Open access • Book Chapter • DOI:10.1002/9780470317082.CH12

Statistical Methods for Meta-Analysis — Source link

Xiao-Hua Zhou, Nancy A. Obuchowski, Donna K. McClish

Institutions: Indiana University, Cleveland Clinic, Virginia Commonwealth University

Published on: 27 May 2008

STATISTICAL METHODS FOR META-ANALYSIS

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

LIFENG LIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISED BY DR. HAITAO CHU

May, 2017

Acknowledgements

I would like to express the greatest 'thank you' to my advisor Dr. Haitao Chu for his continuous support in my four-year PhD study. I feel very fortunate to have met Haitao when I entered the University and to have worked with him since then. Beyond statistical, epidemiological, and clinical knowledge, I am glad that Haitao has spent a lot of time on training my critical thinking, which has been essential for me to become an independent biostatistics researcher. Also, Haitao has given me many opportunities to review other researchers' work as a referee and get involved in preparing grant proposals. These experiences have been greatly valuable for my long-term career.

Many thanks also go to Drs. Bradley Carlin, James Hodges, and Gongjun Xu, who kindly served on my dissertation committee. They provided numerous helpful suggestions on my scientific writing as well as research topics. Moreover, I worked with Dr. Wei Pan as a research assistant for one year; his guidance let me get involved in genetic research and high-dimensional data analysis. In addition, several grants from National Institutes of Health and a Doctoral Dissertation Fellowship from the Graduate School financially supported my research in the last four years. They also provided funds to present my work in various scientific conferences. Last but not least, I would like to thank my family, especially my parents, for their love throughout my life; they are indispensable pieces for my happiness.

Abstract

Meta-analysis has become a widely-used tool to combine findings from independent studies in various research areas. This thesis deals with several important statistical issues in systematic reviews and meta-analyses, such as assessing heterogeneity in the presence of outliers, quantifying publication bias, and simultaneously synthesizing multiple treatments and factors. The first part of this thesis focuses on univariate meta-analysis. We propose alternative measures to robustly describe between-study heterogeneity, which are shown to be less affected by outliers compared with traditional measures. Publication bias is another issue that can seriously affect the validity and generalizability of meta-analysis conclusions. We present the first work to empirically evaluate the performance of seven commonly-used publication bias tests based on a large collection of actual meta-analyses in the Cochrane Library. Our findings may guide researchers in properly assessing publication bias and interpreting test results for future systematic reviews. Moreover, instead of just testing for publication bias, we further consider quantifying it and propose an intuitive publication bias measure, called the skewness of standardized deviates, which effectively describes the asymmetry of the collected studies' results. The measure's theoretical properties are studied, and we show that it can also serve as a powerful test statistic.

The second part of this thesis introduces novel ideas in multivariate meta-analysis. In medical sciences, a disease condition is typically associated with multiple risk and protective factors. Although many studies report results of multiple factors, nearly all meta-analyses separately synthesize the association between each factor and the disease condition of interest. We propose a new concept, multivariate meta-analysis of multiple factors, to synthesize all available factors simultaneously using a Bayesian hierarchical model. By borrowing information across factors, the multivariate method can improve statistical efficiency and reduce biases compared with separate analyses. In addition to synthesizing multiple factors, network meta-analysis has recently attracted much attention in evidence-based medicine because it simultaneously combines both direct and indirect evidence to compare multiple treatments and thus facilitates better decision making. First, we empirically compare two network meta-analysis models, contrast- and arm-based, with respect to their sensitivity to treatment exclusions. The arm-based method is shown to be more robust to such exclusions, mostly because it can use single-arm studies while the contrast-based method cannot. Then, focusing on the currently popular contrast-based method, we theoretically explore the key factors that make network meta-analysis outperform traditional pairwise meta-analyses. We prove that evidence cycles in the treatment network play critical roles in network meta-analysis. Specifically, network meta-analysis produces posterior distributions identical to separate pairwise meta-analyses for all treatment comparisons when a treatment network does not contain cycles. This equivalence is illustrated using simulations and a case study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Systematic reviews and meta-analyses have been frequently used to synthesize findings from multiple independent studies in many areas, including but not limited to evidence-based medicine and health care [1–5]. On an Internet Web search, the term 'meta-analysis' had 112,071 hits in PubMed dated on September 1, 2016, with 62,279 hits within last five years. This thesis deals with several important problems in meta-analysis: assessing heterogeneity (Chapter 2) and publication bias (Chapters 3 and 4), synthesizing multiple risk/protective factors (Chapter 5), and simultaneously comparing multiple treatments (Chapters 6 and 7).

The collected studies in a meta-analysis are called homogeneous if they share a common underlying true effect size; otherwise, they are called heterogeneous. A fixed-effect model is customarily used for studies deemed to be homogeneous, while a random-effects model is used for heterogeneous studies [6, 7]. Assessing heterogeneity is thus a critical issue in meta-analysis because different models may lead to different estimates of overall effect size and different standard errors. Also, the perception of heterogeneity or homogeneity helps clinicians make important decisions, such as whether the collected studies are similar enough to integrate their results and whether a treatment is applicable to all patients [8].

The classical statistic for testing between-study heterogeneity is Cochran's χ^2 test [9], also known as the Q test [10]. However, this test suffers from poor power when the number of collected studies is small, and it may detect clinically unimportant heterogeneity when many studies are pooled [11, 12]. More importantly, since the Q statistic and estimators of between-study variance depend on either the number of collected studies or the scale of effect sizes, they cannot be used to compare degrees of heterogeneity between different meta-analyses. Accordingly, Higgins and Thompson [13] proposed several measures to better describe heterogeneity. Among these, I^2 measures the proportion of total variation between studies that is due to heterogeneity rather than within-study sampling error, and it has been popular in the meta-analysis literature. Higgins and Thompson [3] empirically provided a rough guide to interpretation of I^2: $0 \leq I^2 \leq 0.4$ indicates that heterogeneity might not be important; $0.3 \leq I^2 \leq 0.6$ may represent moderate heterogeneity; $0.5 \leq I^2 \leq 0.9$ may represent substantial heterogeneity; and $0.75 \leq I^2 \leq 1$ implies considerable heterogeneity. These ranges overlap because the importance of heterogeneity depends on several factors and strict thresholds can be misleading [3].

Ideally, if heterogeneity is present in a meta-analysis, it should *permeate* the entire collection of studies instead of being limited to a small number of outlying studies. With this in mind, we may classify meta-analyses into four groups: (i) all collected studies are homogeneous; (ii) a few studies are outlying and the rest are homogeneous; (iii) heterogeneity permeates the entire collection of studies; and (iv) a few studies are outlying and heterogeneity permeates the remaining studies. Outlying studies can have great impact on conventional heterogeneity measures and on the conclusions of a meta-analysis. Several methods have been recently developed for detecting outliers and influential data in meta-analysis [14, 15]. However, no widely accepted guidelines exist for handling outliers in the statistical literature, including the area of meta-analysis. Hedges and Olkin [16] specified two extreme positions about dealing with outlying studies: (i) data are 'sacred', and no study should ever be set aside for any reason; or (ii) data should be tested for outlying studies, and those failing to conform to the hypothesized model should be removed. Neither seems appropriate. Alternatively, if a small number of studies is influential, some researchers usually present sensitivity analyses with and without those studies. However, if the results of sensitivity analysis differ dramatically, clinicians may reach no consensus about which result to use to make decisions. Because of these problems caused by outliers, ideal heterogeneity measures are expected to be robust: they should be minimally affected by outliers and accurately describe heterogeneity. Chapter 2 will introduce new heterogeneity measures that are less affected by outliers than conventional measures.

Like outliers and high heterogeneity between studies, publication bias also seriously threatens the validity and generalizability of conclusions of systematic reviews—studies with statistically significant findings are more likely to be published than those reporting statistically non-significant findings—thus the overall treatment effect may be overestimated [17–21]. Therefore, assessing publication bias has been a critical topic in systematic review and meta-analysis.

A traditional and intuitive method for assessing publication bias is to examine the asymmetry of a funnel plot, which usually plots effect sizes vs. their corresponding precisions or standard errors [22,23]. In the presence of publication bias, the funnel plot is expected to be asymmetric. However, the visual examination is usually subjective. Based on the funnel plot, Begg's rank test, Egger's regression test, and the trim and fill method have been proposed to statistically test publication bias, and they are widely applied [24–26]. The trim and fill method is attractive because it not only detects but also adjusts for publication bias; nevertheless, it makes rather strong assumptions about the treatment effects of potentially suppressed studies, and the adjusted overall effect estimate is generally recommended as a form of sensitivity analysis [27]. Begg's and Egger's tests aim at examining the association between the observed treatment effects and their standard errors; a strong association leads to an asymmetric funnel plot and implies publication bias. The original Egger's test regresses the standardized treatment effect (i.e., effect size divided by its standard error) on the corresponding precision (i.e., the inverse of standard error). This regression can be shown to be equivalent to a weighted regression of the treatment effect on its standard error, weighted by the inverse of its variance [28]. The weighted regression version has become more familiar among meta-analysts, probably because it directly links the treatment effects to their precisions without standardizing. Several modifications of Egger's test also use the technique of weighted regression—the dependent variable is still the treatment effect, but the independent variable differs. For example, Tang and Liu [29] suggested an alternative test motivated by the sample-size-based funnel plot, in which the treatment effect is presented against the total sample size of each study. Tang's regression test basically performs weighted regression of the treatment effect on the inverse of the square root of study-specific sample size.

