. Bowater RJ, Escarela G. Heterogeneity and study size in random-effects meta-analysis. J Appl Stat 2013;40(1):2-16. https://doi.org/10.1080/02664763.2012.700448

18. Wetterslev J, Thorlund K, Brok J, et al. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol. 2009;9(1):1.

19. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001;20(6):825-40. PMID: 11252006. http://dx.doi.org/10.1002/sim.650

20. 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? J Stat Plan Inference. 2010;140(4):961-70. http://dx.doi.org/10.1016/j.jspi.2009.09.017

21. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. Stat Methods Med Res. 2012;21(4):409-26. http://dx.doi.org/10.1177/0962280210392008

22. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60. PMID: 12958120. http://dx.doi.org/10.1136/bmj.327.7414.557

23. Borenstein M, Hedges LV, Higgins JPT, et al. Introduction to meta-analysis. John Wiley & Sons, Ltd, Chichester, UK; 2009.

24. Melsen WG, Bootsma MC, Rovers MM, et al. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clin Microbiol Infect. 2014;20(2):123-9. http://dx.doi.org/10.1111/1469-0691.12494

25. Higgins J, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539-58. PMID: 12111919. http://dx.doi.org/10.1002/sim.1186

26. von Hippel PT. The heterogeneity statistic I2 can be biased in small meta-analyses. BMC Med Res Methodol. 2015;15(1):35. http://dx.doi.org/10.1186/s12874-015-0024-z

27. Rücker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I 2 in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8(1):79. PMID: 19036172 http://dx.doi.org/10.1186/1471-2288-8-79

28. Alba AC, Alexander PE, Chang J, et al. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. J Clin Epidemiol. 2016;70:129-35. http://dx.doi.org/10.1016/j.jclinepi.2015.09.005

29. Rhodes KM, Turner RM, Higgins JP. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. Res Synth Methods. 2015. http://dx.doi.org/10.1002/jrsm.1193 [doi]

30. 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. PMID: 11323066

31. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. PMID: 19621072 http://dx.doi.org/10.1371/journal.pmed.1000097

32. Dahabreh IJ, Trikalinos TA, Lau J, et al. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. Methods Research Report. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I.) AHRQ Publication No 12(13)-EHC136-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2012. www.effectivehealthcare.ahrq.gov/reports/final/cfm.

33. Schulzer M, Mancini GJ. 'Unqualified Success' and 'Unmitigated Failure'Number-Needed-to-Treat-Related Concepts for Assessing Treatment Efficacy in the Presence of Treatment-Induced Adverse Events. Int J Epidemiol. 1996;25(4):704-12.

34. Altman DG. Confidence intervals for the number needed to treat. BMJ. 1998;317(7168):1309. PMID: 9804726

35. Bender R. Calculating confidence intervals for the number needed to treat. Control Clin Trials 2001;22(2):102-10. PMID: 11306148

36. Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am Stat 2000;54(4):280-8.

37. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 1998;17(8):873-90. PMID: 9595617.

38. Lesaffre E, Pledger G. A note on the number needed to treat. Control Clin Trials. 1999;20(5):439-47. https://doi.org/10.1016/S0197-2456(99)00018-5

39. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. 2012;41(2):514-20. PMID: 22253321. http://dx.doi.org/10.1093/ije/dyr218

40. Brockhaus AC, Bender R, Skipka G. The Peto odds ratio viewed as a new effect measure. Stat Med. 2014;33(28):4861-74. http://dx.doi.org/10.1002/sim.6301

41. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002;21(11):1575-600. PMID: 12111921. http://dx.doi.org/10.1002/sim.1188

42. Huang HY, Andrews E, Jones J, et al. Pitfalls in meta-analyses on adverse events reported from clinical trials. Pharmacoepidemiol Drug Saf. 2011;20(10):1014-20. http://dx.doi.org/10.1002/pds.2208

43. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007;26(1):53-77. PMID: 16596572. http://dx.doi.org/10.1002/sim.2528

44. Liu Z, Rich B, Hanley JA. Recovering the raw data behind a non-parametric survival curve. Syst Rev. 2014;3(1):1. PMID: 25551437. http://dx.doi.org/10.1186/2046-4053-3-151

45. Spittal MJ, Pirkis J, Gurrin LC. Meta-analysis of incidence rate data in the presence of zero events. BMC Med Res Methodol. 2015;15(1):42.

46. Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol. 2009; PMID: 19223279 http://dx.doi.org/10.1093/jpepsy/jsp004

47. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edn. Hillsdale, New Jersey: L. Erlbaum; 1988

48. Follmann D, Elliott P, Suh I, et al. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol. 1992;45(7):769-73. PMID: 1619456.

49. Senn S. Covariate imbalance and random allocation in clinical trials. Stat Med. 1989;8(4):467-75.

50. Senn S. Change from baseline and analysis of covariance revisited. Stat Med. 2006;25(24):4334-44.

51. 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. Res Synth Methods. 2015. http://dx.doi.org/10.1002/jrsm.1196

52. 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. PMID: 23326900 Rockville, MD: 2012. http://www.effectivehealthcare.ahrq.gov/ehc/products/344/1322/CorrelationImputation_FinalReport_20121119.pdf

53. Camilli G, de la Torre J, Chiu C. A Noncentral t Regression Model for Meta-Analysis J Educ Behav Stat. 2010;35(2):125-53. http://dx.doi.org/10.3102/1076998609346966

54. Claggett B, Xie M, Tian L. Meta-Analysis With Fixed, Unknown, Study-Specific Parameters. J Am Stat Assoc. 2014;109(508):1660-71. http://dx.doi.org/10.1080/01621459.2014.957288

55. 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. http://dx.doi.org/10.1093/biostatistics/kxt050

56. Doi SA, Barendregt JJ, Khan S, et al. Advances in the Meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. Control Clin Trials. 2015;45(Pt. A):130-8. http://dx.doi.org/10.1016/j.cct.2015.05.009

57. Stanley TD, Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. Stat Med. 2015. http://dx.doi.org/10.1002/sim.6481

58. Mawdsley D, Higgins J, Sutton AJ, et al. Accounting for heterogeneity in meta-analysis using a multiplicative model—an empirical study. Res Synth Methods. 2017;8(1):43-52. PMID: 27259973. http://dx.doi.org/10.1002/jrsm.1216

59. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88. https://doi.org/10.1016/0197-2456(86)90046-2

60. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials 2007;28(2):105-14. PMID: 16807131. http://dx.doi.org/10.1016/j.cct.2006.04.004

61. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Control Clin Trials. 2015;45(Pt. A):139-45. http://dx.doi.org/10.1016/j.cct.2015.09.002

62. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Stat Med. 2002;21(21):3153-9.

63. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Stat Med 2001;20(12):1771-82. PMID: 11406840. http://dx.doi.org/10.1002/sim.791

64. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med 2001;20(24):3875-89. PMID: 11782040

65. Biggerstaff B, Tweedie R. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat Med 1997;16(7):753-68. PMID: 9131763

66. Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. Stat Methods Med Res. 2015. 0962280215583568 [pii]

67. IntHout 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 Med Res Methodol. 2014;14(1):25. http://dx.doi.org/10.1186/1471-2288-14-25

68. Cornell JE, Mulrow CD, Localio R, et al. Random-Effects Meta-analysis of Inconsistent Effects: A Time for Change. Ann of Intern Med. 2014;160(4):267-70. http://dx.doi.org/10.7326/M13-2886

69. Shuster JJ, Walker MA. Low-event-rate meta-analyses of clinical trials: implementing good practices. Stat Med. 2016. http://dx.doi.org/10.1002/sim.6844

70. 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. Stat Med 2004;23(9):1351-75. PMID: 15116347 http://dx.doi.org/10.1002/sim.1761

71. Fleiss J. The statistical basis of meta-analysis. Stat Methods Med Res. 1993;2(2):121-45. PMID: 8261254. http://dx.doi.org/10.1177/096228029300200202

72. Vandermeer B, Bialy L, Hooton N, et al. Meta-analyses of safety data: a comparison of exact versus asymptotic methods. Stat Methods Med Res. 2009;18(4):421-32. http://dx.doi.org/10.1177/0962280208092559

73. Bhaumik DK, Amatya A, Normand SL, et al. Meta-Analysis of Rare Binary Adverse Event Data. J Am Stat Assoc. 2012;107(498):555-67. http://dx.doi.org/10.1080/01621459.2012.664484

74. 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.

75. Mehta CR. The exact analysis of contingency tables in medical research. Recent Advances in Clinical Trial Design and Analysis. Springer; 1995:177-202.