When study outcomes are binary, the commonly-used effect size, log odds ratio, is mathematically associated with its standard error, even in the absence of publication bias [30, 31]. Although it is infeasible to accurately evaluate this association's strength, several authors have concerns that Begg's and Egger's methods may have inflated type I error rates for binary outcomes due to the potential association, and alternative regression tests have been designed specifically to deal with this issue [31–33]. For example, Macaskill et al. [32] regressed the log odds ratio on the study-specific total sample size. Deeks et al. [31] used the 'effective sample size' (see its definition in Table 3.1) as the regression independent variable, and Peters et al. [33] slightly modified Macaskill's regression and used the inverse of the total sample size as the independent variable. Table 3.1 briefly describes these approaches; more details are provided by Sterne et al. [34].

The various approaches have been widely applied to assess publication bias in systematic reviews, and some of them have been compared in extensive simulation studies [33,35,36]. It is generally recognized that Begg's rank test has lower statistical power than the others based on their performance in simulations [28, 30, 32]. However, most meta-analysis articles only perform one or two publication bias tests, and so far the performance of the various tests has not been systematically and comprehensively evaluated using published meta-analysis datasets. In addition, some simulation settings can be fairly unrealistic; for example, studies may be suppressed because of non-significant P-values [24], or negative effect sizes [26], or other obscure editorial criteria, and the exact mechanism of publication bias in a real meta-analysis can never be reproduced by simulations. Instead of just conducting simulation studies, Chapter 3 evaluates seven commonly-used publication bias tests using a large collection of actual meta-analyses published in the Cochrane Library.

Besides the aforementioned funnel-plot-based methods, another class of approaches to detecting publication bias is based on selection models. These approaches typically use weighted distribution theory to model the selection (i.e., publication) process and develop estimation procedures that account for the selection process; see, e.g., [37–40]. Sutton et al. [41] provide a comprehensive review. The selection models are usually complicated, limiting their applicability. Moreover, they incorporate weight functions in an effort to correct publication bias, but strong and largely untestable assumptions are often made [41]. Therefore, the validity of their adjusted results may be doubtful, and these methods are usually employed as sensitivity analyses.

In addition to detecting publication bias using selection models and funnel-plot-based methods, it is also important to *quantify* publication bias by measures that permit comparisons between different meta-analyses. A candidate measure is the intercept of the regression test [25]. However, as a measure of asymmetry of the collected study results, the regression intercept lacks a clear interpretation; for example, it is difficult to provide a range guideline to determine mild, moderate, or substantial publication bias based on the regression intercept. Due to this limitation, meta-analysts usually report the P-value of Egger's regression test, but not the magnitude of the intercept. We will show that the regression intercept basically estimates the average of study-specific standardized deviates; it does not account for the shape of the deviates' distribution, which is skewed in the presence of publication bias. This may limit the statistical power of Egger's regression test. Chapter 4 introduces a new measure of publication bias: the skewness of the standardized deviates. It not only has an intuitive interpretation as the asymmetry of the collected study results but also can serve as a powerful test statistic.

Beyond univariate meta-analysis, statistical methods for multivariate meta-analysis are increasingly popular in the era of big data. This thesis will introduce innovative ideas when synthesizing multiple factors and treatments. As a disease condition is typically associated with many risk and protective factors in medical sciences, many randomized controlled trials and observational studies considered multiple factors [42–45]. Reliable summaries of association between each factor and the disease condition are crucial for the design of a multi-factor intervention program. The growth of interest in evidence-based medicine has led to a dramatic increase in attention paid to systematic reviews and meta-analyses. In prevention studies, it has become increasingly popular to perform meta-analyses on multiple risk and protective factors to summarize existing evidence; however, currently, nearly all meta-analyses are performed on each factor separately [46–48]. Different studies usually focus on different subsets of all risk and protective factors, and may only selectively report some significant factors in peer-reviewed articles; some factors may be reported by only a few studies. Hence, if we organize the collected data in a matrix with rows and columns indexing studies and factors respectively, then the data matrix is expected to contain many missing entries; see the example in Table 5.1. A conventional meta-analysis separately estimates each factor's association with the disease condition, so it cannot borrow information from the correlations between factors. Moreover, results from separate meta-analyses may not be directly comparable because they may be based on different subpopulations. This limits medical investigators as they select most important factors for the design of a multi-factor intervention program.

Recently, Serghiou et al. [49] introduced field-wide systematic review and meta-analysis to report and assess the entire field of putative factors for a disease condition. Based on this concept, researchers can learn the selective availability and different adjustments of factors and the patterns of modeling. Although multiple factors were collected, the authors pooled the results for each factor separately; this is not efficient to analyze the multivariate data from a field-wide systematic review. Chapter 5 proposes *multivariate meta-analysis of multiple factors* to jointly synthesize all risk and protective factors in the field-wide systematic review. This method is shown to produce better estimates of association measures between the factors and the disease condition, compared with separate meta-analyses.

A disease condition can also have multiple treatments in medical sciences. Comparative effectiveness research is aimed at informing health care decisions concerning the benefits and risks of different diagnostic and intervention options. The growing number of treatment options for a given condition, as well as the rapid escalation in their cost, has created a greater need for rigorous comparisons of multiple treatments in clinical practice. To simultaneously compare multiple treatments for a given condition, network meta-analysis methods, also known as mixed treatment comparisons, have recently been developed, expanding the scope of conventional pairwise meta-analysis. Network meta-analysis simultaneously synthesizes both direct comparisons of interventions within randomized controlled trials and indirect comparisons across trials [50–56]. Based on an Internet Web search, the prestigious medical journals *BMJ*, *JAMA*, and *Lancet* have published more than 100 research articles with the term ‘network meta-analysis’ in their titles since 2010.

Currently, much effort in network meta-analysis has been devoted to contrast-based approaches, which focus on investigating relative treatment effects, such as odds ratios when the outcome is binary. However, population-averaged absolute risks may be preferred in some situations such as cost-effectiveness analysis [57,58]. In addition, relative treatment effects are sometimes insufficient for patients to make decisions. For instance, consider a patient's choice between treatments A and B with the following two sets of one-year survival rates: (i) $\pi_{\text{A}} = 0.8$ vs. $\pi_{\text{B}} = 0.5$; (ii) $\pi_{\text{A}} = 0.004$ vs. $\pi_{\text{B}} = 0.001$. Most likely, patients will strongly prefer treatment A in scenario (i) but have little preference in scenario (ii), yet both have odds ratio 4.0.

Contrast-based network meta-analysis can back-transform odds ratios to population-averaged absolute risks only if the absolute risk of a given reference treatment group can be accurately estimated from external data, or can be estimated using a separate model to analyze responses for the reference arm from the network [57, 58]. Both approaches depend on strong assumptions. For the approach using external data, even if such external data are available, they may come from a population different from the one represented in the network meta-analysis, and the assumption of transitivity of relative effects on the odds ratio scale (i.e., that treatment effects are independent of baseline risks) is rather strong. The choice of the odds ratio scale is generally arbitrary or conventional, and there is no particular reason to expect effects in different trials to be exchangeable on the odds ratio scale. For the approach using a distinct model for the reference arm, under the theory of missing data, this analysis is unbiased only under a strong assumption of missing completely at random, i.e., that each study randomly chooses which treatment arms to include. In addition, if the estimation of absolute effects uses a subset of the same trials used for the estimation of relative effects, then the estimated absolute and relative effects are not independent. Thus, one would need to model the correlations among the two sets of estimates, which is not straightforward. Finally, the back-transformed relative risks and risk differences can be noticeably different depending on which treatment is chosen as the reference group, even with exactly the same model and priors (Appendix A.7 gives an example). These considerations suggest methodological limitations in contrast-based methods for estimating population-averaged absolute risks.

When performing a network meta-analysis, selecting appropriate treatments for the systematic review is crucial, as this will influence the validity and generalizability of both the direct and indirect evidence summarized in the analysis. However, no guidelines exist for treatment selection. Because the control treatment may not be defined consistently across trials, some have suggested excluding such control treatments from a network meta-analysis [59–61], but others have argued that having no comparison between an active intervention and placebo is problematic [62–64]. Moreover, the treatments of interest may differ in different countries, and may vary over time due to introduction of new drugs [65]. Therefore, the treatment arms included in a network meta-analysis usually consist of a subset of a more extensive network. Using a contrast-based network meta-analysis [50, 51], Mills et al. [66] examined the sensitivity of estimated effect sizes such as odds ratio to removal of certain treatments. They concluded that excluding a treatment sometimes has substantial influence on estimated effect sizes. Consequently, selection of treatment arms should be carefully considered when applying network meta-analysis.

Chapter 6 examines the sensitivity to treatment exclusion of an alternative arm-based approach to network meta-analysis, which has recently been developed from the perspective of missing data analysis [67]. The difference between the contrast- and arm-based approaches is substantial, and it is almost entirely due to single-arm trials. When a treatment is removed from a contrast-based network meta-analysis, it is necessary to exclude other treatments in two-arm studies that investigated the excluded treatment; such exclusions are not necessary in arm-based network meta-analysis, leading to substantial gain in performance.

As mentioned above, the Lu–Ades contrast-based Bayesian hierarchical model [51,54] is still the most popular method to implement network meta-analysis, and it is generally considered more powerful than conventional pairwise meta-analysis, leading to more accurate effect estimates with narrower confidence intervals. However, the improvement of effect estimates produced by Lu–Ades network meta-analysis has never been studied theoretically. Chapter 7 shows that such improvement depends highly on evidence cycles in the treatment network. Specifically, Lu–Ades network meta-analysis produces posterior distributions identical to separate pairwise meta-analyses for all treatment comparisons when a treatment network does not contain cycles. Even in a general network with cycles, treatment comparisons that are not contained in any cycles do not benefit from Lu–Ades network meta-analysis. Simulations and a case study will be used to illustrate the equivalence of Lu–Ades network meta-analysis and pairwise meta-analysis in certain networks.

Chapter 8 summarizes the major findings in this thesis and introduces some related future topics.

Chapter 2

Alternative Measures of Between-Study Heterogeneity in Meta-Analysis: Reducing the Impact of Outlying Studies

This chapter introduces several new heterogeneity measures, which are designed to be less affected by outliers than conventional measures. The basic idea comes from least absolute deviations (LAD) regression, which is known to have significant robustness advantages over classical least squares (LS) regression [68]. Specifically, LS regression aims at minimizing the sum of *squared* errors $\sum(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2$, where $\boldsymbol{x}_i$ represents predictors, y_i is the response, and $\boldsymbol{\beta}$ contains the regression coefficients. LAD regression minimizes the sum of *absolute* errors $\sum |y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}|$. The impact of outliers is diminished by using absolute values in LAD regression, compared to using squared values in LS regression. In meta-analysis, the conventional Q statistic has the form $Q = \sum w_i(y_i - \bar{\mu})^2$, where the y_i's are the observed effect sizes, the w_i's are study-specific weights, and $\bar{\mu}$ is the weighted average effect size. Analogously, we consider a new measure $Q_r = \sum \sqrt{w_i}|y_i - \bar{\mu}|$, which is expected to be more robust against outliers than the conventional Q. An estimate of the between-study variance can be obtained based on Q_r. Also, since Q_r depends on the number of collected studies, we further derive two statistics to quantify heterogeneity, which are counterparts of I^2 and another statistic H also proposed by Higgins and Thompson [13].

This chapter is organized as follows. Section 2.1 gives a brief review of conventional measures and discusses the dilemma of handling outliers in meta-analysis. Section 2.2 proposes several new heterogeneity measures designed to be robust to outliers. Section 2.3 uses theoretical properties to compare the proposed and conventional measures. Section 2.4 presents simulations to compare the various approaches empirically, and Section 2.5 applies the approaches to two actual meta-analyses. Section 2.6 provides a brief discussion.

2.1 The conventional methods

2.1.1 Measures of between-study heterogeneity

Suppose that a meta-analysis contains n independent studies. Let μ_i be the underlying true effect size, such as log odds ratio, in study i ($i = 1, \ldots, n$). Typically, published studies report estimates of the effect sizes and their within-study variances, which we will call y_i and s_i^2. It is customary to assume that the y_i's are approximately normally distributed with mean μ_i and variance σ_i^2, respectively. Since the unknown σ_i^2 can be estimated by s_i^2, these data are commonly modeled as $y_i \sim N(\mu_i, s_i^2)$ with s_i^2 treated as known. Also, we assume that the true μ_i's are independently distributed as $\mu_i \sim N(\mu, \tau^2)$, where μ is the true overall mean effect size across studies and τ^2 is the between-study variance. The collected n studies are defined to be homogeneous if their underlying true effect sizes are equal, that is, $\mu_i = \mu$ for all $i = 1, \ldots, n$, or equivalently $\tau^2 = 0$. On the other hand, the studies are heterogeneous if their underlying true effect sizes vary, that is, $\tau^2 > 0$.

To test the homogeneity of the y_i's (i.e., H_0: $\tau^2 = 0$ vs. H_A: $\tau^2 > 0$), the well-known Q statistic [10] is defined as

$$Q = \sum_{i=1}^{n} w_i (y_i - \bar{\mu})^2,$$

which follows a χ^2_{n-1} distribution under the null hypothesis. Here, $w_i = 1/s_i^2$ is the reciprocal of the within-study variance of y_i, and $\bar{\mu} = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$ is the pooled fixed-effect estimate of μ. Based on the Q statistic, DerSimonian and Laird [69] introduced a method of moments estimate of the between-study variance,

$$\widehat{\tau}^2_{\mathrm{DL}} = \max\left\{0, \frac{Q-(n-1)}{\sum_{i=1}^n w_i - \sum_{i=1}^n w_i^2/\sum_{i=1}^n w_i}\right\}.$$

Note that the Q statistic depends on the number of collected studies n and the estimate of between-study variance depends on the scale of effect sizes. Hence, neither Q nor $\widehat{\tau}^2_{\mathrm{DL}}$ can be used to compare degrees of heterogeneity between different meta-analyses. To allow such comparisons, Higgins and Thompson [13] proposed the measures H and I^2:

$$H = \sqrt{Q/(n-1)}, \quad I^2 = [Q-(n-1)]/Q.$$

The H statistic is interpreted as the ratio of the standard deviation of the estimated overall effect size from a random-effects meta-analysis compared to the standard deviation from a fixed-effect meta-analysis; I^2 describes the proportion of total variance between studies that is attributed to heterogeneity rather than sampling error. In practice, meta-analysts truncate H at 1 when $H < 1$ and truncate I^2 at 0 when $I^2 < 0$; therefore, $H \geq 1$ and I^2 lies between 0 and 1. Since I^2 is interpreted as a proportion, it is usually expressed as a percent. Both measures have been widely adopted in practice.

2.1.2 Outlier detection

As in many other statistical applications, outliers frequently appear in meta-analysis. Outliers may arise from at least three sources:

(i) *The quality of collected studies and systematic review.* The published results (y_i, s_i^2) in a clinical study could be outlying due to errors in the process of recording, analyzing, or reporting data. Also, the populations in certain clinical studies may not meet the systematic review's inclusion criteria; hence, such studies may be outlying compared to most other collected studies.

(ii) *A heavy-tailed distribution of study-specific underlying effect sizes.* Conventionally, at the between-study level, the study-specific underlying effect sizes μ_i are assumed to have a normal distribution. However, the true distribution of the μ_i's may greatly depart from the normality assumption and have heavy tails, such as the t-distribution with small degrees of freedom.

(iii) *Small sample sizes in certain studies.* The true within-study variances σ_i^2 could be poorly estimated by the sample variances s_i^2 if the sample sizes are small. In some situations, effect sizes in small studies may be more informative than large studies due to 'small study effects' [70]; if their true within-study variances σ_i^2 are seriously underestimated, then small studies could be outlying.

Hedges and Olkin [16] and Viechtbauer and Cheung [14] introduced outlier detection methods for fixed-effect and random-effects meta-analyses, respectively. Both methods use a 'leave-one-study-out' technique so that a potential outlier could have little influence on the residuals of interest. Specifically, the residual of study i is calculated as $e_i = y_i - \bar{\mu}_{(-i)}$. Here, $\bar{\mu}_{(-i)}$ is the estimated overall effect size using the data without study i; that is, $\bar{\mu}_{(-i)} = \frac{\sum_{j \neq i} y_j / s_j^2}{\sum_{j \neq i} 1/s_j^2}$ under the fixed-effect setting, and $\bar{\mu}_{(-i)} = \frac{\sum_{j \neq i} y_j/(s_j^2 + \hat{\tau}^2_{(-i)})}{\sum_{j \neq i} 1/(s_j^2 + \hat{\tau}^2_{(-i)})}$ under the random-effects setting, where $\hat{\tau}^2_{(-i)}$ can be the DerSimonian and Laird estimate using the data without study i. The variance of e_i is estimated as $v_i = s_i^2 + (\sum_{j \neq i} 1/s_j^2)^{-1}$ and $v_i = s_i^2 + \hat{\tau}^2_{(-i)} + [\sum_{j \neq i} 1/(s_j^2 + \hat{\tau}^2_{(-i)})]^{-1}$ under the fixed-effect and random-effects settings, respectively. The standardized residuals $\epsilon_i = e_i/\sqrt{v_i}$ are expected to follow the standard normal distribution and studies with ϵ_i's greater than 3 in absolute magnitude are customarily considered outliers.