76. Vazquez FJ, Moreno E, Negrin MA, et al. Bayesian robustness in meta-analysis for studies with zero responses. Pharm Stat. 2016. http://dx.doi.org/10.1002/pst.1741

77. Bai O, Chen M, Wang X. Bayesian Estimation and Testing in Random Effects Meta-Analysis of Rare Binary Adverse Events. Stat Biopharm Res. 2016;8(1):49-59. PMID: 27127551. http://dx.doi.org/10.1080/19466315.2015.1096823

78. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. Stat Med. 2010;29(12):1259-65. http://dx.doi.org/10.1002/sim.3607

79. Spittal MJ, Pirkis J, Gurrin LC. Meta-analysis of incidence rate data in the presence of zero events. BMC Med Res Methodol. 2015;15(1):1.

80. Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. Stat Med. 2015;34(7):1097-116. http://dx.doi.org/10.1002/sim.6383

81. Deeks JJ, Higgins JPT, Altman DG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). [s.l.]: Cochrane Collaboration; 2011.

82. Mehta CR, Patel NR. Exact logistic regression: theory and examples. Stat Med 1995;14(19):2143-60. PMID: 8552893

83. 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. http://dx.doi.org/10.1093/biostatistics/kxn034

84. Liu D, Liu RY, Xie M. Exact Meta-Analysis Approach for Discrete Data and its Application to 2 × 2 Tables With Rare Events. J Am Stat Assoc. 2014;109(508):1450-65. http://dx.doi.org/10.1080/01621459.2014.946318

85. Yang G, Liu D, Wang J, et al. Meta-analysis framework for exact inferences with application to the analysis of rare events. Biometrics. 2016.

86. 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. Med Decis Making. 2013;33(5):607-17. http://dx.doi.org/10.1177/0272989X12458724

87. 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. Stat Med. 2011;30(20):2511-35. http://dx.doi.org/10.1002/sim.4284

88. Schmid CH. Using Bayesian inference to perform meta-analysis. Eval Health Prof. 2001;24(2):165-89.

89. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation: John Wiley & Sons; 2004.

90. 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. Stat Med. 2005;24(15):2401-28. PMID: 16015676. http://dx.doi.org/10.1002/sim.2112

91. 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. Med Decis Making. 2013;33(3):356-68. http://dx.doi.org/10.1177/0272989X12453504

92. J H, Anne W. Borrowing strength from external trials in a meta-analysis. Stat Med. 1996;15(24):2733-49. PMID: 8981683. http://dx.doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0

93. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172(1):137-59. http://dx.doi.org/10.1111/j.1467-985X.2008.00552.x

94. 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. Int J Epidemiol. 2012;41(3):818-27. http://dx.doi.org/10.1093/ije/dys041

95. Ades A, Lu G, Higgins J. The interpretation of random-effects meta-analysis in decision models. Med Decis Making. 2005;25(6):646-54. PMID: 16282215. http://dx.doi.org/10.1177/0272989X05282643

96. Unit WMB. WinBUGS | MRC Biostatistics Unit. 2016. http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/2016.

97. OpenBUGS. OpenBUGS. 2016. http://www.openbugs.net/w/FrontPage2016.

98. . 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.

99. JAGS - Just Another Gibbs Sampler. JAGS - Just Another Gibbs Sampler.; 2016. http://mcmc-jags.sourceforge.net/2016.

100. Carpenter B, Gelman A, Hoffman M, et al. Stan: A probabilistic programming language. J Stat Softw. 2016.

101. de Valpine P, Turek D, Paciorek CJ, et al. Programming with models: writing statistical algorithms for general model structures with NIMBLE. J Comput Graph Stat. 2016(just-accepted):1-28. https://doi.org/10.1080/10618600.2016.1172487

102. Nimble-admin. NIMBLE | An R package for programming with BUGS models and compiling parts of R. 2016. http://r-nimble.org/2016.

103. Thompson J, Palmer T, Moreno S. Bayesian analysis in Stata using WinBUGS. Stata J. 2006;6(4):530-49.

104. bmeta. bmeta - Bayesian meta-analysis & meta-regression in R - Gianluca Baio. . 2016. https://sites.google.com/a/statistica.it/gianluca/bmeta2016.

105. van Valkenhoef G, Lu G, de Brock B, et al. Automating network meta-analysis. Res Synth Methods. 2012;3(4):285-99. http://dx.doi.org/10.1002/jrsm.1054

106. van Valkenhoef G, Dias S, Ades AE, et al. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Res Synth Methods. 2016;7(1):80-93. http://dx.doi.org/10.1002/jrsm.1167

107. 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.