Outliers may be masked if the above approaches are used in an inappropriate setting. For example, Figures 2.3(b) and 2.3(d) in Section 2.5 show standardized residuals of two actual meta-analyses; different outlier detection methods identify different outliers. Hence, one must assess the heterogeneity of collected studies to correctly apply the foregoing approaches to detect outliers. However, outliers may cause heterogeneity to be overestimated and thus affect procedures to detect them. Additionally, even if outliers are identified, there is no consensus in the statistical literature on what to do about them unless these studies are evidently erroneous [71]. To avoid the dilemmas of detecting and handling outliers, we propose robust measures to assess heterogeneity.

2.2 The proposed alternative heterogeneity measures

2.2.1 Measures based on absolute deviations and weighted average

In linear regression, it is well-known that least absolute deviations regression is more robust to outliers than classical least squares regression [68]. The former method minimizes $\sum |y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}|$ and the latter minimizes $\sum (y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2$, where $\boldsymbol{x}_i$ and y_i are predictors and response respectively and $\boldsymbol{\beta}$ contains the regression coefficients. In the context of meta-analysis, the conventional Q statistic is analogous to least squares regression, because Q is a weighted sum of *squared* deviations. To reduce the impact of outlying studies, we propose a new measure Q_r which is analogous to least absolute deviations regression. This measure is the weighted sum of *absolute* deviations, and is defined as

$$Q_r = \sum_{i=1}^{n} \sqrt{w_i}|y_i - \bar{\mu}|.$$

For random-effects meta-analysis, $\mathrm{E}[Q_r] = \sum_{i=1}^{n} \sqrt{2v_i/\pi}$, where $v_i = 1 - w_i/\sum_{j=1}^{n} w_j + \tau^2[w_i - 2w_i^2/\sum_{j=1}^{n} w_j + w_i \sum_{j=1}^{n} w_j^2/(\sum_{j=1}^{n} w_j)^2]$.

DerSimonian and Laird [69] derived an estimate of the between-study variance τ^2 based on the Q statistic by the method of moments, i.e., equating the observed Q with its expectation. We can similarly obtain a new estimate of τ^2, denoted as $\widehat{\tau}_r^2$, from the proposed Q_r statistic. Specifically, $\widehat{\tau}_r^2$ is the solution to the following equation in τ^2:

$$Q_r\sqrt{\frac{\pi}{2}} = \sum_{i=1}^{n} \left\{ 1 - \frac{w_i}{\sum_{j=1}^{n} w_j} + \tau^2 \left[w_i - \frac{2w_i^2}{\sum_{j=1}^{n} w_j} + \frac{w_i \sum_{j=1}^{n} w_j^2}{(\sum_{j=1}^{n} w_j)^2} \right] \right\}^{1/2}. \tag{2.1}$$

If this equation has no nonnegative solution, set $\widehat{\tau}_r^2 = 0$. Note that the right-hand side of Equation (2.1) is monotone increasing in τ^2, so the solution is unique.

The Q_r statistic, like Q, is dependent on the number of studies; $\widehat{\tau}_r^2$, like $\widehat{\tau}_{\mathrm{DL}}^2$, is dependent on the scale of effect sizes. Following the approach of Higgins and Thompson [13], we tentatively assume that all studies share a common within-study variance σ^2 and explore heterogeneity measures that are independent of both the number of studies and the scale of effect sizes, so that they can be used to compare degrees of heterogeneity between meta-analyses. Suppose the target heterogeneity measure can be written as $f(\mu, \tau^2, \sigma^2, n)$, which is a function of the true overall mean effect size μ, the between-study variance τ^2, the within-study variance σ^2, and the number of studies

n. Higgins and Thompson [13] suggested that this measure should satisfy the following three criteria:

(i) (Dependence on the magnitude of heterogeneity) $f(\mu, \tau'^2, \sigma^2, n) > f(\mu, \tau^2, \sigma^2, n)$ for any $\tau'^2 > \tau^2$. This criterion is self-evident.

(ii) (Scale invariance) $f(a + b\mu, b^2\tau^2, b^2\sigma^2, n) = f(\mu, \tau^2, \sigma^2, n)$ for any constants a and b. This criterion 'standardizes' comparisons between meta-analyses using different scales of measurement and different types of outcome data.

(iii) (Size invariance) $f(\mu, \tau^2, \sigma^2, n') = f(\mu, \tau^2, \sigma^2, n)$ for any positive integers n and n'. This criterion indicates that the number of studies collected in meta-analysis does not systematically affect the magnitude of the heterogeneity measure.

Monotone increasing functions of $\rho = \tau^2/\sigma^2$ can be easily shown to satisfy these three criteria. Plugging $w_i = 1/\sigma^2$ into Equation (2.1), we have $\rho + 1 = \pi Q_r^2/[2n(n-1)]$. This implies that

$$H_r^2 = \frac{\pi Q_r^2}{2n(n-1)}$$

is a candidate measure. Further, considering $\rho/(\rho+1) = \tau^2/(\tau^2+\sigma^2)$, commonly called the intraclass correlation, Equation (2.1) yields another candidate:

$$I_r^2 = \frac{Q_r^2 - 2n(n-1)/\pi}{Q_r^2}.$$

In practice, H_r would be truncated at 1 when $H_r < 1$ and I_r^2 would be truncated at 0 when $I_r^2 < 0$. These two measures, H_r^2 and I_r^2, are analogous to and have the same interpretations as H^2 and I^2, respectively. Higgins and Thompson [13] also introduced a so-called R^2 statistic; since it has interpretation and performance similar to H^2, we do not present a version of R^2 based on the new Q_r statistic.

Since standard deviations are used more frequently in clinical practice, Higgins and Thompson [13] suggested reporting H, instead of H^2, for meta-analyses. For the proposed measures, we also recommend reporting H_r rather than H_r^2. However, we suggest presenting I^2 and I_r^2 instead of their square roots because their interpretation of 'proportion of variance explained' is widely familiar to clinicians. $H_r = 1$ or $I_r^2 = 0$ implies perfect homogeneity. Also, since the expressions for H_r and I_r^2 only involve Q_r and

n but not within-study variances, these two measures can be easily generalized to a situation where the within-study variances s_i^2 vary across studies.

2.2.2 Measures based on absolute deviations and weighted median

The proposed Q_r statistic uses the weighted average $\bar{\mu}$ to estimate overall effect size under the null hypothesis; it may be sensitive to potential outliers. To derive an even more robust heterogeneity measure, we may replace the weighted average with the weighted median $\widehat{\mu}_m$, which is defined as the solution to the following equation in θ:

$$\sum_{i=1}^{n} w_i \left[\mathbb{I}(\theta \geq y_i) - 0.5\right] = 0, \tag{2.2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This weighted median leads to a new test statistic, $Q_m = \sum_{i=1}^{n} \sqrt{w_i}|y_i - \widehat{\mu}_m|$. Note that the solution to Equation (2.2) may be not unique; to avoid this problem, we will approximate the indicator function by a monotone increasing smooth function [72]. Section 2.2.3 introduces the details.

The expectation of Q_m may not be explicitly calculated because the distribution of weighted median of finite samples is unclear. By the theory of M-estimation [73], the weighted median is a $\sqrt{n}$-consistent estimator of the true overall effect size μ. Suppose that the weights w_i have finite first-order moment, then it can be shown that

$$\left| Q_m/n - \frac{1}{n}\sum_{i=1}^{n} \sqrt{w_i}|y_i - \mu| \right| \leq |\widehat{\mu}_m - \mu| \cdot \frac{1}{n}\sum_{i=1}^{n} \sqrt{w_i} = O_p(n^{-1/2}).$$

Therefore, when the number of collected studies n is large,

$$\mathrm{E}[Q_m/n] \approx \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n} \sqrt{w_i}|y_i - \mu|\right] = \frac{1}{n}\sqrt{2/\pi}\sum_{i=1}^{n} \sqrt{(s_i^2 + \tau^2)/s_i^2}.$$

By equating the Q_m statistic to its approximated expectation, a new estimator of between-study variance $\widehat{\tau}_m^2$ can be derived as the solution to

$$Q_m\sqrt{\pi/2} = \sum_{i=1}^{n} \sqrt{(s_i^2 + \tau^2)/s_i^2}$$

in τ^2. If all within-study variances are further assumed to be equal to a common value σ^2 as in Section 2.2.1, $\mathrm{E}[Q_m/n] \approx \sqrt{2/\pi}\sqrt{(\sigma^2 + \tau^2)/\sigma^2}$. Based on Q_m, the counterparts of H_r^2 and I_r^2—which assess $(\sigma^2+\tau^2)/\sigma^2$ and $\tau^2/(\sigma^2+\tau^2)$ respectively—are defined as

$$H_m^2 = \frac{\pi Q_m^2}{2n^2}, \quad I_m^2 = \frac{Q_m^2 - 2n^2/\pi}{Q_m^2}.$$

Note that many meta-analyses only collect a small number of studies; however, the derivation of $\hat{\tau}_m^2$, H_m^2, and I_m^2 assumes a large n. The finite-sample performance of these heterogeneity measures will be studied using simulations.

2.2.3 Calculation of P-values and confidence intervals

Due to the difficulty caused by summing the absolute values of correlated random variables in the expression of Q_r and the intractable distribution of weighted median in Q_m, it is not feasible to explicitly derive the probability and cumulative density functions for the proposed statistics. Instead, resampling method can be used to calculate P-values and 95% confidence intervals (CIs). Since the weighted median in Q_m is discontinuous and may be not unique due to the indicator function in Equation (2.2), we apply the approach by Horowitz [72] to approximate the indicator function $\mathbb{I}(t > 0)$ by a smooth function $J(t)$ in the following simulations and case studies. For example, $J(t)$ can be the scaled expit function $J_\epsilon(t) = 1/[1 + \exp(-t/\epsilon)]$, where ϵ is a pre-specified small constant. We use $\epsilon = 10^{-4}$; various choices of ϵ are shown to produce stable results in Appendix A.1.

Parametric resampling can be used to calculate a P-value for Q_r; similar procedures can also be used for Q and Q_m. First, estimate the overall effect size $\bar{\mu}$ under $H_0 : \tau^2 = 0$ (i.e., the fixed-effect setting) and calculate the Q_r statistic based on the original data. Second, draw n samples under H_0, $y_i^\star \sim N(\bar{\mu}, s_i^2)$, and repeat this for B (say 10,000) iterations. Here, the weighted average $\bar{\mu}$ is used to estimate μ because it is unbiased and may have smaller variance than the weighted median under the null hypothesis. Third, based on the B sets of bootstrap samples, calculate the Q_r statistic as $Q_r^{(b)}$ for $b = 1, \ldots, B$. Finally, the P-value is calculated as $P = \left[\sum_{b=1}^B \mathbb{I}(Q_r^{(b)} > Q_r) + 1\right]/(B+1)$. Here, 1 is added to both numerator and denominator to avoid calculating $P = 0$. To derive 95% CIs for the various heterogeneity measures, the nonparametric bootstrap can be used by taking samples of size n with replacement from the original data $\{(y_i, s_i^2)\}_{i=1}^n$ and calculating 2.5% and 97.5% quantiles for each of the measures over the bootstrap samples.

2.3 The relationship between I^2, I_r^2, and I_m^2

2.3.1 When the number of studies is fixed

Since I_r^2 and I_m^2 are designed to be robust compared to the conventional I^2, they are expected to be smaller than I^2 in the presence of outliers. Applying the Cauchy-Schwarz Inequality, $Q_r^2 \leq nQ$, and the equality holds if and only if each $w_i(y_i - \bar{\mu})^2$ equals a common value for all studies, in which case outliers are unlikely to appear. The foregoing inequality further implies $H_r \leq H\sqrt{\pi/2}$ and $I_r^2 \leq I^2 + (1 - 2/\pi)(1 - I^2)$. Therefore, the proposed H_r and I_r^2 are not always smaller than H and I^2, respectively; I_r^2 may be greater than I^2 by up to $(1 - 2/\pi)(1 - I^2)$. Appendix A.2 provides artificial meta-analyses to illustrate how the proposed measures may have better interpretations even when no outliers are present; I_r^2 and I_m^2 are larger than I^2 in those examples. As I_m^2 is based on the intractable weighted median, determining its relationship with I^2 and I_r^2 is not feasible in finite samples except by simulations. Alternatively, the asymptotic values of the three measures can be derived as $n \to \infty$; Section 2.3.2 considers this case.

2.3.2 When the number of studies becomes large

This section focuses on the asymptotic properties of the three heterogeneity measures as the number of collected studies $n \to \infty$. Denote $\xrightarrow{P}$ as convergence in probability, and let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. We have the following two propositions if no outliers are present.

Proposition 1. *Under the fixed-effect setting, the observed effect sizes are $y_i \sim N(\mu, s_i^2)$. Assume that the weights $w_i = 1/s_i^2$ are independent and identically distributed with finite positive mean, and independent of the y_i's. Then I^2, I_r^2, and I_m^2 converge to 0 in probability as $n \to \infty$.*

Proposition 2. *Assume that all studies share a common within-study variance σ^2. Under the random-effects setting, the observed effect sizes are $y_i \sim N(\mu_i, \sigma^2)$ and $\mu_i \sim N(\mu, \tau^2)$; hence, the true proportion of total variation between studies due to heterogeneity is $I_0^2 = \tau^2/(\sigma^2 + \tau^2)$. Then I^2, I_r^2, and I_m^2 converge to the true I_0^2 in probability as $n \to \infty$.*

Propositions 1 and 2 show that, for either homogeneous or heterogeneous studies, all three heterogeneity measures converge to the true value and correctly indicate homogeneity or heterogeneity. Proposition 1 does not require that the n studies have a common within-study variance; Proposition 2 makes this assumption to facilitate definition of the true I_0^2. The following proposition compares the three measures when the collection of studies is contaminated by a certain proportion of outlying studies.

Proposition 3. *Assume that all studies share a common within-study variance σ^2. The observed effect sizes are $y_i \sim N(\mu_i, \sigma^2)$. The meta-analysis is supposed to focus on a certain population of interest, and in this population, the study-specific underlying effect sizes are $\mu_i \sim N(\mu, \tau^2)$; therefore, the true proportion of total variation between studies in this population that is due to heterogeneity is $I_0^2 = \tau^2/(\sigma^2 + \tau^2)$. However, 100η percent of the n studies are mistakenly included, having been conducted on inappropriate populations; their study-specific underlying effect sizes are $\mu_i \sim N(\mu + C, \tau^2)$, where C is a constant, representing the discrepancy of outliers. Then, as $n \to \infty$,*

$$I^2 \xrightarrow{P} 1 - [(1 - I_0^2)^{-1} + r_1 r_2]^{-1};$$

$$I_r^2 \xrightarrow{P} h(r_1, r_2; \eta, I_0^2);$$

$$I_m^2 \xrightarrow{P} h(s_1, s_2; \eta, I_0^2).$$

Here, $h(\cdot, \cdot; \eta, I_0^2)$ is a function depending on η and I_0^2 defined as

$$\begin{aligned} h(t_1, t_2; \eta, I_0^2) = 1 - \Bigg\{ & \eta \left[(1 - I_0^2)^{-1/2} \exp\left(-\frac{1}{2} t_1^2 (1 - I_0^2) \right) + \sqrt{\frac{\pi}{2}} t_1 \left(1 - 2\Phi\left(-t_1 (1 - I_0^2)^{1/2} \right) \right) \right] \\ & + (1 - \eta) \left[(1 - I_0^2)^{-1/2} \exp\left(-\frac{1}{2} t_2^2 (1 - I_0^2) \right) - \sqrt{\frac{\pi}{2}} t_2 \left(1 - 2\Phi\left(t_2 (1 - I_0^2)^{1/2} \right) \right) \right] \Bigg\}^{-2}; \end{aligned}$$

also, $r_1 = (1 - \eta) C/\sigma$, $r_2 = \eta C/\sigma$, $s_2 = C/\sigma - s_1$, and s_1 is the solution to

$$\eta \Phi\left(-s_1 (1 - I_0^2)^{1/2} \right) + (1 - \eta) \Phi\left((C/\sigma - s_1)(1 - I_0^2)^{1/2} \right) = 0.5.$$

Appendix B.1 gives proofs of the three propositions. Proposition 3 suggests that all three heterogeneity measures are affected by outlying studies, though to different degrees. Specifically, their asymptotic values are determined by three factors: the true proportion of total variation between studies that is due to heterogeneity I_0^2, the proportion of outliers η, and the ratio of the discrepancy of the outliers C compared to the within-study standard deviation σ, that is, $R = C/\sigma$. Outliers are usually present in small quantities, so the proportion of outliers η is usually not large. Also, an observation is customarily considered an outlier if the distance to the overall mean is greater than three times the standard deviation σ; therefore, the ratio R is usually greater than 3.

Figure 2.1 compares the asymptotic values of the three heterogeneity measures derived in Proposition 3. The upper panels show the setting of true homogeneity ($I_0^2 = 0$) and the lower panels show the setting of true heterogeneity ($I_0^2 = 0.5$). Under each setting, the proportion of outliers is 1%, 5%, or 10%. Clearly, all panels present a common trend: the three heterogeneity measures increase as R increases. When η is 1%, I_r^2 and I_m^2 are much less affected by outliers than I^2, indicating the robustness of the proposed measures. Also, I_m^2 is a bit smaller than I_r^2. As η increases, the difference between I^2 and I_r^2 becomes smaller, while the difference between I_r^2 and I_m^2 becomes larger though it is never substantial. This implies that I_m^2 is the most robust measure when a meta-analysis is contaminated by a large proportion of outliers.