108. Bayesian “random-effects” models. Bayesian “random-effects” models | Stata News. . 2016. http://www.stata.com/stata-news/news30-2/bayesian-random-effects/2016.

109. 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. Stat Med. 2010;29(29):3046-67. http://dx.doi.org/10.1002/sim.4040

110. Senn S. Trying to be precise about vagueness. Stat Med. 2007;26(7):1417.

111. Moreno E, Vázquez-Polo FJ, Negrin MA. Objective Bayesian meta-analysis for sparse discrete data. Stat Med. 2014;33(21):3676-92. http://dx.doi.org/10.1002/sim.6163

112. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol. 2008;37(5):1158-60. https://doi.org/10.1093/ije/dyn204

113. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002. http://dx.doi.org/10.1136/bmj.d4002

114. Langan D, Higgins JP, Simmonds M. An empirical comparison of heterogeneity variance estimators in 12,894 meta-analyses. Res Synth Methods. 2015;6(2):195-205. http://dx.doi.org/10.1002/jrsm.1140

115. Anzures-Cabrera J, Higgins JPT. Graphical displays for meta-analysis: An overview with suggestions for practice. Res Synth Methods. 2010;1(1):66-80. http://dx.doi.org/10.1002/jrsm.6

116. IntHout J, Ioannidis JP, Rovers MM, et al. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247. PMID: 27406637. http://dx.doi.org/10.1136/bmjopen-2015-010247

117. Guddat C, Grouven U, Bender R, et al. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev. 2012;1(1):34. http://dx.doi.org/10.1186/2046-4053-1-34

118. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol. 2005;58(9):894-901. https://doi.org/10.1016/j.jclinepi.2005.01.006

119. Tang J-L, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. J Clin Epidemiol. 2000;53(5):477-84. https://doi.org/10.1016/S0895-4356(99)00204-8

120. Lau J, Ioannidis JP, Terrin N, et al. Evidence based medicine: The case of the misleading funnel plot. BMJ. 2006;333(7568):597. http://dx.doi.org/10.1136/bmj.333.7568.597

121. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods. 2016;7(1):55-79. http://dx.doi.org/10.1002/jrsm.1164

122. Hoaglin DC. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. Stat Med. 2016;35(4):485-95. http://dx.doi.org/10.1002/sim.6632

123. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, et al. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? Psychol Methods. 2006;11(2):193. PMID: 16784338. http://dx.doi.org/10.1037/1082-989X.11.2.193

124. Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. Stat Med. 2006;25(24):4321-33. PMID: 16991104. http://dx.doi.org/10.1002/sim.2692

125. Kulinskaya E, Dollinger MB, Bjorkestol K. Testing for homogeneity in meta-analysis I. The one parameter case: Standardized mean difference. Biometrics. 2011;67(1):203-12. http://dx.doi.org/10.1111/j.1541-0420.2010.01442.x

126. Kulinskaya E, Dollinger MB. An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. BMC Med Res Methodol. 2015;15(1):49. http://dx.doi.org/10.1186/s12874-015-0034-x

127. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17(8):841-56. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8<841::AID-SIM781>3.0.CO;2-D

128. Baujat B, Mahé 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;21(18):2641-52. http://dx.doi.org/10.1002/sim.1221

129. IntHout J, Ioannidis JP, Borm GF, et al. Small studies are more heterogeneous than large ones: a meta-meta-analysis. J Clin Epidemiol. 2015;68(8):860-9. http://dx.doi.org/10.1016/j.jclinepi.2015.03.017

130. Paule RC, Mandel J. Consensus values and weighting factors. J Res Natl Bur Stand. 1982;87(5):377-85.

131. Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. . Res Synth Methods. 2017;8:181-98.

132. Raudenbush SW. Analyzing effect sizes: Random-effects models. In: Cooper H, Hedges LV, Valentine JC, eds. The Handbook of Research Synthesis and Meta-analysis. Vol. 2nd. New York, NY: Russell Sage Foundation; 2009:295-315.