2.4 Simulations

Simulations were conducted to investigate the finite-sample performance of the various approaches to assessing heterogeneity. Without loss of generality, the true overall mean effect size was fixed as $\mu = 0$. The number of studies in these simulated meta-analyses was set to $n = 10$ or 30, and the between-study variance was $\tau^2 = 0$ (homogeneity) or 1 (heterogeneity). Under the homogeneity setting, the within-study standard errors s_i were sampled from $U(0.5, 1)$; under the heterogeneity setting, we sampled s_i's from $U(s_{\min}, s_{\max})$, where $(s_{\min}, s_{\max}) = (0.5, 1)$, $(1, 2)$, or $(2, 5)$ to represent different proportions of total variation between studies that is due to heterogeneity. The observed effect sizes were drawn from $y_i \sim N(\mu_i, s_i^2)$, where μ_i's are study-specific underlying effect sizes. Regarding the μ_i, we considered the following two different scenarios to produce outliers.

(I) (Contamination) The μ_i's are normally distributed, $\mu_i \sim N(\mu, \tau^2)$; however, m out of the n studies were contaminated by a certain outlying discrepancy, as in Proposition 3. We set $m = 0$, 1, 2, and 3, and five outlier patterns were considered: the m studies were created as outliers by artificially adding C, (C, C), $(C, -C)$,

(C, C, C), or $(C, C, -C)$ to the original effect sizes for $m = 1$, 2, 2, 3, and 3 respectively. The discrepancy of outliers was set to $C = 3\sqrt{s^2_{\max} + \tau^2}$.

(II) (Heavy tail) The μ_i's are drawn from a heavy-tailed distribution. We considered a location-scale transformed t distribution with degrees of freedom $\mathrm{df} = 3$, 5, and 10; that is, $\mu_i = \mu + z_i\sqrt{(\mathrm{df} - 2)/\mathrm{df}}$, where $z_i \sim t_{\mathrm{df}}$. Note that the between-study variance $\tau^2 = \mathrm{Var}[\mu_i] = 1$ in this scenario, so the generated studies are heterogeneous. Also, as degrees of freedom increases, the distribution of μ_i's converges to the normal distribution and outliers are less likely.

Table 2.1 presents some results for $n = 30$, including statistical sizes (type I error rates) and powers of the statistics Q, Q_r, and Q_m for testing $H_0 : \tau^2 = 0$ vs. $H_A : \tau^2 > 0$, and the root mean squared errors (RMSEs) and coverage probabilities of 95% CIs of $\widehat{\tau}^2_{\mathrm{DL}}$, $\widehat{\tau}^2_r$, and $\widehat{\tau}^2_m$. Appendix A.3 contains complete simulation results. When the studies are homogeneous, each of the three test statistics controls type I error rate well if no outliers are present. Also, the RMSEs of the three estimators of τ^2 are close and their coverage probabilities are fairly high. However, when outliers appear, the type I error rate of Q inflates dramatically compared to Q_r and Q_m. The RMSE of $\widehat{\tau}^2_{\mathrm{DL}}$ becomes larger than those of $\widehat{\tau}^2_r$ and $\widehat{\tau}^2_m$; also, the coverage probability of $\widehat{\tau}^2_{\mathrm{DL}}$ is lower, especially when $m = 3$. As the number of outliers increases, the weighted-median-based $\widehat{\tau}^2_m$ has smaller RMSE and its 95% CI has higher coverage probability than the weighted-mean-based $\widehat{\tau}^2_r$.

For heterogeneous studies, the conventional Q statistic is more powerful than Q_r or Q_m, but the differences are not large; this is expected because Q sacrifices type I error in the presence of outliers. In spite of this disadvantage of Q_r and Q_m, the proposed estimators of τ^2 still perform better than the conventional $\widehat{\tau}^2_{\mathrm{DL}}$ in both Scenarios I and II.

Figure 2.2 compares the impact of a single outlier in Scenario I with $m = 1$ on the heterogeneity measures I^2, I^2_r, and I^2_m. As expected, these heterogeneity measures generally increase due to the outlying study, so their changes are mostly greater than 0. However, for both homogeneous and heterogeneous studies, the changes of I^2_r and I^2_m are generally smaller than the changes of I^2, indicating that the proposed measures are indeed less affected by outliers than the conventional I^2.

2.5 Two case studies

2.5.1 Homogeneous studies with outliers

Ismail et al. [74] reported a meta-analysis consisting of 29 studies to evaluate the effect of aerobic exercise (AEx) on visceral adipose tissue (VAT) content/volume in overweight and obese adults, compared to control treatment. Figure 2.3(a) shows the forest plot with the observed effect sizes and their within-study 95% CIs; studies 1, 3, 19, and 29 seem to be outlying. If these four studies are removed, the remaining studies are much more homogeneous. Figure 2.3(b) presents the standardized residuals using both the fixed-effect and random-effects approaches described in Section 2.1.2. Studies 1, 19, and 29 have standardized residuals (under the fixed-effect setting) greater than 3 in absolute magnitude; hence, they may be considered outliers. We conducted sensitivity analysis by removing the following studies: (i) 1; (ii) 19; (iii) 29; (iv) 1 and 19; (v) 1 and 29; (vi) 19 and 29; and (vii) 1, 19, and 29.

Table 2.2 presents the results for the original meta-analysis and for alternate meta-analyses removing certain outlying studies. For the original meta-analysis, $I_r^2 = 0.44$ and $I_m^2 = 0.45$, compared to $I^2 = 0.59$. Also, $\widehat{\tau}_r$ and $\widehat{\tau}_m$ are smaller than $\widehat{\tau}_{\mathrm{DL}}$. To test $H_0 : \tau^2 = 0$ vs. $H_A : \tau^2 > 0$, the P-value of the Q statistic is smaller than 0.001, and those of the Q_r and Q_m statistics are 0.013 and 0.006, respectively. When study 29 is removed, the Q statistic is still significant (P-value $= 0.008$), while the P-values of the Q_r and Q_m statistics are larger than the commonly used significance level $\alpha = 0.05$. After removing all three outlying studies, the P-values of the three test statistics are much larger than 0.05; also, $I_r^2 = I_m^2 = 0$ and $I^2 = 0.11$. Hence, the heterogeneity presented in the original meta-analysis is mainly caused by the few outliers. Note that I_r^2 and I_m^2 are still noticeably smaller than I^2 after removing the three identified outliers. This may be because some studies other than studies 1, 19, and 29 are potentially outlying. Figure 2.3(b) shows that the absolute values of the standardized residuals of studies 3 and 28 are fairly close to 3. Although some outliers may not be clearly detected, I_r^2 and I_m^2 automatically reduce their impact without removing them.

2.5.2 Heterogeneous studies with outliers

Haentjens et al. [75] investigated the magnitude and duration of excess mortality after hip fracture among older men by performing a meta-analysis consisting of 17 studies. Figure 2.3(c) shows the forest plot with the observed effect sizes (log hazard ratios) and their 95% within-study CIs. The forest plot indicates that the collected studies tend to be heterogeneous. Despite this, we used both the fixed-effect and random-effects diagnostic procedure in Section 2.1.2 to detect potential outliers. Figure 2.3(d) shows the study-specific standardized residuals, indicating that study 17 is apparently outlying. Although study 9's standardized residual is smaller than 2 in absolute magnitude when using the random-effects approach, its standardized residual under the fixed-effect setting is fairly large. To take all potential outliers into account, we conducted sensitivity analysis by removing the following studies: (i) 9; (ii) 17; and (iii) 9 and 17.

The results are in Table 2.2. For the original meta-analysis, the P-values of all three test statistics are smaller than 0.001, rejecting the null hypothesis of homogeneity. Also, $I^2 = 0.74$, $I_r^2 = 0.66$ and $I_m^2 = 0.63$, indicating substantial heterogeneity. If study 9 is removed, the results seem to change little, implying that this study is not influential. If study 17 is removed, the P-values of the test statistics change noticeably; also, each of I^2, I_r^2, and I_m^2 is reduced by more than 0.10. The three heterogeneity measures are still fairly high (larger than or close to 0.5); therefore, meta-analysts may keep paying attention to the heterogeneity of the remaining studies.

2.6 Discussion

This paper proposed several alternative measures of heterogeneity in meta-analysis. Large-sample properties and finite-sample studies showed that the new measures are robust to outliers compared with conventional measures. Since outliers frequently appear in meta-analysis and may not simply be removed without sound evidence, the proposed robust measures can provide useful information describing heterogeneity. The robustness of the new approaches mainly arises from using the absolute deviations in the Q_r and Q_m statistics; Q_r summarizes the deviations using the weighted average, and Q_m summarizes the deviations using the weighted median. Note that the number of studies is assumed to be large in deriving $\widehat{\tau}_m^2$, H_m, and I_m^2. However, many meta-analyses may only collect a few studies [76]; these three measures need to be used with caution for small meta-analyses.

When study-level covariates are collected in meta-analysis, meta-regression is widely applied to investigate whether study characteristics explain heterogeneity [77]. To improve robustness to outliers, instead of performing least squares regression, researchers may consider least absolute deviations regression [68], which is related to the heterogeneity measures proposed in this chapter.