133. 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. http://dx.doi.org/10.1371/journal.pone.0039471

134. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(7626):914. PMID: 17974687. http://dx.doi.org/10.1136/bmj.39343.408449.80

135. Thompson SG, Higgins J. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21(11):1559-73. http://dx.doi.org/10.1002/sim.1187

136. Berkey CS, Hoaglin DC, Mosteller F, et al. A random-effects regression model for meta-analysis. Stat Med 1995;14(4):395-411. PMID: 7746979

137. Higgins J, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004;23(11):1663-82. http://dx.doi.org/10.1002/sim.1752

138. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med 2003;22(17):2693-710. http://dx.doi.org/10.1002/sim.1482

139. 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;21(3):371-87. http://dx.doi.org/10.1002/sim.1023

140. Borenstein M, Higgins JPT. Meta-Analysis and Subgroups. Prev Sci. 2013;14(2):134-43. http://dx.doi.org/10.1007/s11121-013-0377-7

141. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med. 1992;327(4):248-54. http://dx.doi.org/10.1056/NEJM199207233270406

142. Viechtbauer W, Cheung MW. Outlier and influence diagnostics for meta-analysis. Res Synth Methods. 2010;1(2):112-25. http://dx.doi.org/10.1002/jrsm.11

143. M.W. M. The population risk as an explanatory variable in research synthesis of clinical trials. Stat Med. 1996;15(16):1713-28. http://dx.doi.org/10.1002/(SICI)1097-0258(19960830)15:16<1713::AID-SIM331>3.0.CO;2-D

144. 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.

145. Glasziou PILM. An evidence based approach to individualising treatment. BMJ. 1995;311:1356. http://dx.doi.org/10.1136/bmj.311.7016.1356

146. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med. 1997;16(23):2741-58.

147. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002;21(4):589-624. http://dx.doi.org/10.1002/sim.1040

148. Riley RD, Abrams K, Lambert P, et al. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med. 2007;26(1):78-97. http://dx.doi.org/10.1002/sim.2524

149. Nam IS, Mengersen K, Garthwaite P. Multivariate meta-analysis. Stat Med 2003;22(14):2309-33. PMID: 12854095. http://dx.doi.org/10.1002/sim.1410

150. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. Stat Med. 2011;30(20):2481-98. http://dx.doi.org/10.1002/sim.4172

151. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. Stat Med. 2010;29(12):1282-97. http://dx.doi.org/10.1002/sim.3602

152. 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 J. Biometrische Zeitschrift. 2013;55(2):231-45. http://dx.doi.org/10.1002/bimj.201200152

153. Jackson D, Rollins K, Coughlin P. A multivariate model for the meta-analysis of study level survival data at multiple times. Res Synth Methods. 2014;5(3):264-72. http://dx.doi.org/10.1002/jrsm.1112

154. Chen H, Manning AK, Dupuis J. A method of moments estimator for random effect multivariate meta-analysis. Biometrics. 2012;68(4):1278-84. http://dx.doi.org/10.1111/j.1541-0420.2012.01761.x

155. Van den Noortgate W, López-López JA, Marín-Martínez F, et al. Meta-analysis of multiple outcomes: a multilevel approach. Behav Res Methods. 2015;47(4):1274-94. http://dx.doi.org/10.3758/s13428-014-0527-2

156. Hurtado Rua SM, Mazumdar M, Strawderman RL. The choice of prior distribution for a covariance matrix in multivariate meta-analysis: a simulation study. Stat Med. 2015;34(30):4083-104. http://dx.doi.org/10.1002/sim.6631

157. Hedges LV. Effect sizes in three-level cluster-randomized experiments. J Educ Behav Stat. 2011;36(3):346-80. http://dx.doi.org/10.3102/1076998610376617

158. Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. Stat Med. 2012;31(20):2179-95. http://dx.doi.org/10.1002/sim.5356

159. Frosi G, Riley RD, Williamson PR, et al. Multivariate meta-analysis helps examine the impact of outcome reporting bias in Cochrane rheumatoid arthritis reviews. J Clin Epidemiol. 2014;68(5):542-50. http://dx.doi.org/10.1016/j.jclinepi.2014.11.017

160. Bland JM. Comments on 'Multivariate meta-analysis: Potential and promise'by Jackson et al., Statistics in Medicine. Stat Med. 2011;30(20):2502-3. http://dx.doi.org/10.1002/sim.4223

161. Chen Y, Hong C, Riley RD. An alternative pseudolikelihood method for multivariate random-effects meta-analysis. Stat Med. 2015;34(3):361-80. http://dx.doi.org/10.1002/sim.6350

162. Chen Y, Cai Y, Hong C, et al. Inference for correlated effect sizes using multiple univariate meta-analyses. Stat Med. 2016;35(9):1405-22. http://dx.doi.org/10.1002/sim.6789

163. Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. Stat Med. 2012;31(29):3805-20. http://dx.doi.org/10.1002/sim.5453

164. Trikalinos TA, Hoaglin DC, Schmid CH. An empirical comparison of univariate and multivariate meta-analyses for categorical outcomes. Stat Med. 2014;33(9):1441-59.

165. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol. 1992;135(11):1301-9. https://doi.org/10.1093/oxfordjournals.aje.a116237

166. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. Epidemiology. 1993;4(3):218-28.

167. Crippa A, Orsini N. Dose-response meta-analysis of differences in means. BMC Med Res Methodol. 2016;16(1):91. https://doi.org/10.1186/s12874-016-0189-0

168. Hamling J, Lee P, Weitkunat R, et al. Facilitating meta-analyses by deriving relative effect and precision estimates for alternative comparisons from a set of estimates presented by exposure level or disease category. Stat Med. 2008;27(7):954-70. http://dx.doi.org/10.1002/sim.3013

169. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol. 2011;64(12):1294-302. PMID: 21803546. http://dx.doi.org/10.1016/j.jclinepi.2011.03.017

170. Mills EJ, Ioannidis JP, Thorlund K, et al. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA 2012;308(12):1246-53. PMID: 23011714. http://dx.doi.org/10.1001/2012.jama.11228

171. Higgins JPT, Jackson D, Barrett JK, et al. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Methods. 2012;3(2):98-110. http://dx.doi.org/10.1002/jrsm.1044

172. 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. Res Synth Methods. 2012;3(2):80-97. http://dx.doi.org/10.1002/jrsm.1037

173. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol. 2008;61(5):455-63. PMID: 18394538. http://dx.doi.org/10.1016/j.jclinepi.2007.06.006

174. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997;50(6):683-91. PMID: 9250266

175. Glenny A, Altman D, Song F, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005; PMID: 16014203

176. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002;21(16):2313-24. PMID: 12210616. http://dx.doi.org/10.1002/sim.1201

177. Salanti G, Higgins JP, Ades A, et al. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008;17(3):279-301. PMID: 17925316. http://dx.doi.org/10.1177/0962280207080643

178. White IR, Barrett JK, Jackson D, et al. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods. 2012;3(2):111-25. http://dx.doi.org/10.1002/jrsm.1045

179. Greco T, Edefonti V, Biondi-Zoccai G, et al. A multilevel approach to network meta-analysis within a frequentist framework. Control Clin Trials. 2015;42:51-9. http://dx.doi.or/10.1016/j.cct.2015.03.005

180. Lu G, Ades A. Assessing Evidence Inconsistency in Mixed Treatment Comparisons J Am Stat Assoc. 2006;101(20):447-59 https://doi.org/10.1198/016214505000001302

181. 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. Stat Med. 2009;28(14):1861-81. PMID: 19399825. http://dx.doi.org/10.1002/sim.3594

182. Salanti G, Dias S, Welton NJ, et al. Evaluating novel agent effects in multiple-treatments meta-regression. Stat Med. 2010;29(23):2369-83. http://dx.doi.org/10.1002/sim.4001

183. Schmid CH, Trikalinos TA, Olkin I. Bayesian network meta-analysis for unordered categorical outcomes with incomplete data. Res Synth Methods. 2014;5(2):162-85. http://dx.doi.org/10.1002/jrsm.1103

184. Higgins J, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med. 1996;15(24):2733-49. PMID: 8981683. http://dx.doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0

185. Turner RM, Jackson D, Wei Y, et al. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. Stat Med. 2014. http://dx.doi.org/10.1002/sim.6381

186. Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004;23(20):3105-24. PMID: 15449338. http://dx.doi.org/10.1002/sim.1875

187. Greco T, Landoni G, Biondi-Zoccai G, et al. A Bayesian network meta-analysis for binary outcome: how to do it. Stat Methods Med Res. 2013. http://dx.doi.org/10.1177/0962280213500185

188. Hong H, Carlin BP, Shamliyan TA, et al. Comparing Bayesian and Frequentist Approaches for Multiple Outcome Mixed Treatment Comparisons. Med Decis Making. 2013;33(5):702-14. http://dx.doi.org/10.1177/0272989X13481110