Heterogeneity measures are customarily used to select a fixed-effect or random-effects model, but both models have limitations in certain situations. Some researchers believe that heterogeneity is to be expected in any meta-analysis because the collected studies were performed by different teams in different places using different methods [78]. Also, the fixed-effect model produces confidence intervals with poor coverage probability when the collected studies have different true effect sizes [79], so some researchers recommend routinely using the random-effects model to yield conservative results [80]. However, the random-effects model is not always better than the fixed-effect model, especially in the presence of publication bias [81–83]. Besides robustly assessing heterogeneity, alternative approaches to robustly estimating overall effects size in the presence of outliers remain to be studied.

The `R` code for the proposed methods are organized in the package ‘altmeta’ and available at `http://cran.r-project.org/package=altmeta`.

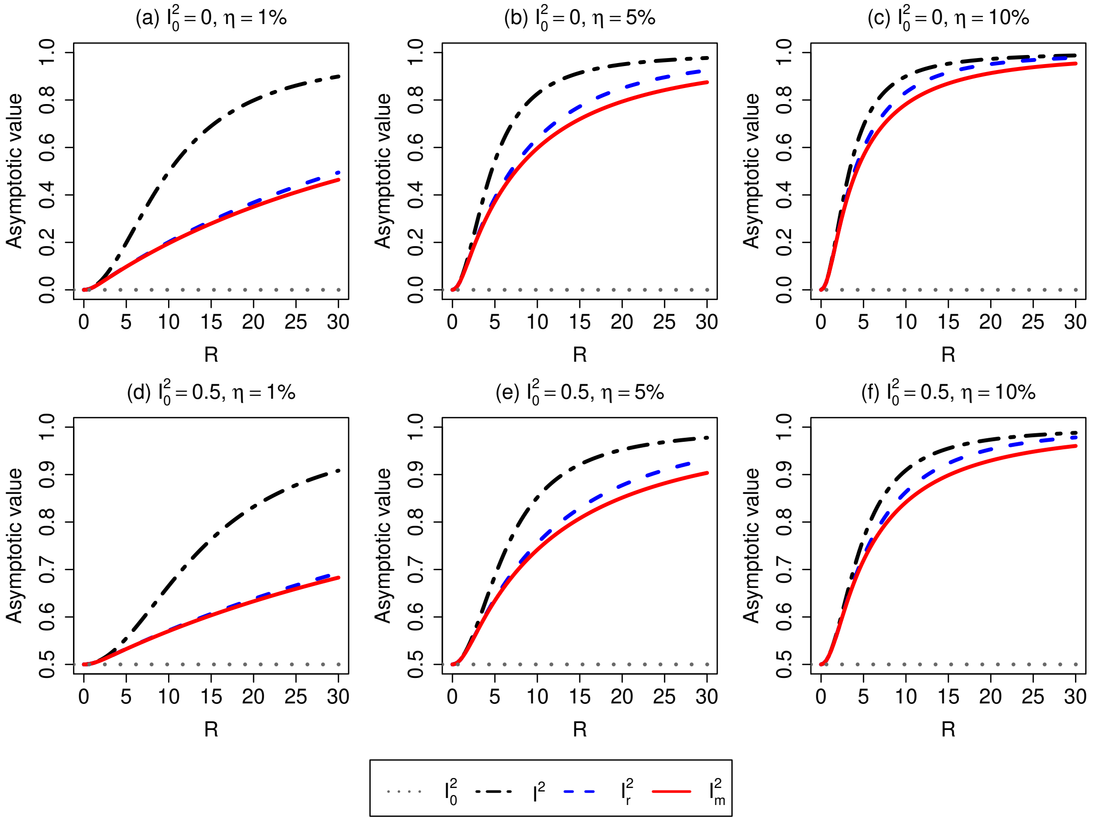


Figure 2.1: The asymptotic values of I^2, I_r^2, and I_m^2 as $n \to \infty$. The horizontal axis represents the ratio (R) of discrepancy of outliers (C) compared to within-study standard deviation (σ), that is, $R = C/\sigma$. The true proportion of total variation between studies that is due to heterogeneity I_0^2 is 0 (homogeneity, top row) or 0.5 (heterogeneity, bottom row). The proportion of outlying studies η varies from 1% (left panels) to 10% (right panels).

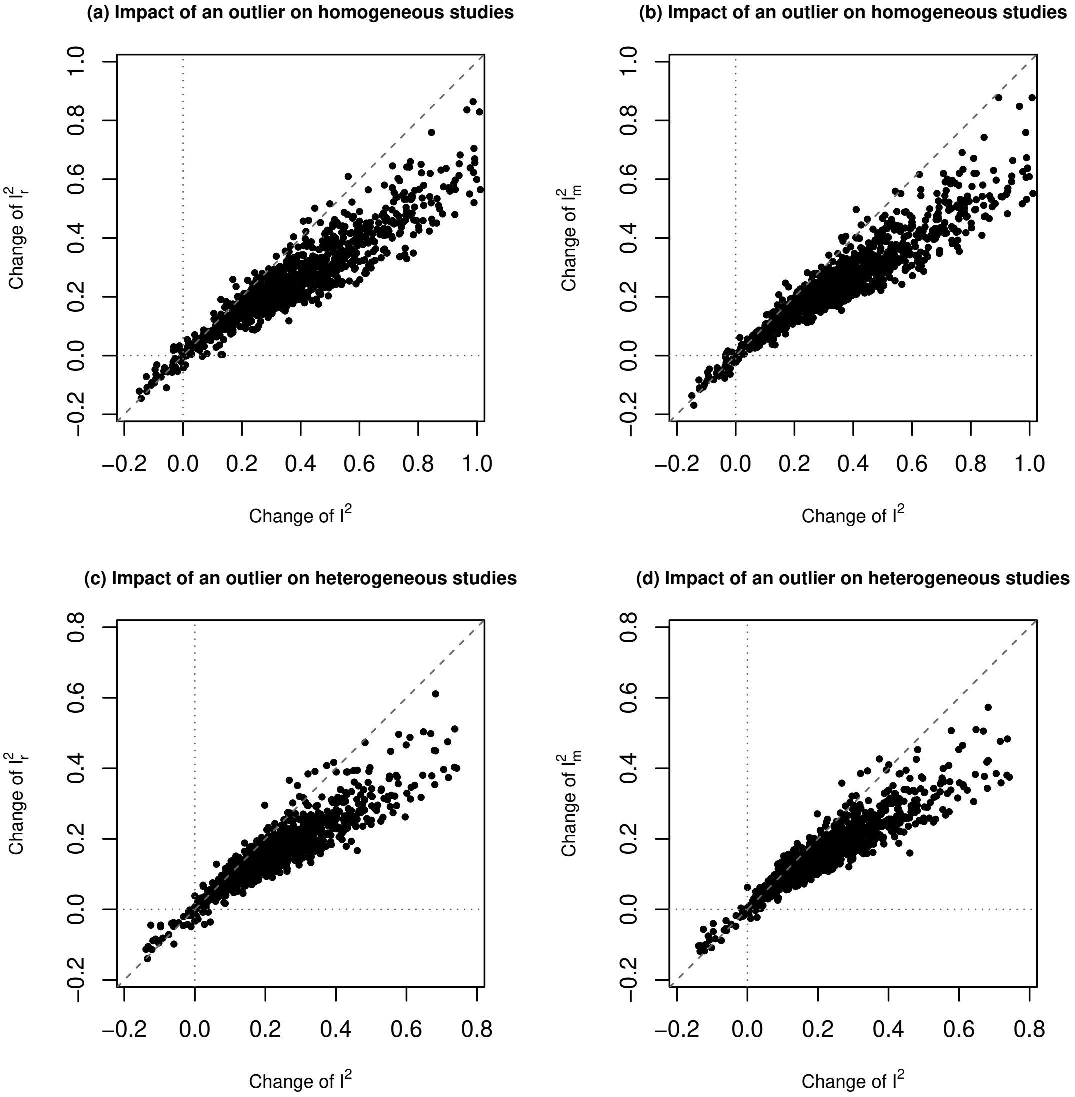
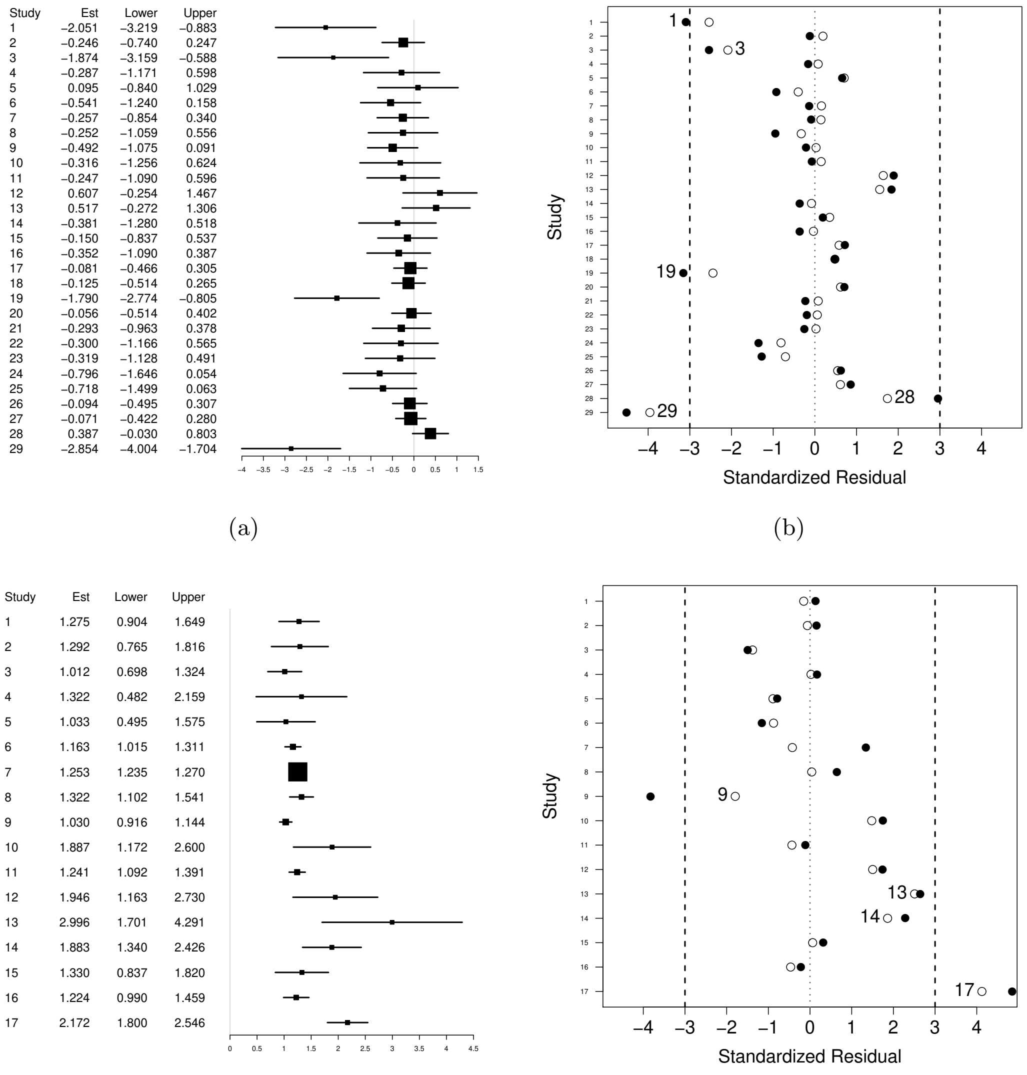