189. Hong H, Chu H, Zhang J, et al. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. Res Synth Methods. 2016;7(1):6-22. PMID: 26536149. http://dx.doi.org/10.1002/jrsm.1153

190. Hawkins N, Scott DA, Woods B. 'Arm-based' parameterization for network meta-analysis. Res Synth Methods. 2015. http://dx.doi.org/10.1002/jrsm.1187 [doi]

191. Zhang J, Chu H, Hong H, et al. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. Stat Methods Med Res. 2015; PMID: 26220535. http://dx.doi.org/10.1177/0962280215596185

192. Zhang J, Carlin BP, Neaton JD, et al. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. Clin Trials. 2014;11(2):246-62. http://dx.doi.org/10.1177/1740774513498322

193. Dias S, Ades A. Absolute or relative effects? Arm-based synthesis of trial data. Res Synth Methods. 2016;7(1):23-8. PMID: 26461457. http://dx.doi.org/10.1002/jrsm.1184

194. White IR. Network meta-analysis. Stata J. 2015;15(4):951-85.

195. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2012;101(474):447-59. http://dx.doi.org/10.1198/016214505000001302

196. Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29(7-8):932-44. http://dx.doi.org/10.1002/sim.3767

197. Chaimani A, Higgins JP, Mavridis D, et al. Graphical Tools for Network Meta-Analysis in STATA. PloS One. 2013;8(10):e76654. http://dx.doi.org/10.1371/journal.pone.0076654

198. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. BMC Med Res Methodol. 2013;13:35. http://dx.doi.org/10.1186/1471-2288-13-35

199. Donegan S, Williamson P, D'Alessandro U, et al. Assessing key assumptions of network meta-analysis: a review of methods. Res Synth Methods. 2013;4(4):291-323. http://dx.doi.org/10.1002/jrsm.1085

200. Piepho HP. Network-meta analysis made easy: Detection of inconsistency using factorial analysis-of-variance models. BMC Med Res Methodol. 2014;14(1):61. http://dx.doi.org/10.1186/1471-2288-14-61

201. van der Valk R, Webers CAB, Lumley T, et al. A network meta-analysis combined direct and indirect comparisons between glaucoma drugs to rank effectiveness in lowering intraocular pressure. J Clin Epidemiol. 2009;62(12):1279-83. http://dx.doi.org/10.1016/j.jclinepi.2008.04.012

202. 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 Med Res Methodol. 2002;2(1):13. PMID: 12429069.

203. Caldwell DM, Ades A, Higgins J. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005;7521:897. PMID: 16223826 http://dx.doi.org/10.1136/bmj.331.7521.897

204. Song F, Glenny A-M, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. Control Clin Trials. 2000;21(5):488-97.

205. 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. Lancet. 2006;368(9546):1503-15. PMID: 17071284. http://dx.doi.org/10.1016/S0140-6736(06)69638-4

206. Efthimiou O, Debray TP, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. Res Synth Methods. 2016. http://dx.doi.org/10.1002/jrsm.1195

207. Neupane B, Richer D, Bonner AJ, et al. Network meta-analysis using R: a review of currently available automated packages. PloS One. 2014;9(12):e115065. http://dx.doi.org/10.1371/journal.pone.0115065

208. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64(12):1277-82. http://dx.doi.org/10.1016/j.jclinepi.2011.01.011

209. 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. http://dx.doi.org/10.1371/journal.pone.0099682

210. 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. PMID: 26085374. http://dx.doi.org/10.1136/bmj.h3326

211. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol. 2011;64(2):163-71. http://dx.doi.org/10.1016/j.jclinepi.2010.03.016

212. 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. PMID: 25005654. http://dx.doi.org/10.1001/jama.2014.5559

213. 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 J. 2012;141(2_suppl):e185S-e94S. PMID: 22315260 http://dx.doi.org/10.1378/chest.11-2289

214. Bafeta A, Trinquart L, Seror R, et al. Reporting of results from network meta-analyses: methodological systematic review. BMJ. 2014;348; PMID: 24618053. http://dx.doi.org/10.1136/bmj.g1741.

215. 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.

216. 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. http://dx.doi.org/10.1371/journal.pone.0092508

217. 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. Ann of Intern Med. 2015;162(11):777-84. http://dx.doi.org/10.7326/M14-2385