Figure 2.2: Scatter plots of the changes of I_r^2 and I_m^2 due to an outlier against the changes of I^2. For the upper panels, $\tau^2 = 0$ (homogeneous studies) and $s_i \sim U(0.5, 1)$; for the lower panels, $\tau^2 = 1$ (heterogeneous studies) and $s_i \sim U(1, 2)$. The left panels compare I_r^2 with I^2; the right panels compare I_m^2 with I^2.



Study	Est	Lower	Upper
1	−2.051	−3.219	−0.883
2	−0.246	−0.740	0.247
3	−1.874	−3.159	−0.588
4	−0.287	−1.171	0.598
5	0.095	−0.840	1.029
6	−0.541	−1.240	0.158
7	−0.257	−0.854	0.340
8	−0.252	−1.059	0.556
9	−0.492	−1.075	0.091
10	−0.316	−1.256	0.624
11	−0.247	−1.090	0.596
12	0.607	−0.254	1.467
13	0.517	−0.272	1.306
14	−0.381	−1.280	0.518
15	−0.150	−0.837	0.537
16	−0.352	−1.090	0.387
17	−0.081	−0.466	0.305
18	−0.125	−0.514	0.265
19	−1.790	−2.774	−0.805
20	−0.056	−0.514	0.402
21	−0.293	−0.963	0.378
22	−0.300	−1.166	0.565
23	−0.319	−1.128	0.491
24	−0.796	−1.646	0.054
25	−0.718	−1.499	0.063
26	−0.094	−0.495	0.307
27	−0.071	−0.422	0.280
28	0.387	−0.030	0.803
29	−2.854	−4.004	−1.704

(a)

(b)

Study	Est	Lower	Upper
1	1.275	0.904	1.649
2	1.292	0.765	1.816
3	1.012	0.698	1.324
4	1.322	0.482	2.159
5	1.033	0.495	1.575
6	1.163	1.015	1.311
7	1.253	1.235	1.270
8	1.322	1.102	1.541
9	1.030	0.916	1.144
10	1.887	1.172	2.600
11	1.241	1.092	1.391
12	1.946	1.163	2.730
13	2.996	1.701	4.291
14	1.883	1.340	2.426
15	1.330	0.837	1.820
16	1.224	0.990	1.459
17	2.172	1.800	2.546

(c)

(d)

Figure 2.3: Forest plots and standardized residual plots of two actual meta-analyses. The upper panels show the meta-analysis conducted by Ismail et al.; the lower panels show that conducted by Haentjens et al. In (a) and (c), the columns 'Lower' and 'Upper' are the lower and upper bounds of 95% CIs of the effect sizes within each study. In (b) and (d), the filled dots represent standardized residuals obtained under the fixed-effect setting; the unfilled dots represent those obtained under the random-effects setting.

Table 2.1: Type I error rates and powers of three heterogeneity tests for the simulated meta-analyses containing 30 studies.

Outlier pattern	Size/power[†]			RMSE			CP (%)		
	$Q^{\ddagger}$	Q_r	Q_m	$\widehat{\tau}^2_{\mathrm{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$	$\widehat{\tau}^2_{\mathrm{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$
Scenario I (contamination) with $\tau^2 = 0$ (homogeneity) and $s_i \sim U(0.5, 1)$:									
No outliers	0.05 (0.06)	0.05	0.05	0.10	0.12	0.10	98	99	99
C	0.55 (0.55)	0.27	0.25	0.37	0.24	0.20	97	97	98
(C, C)	0.89 (0.89)	0.66	0.60	0.63	0.42	0.35	88	90	94
$(C, -C)$	0.92 (0.92)	0.61	0.61	0.68	0.40	0.36	89	90	94
(C, C, C)	0.98 (0.98)	0.90	0.87	0.88	0.64	0.53	65	74	83
$(C, C, -C)$	0.99 (0.98)	0.89	0.88	0.99	0.61	0.55	64	73	83
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(0.5, 1)$:									
No outliers	0.98 (0.99)	0.98	0.98	0.40	0.43	0.41	88	93	91
C	1.00 (1.00)	1.00	1.00	0.84	0.63	0.55	97	97	98
(C, C)	1.00 (1.00)	1.00	1.00	1.37	1.00	0.85	93	94	96
$(C, -C)$	1.00 (1.00)	1.00	1.00	1.45	0.97	0.85	93	94	96
(C, C, C)	1.00 (1.00)	1.00	1.00	1.86	1.44	1.22	76	83	90
$(C, C, -C)$	1.00 (1.00)	1.00	1.00	2.05	1.40	1.25	77	84	91
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(1, 2)$:									
No outliers	0.48 (0.49)	0.42	0.43	0.74	0.81	0.75	89	93	91
C	0.89 (0.89)	0.78	0.77	1.97	1.36	1.17	98	97	98
(C, C)	0.99 (0.99)	0.94	0.94	3.33	2.29	1.93	91	92	96
$(C, -C)$	0.99 (0.99)	0.94	0.94	3.50	2.17	1.93	91	92	96
(C, C, C)	1.00 (1.00)	0.99	0.99	4.60	3.41	2.85	70	80	88
$(C, C, -C)$	1.00 (1.00)	0.99	0.99	5.03	3.24	2.90	71	81	88
Scenario II (heavy tail) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(0.5, 1)$:									
df $= 3$	0.92 (0.92)	0.89	0.88	1.45	0.59	0.56	72	79	73
df $= 5$	0.98 (0.98)	0.95	0.95	0.55	0.45	0.45	84	90	86
df $= 10$	0.98 (0.98)	0.97	0.97	0.43	0.43	0.42	88	93	90
Scenario II (heavy tail) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(1, 2)$:									
df $= 3$	0.41 (0.40)	0.35	0.35	1.53	0.88	0.82	83	90	87
df $= 5$	0.46 (0.46)	0.40	0.40	0.82	0.82	0.77	88	93	90
df $= 10$	0.48 (0.49)	0.42	0.42	0.76	0.82	0.77	88	94	90

RMSE: root mean squared error; CP: coverage probability of 95% confidence interval.

[†] Size (type I error rate) for homogeneous studies ($\tau^2 = 0$) and power for heterogeneous studies ($\tau^2 > 0$) at the significance level $\alpha = 0.05$.

[‡] The sizes/powers outside the parentheses are produced by the resampling method; those inside the parentheses are obtained using Q's theoretical distribution under the null hypothesis.

Table 2.2: Results of assessing heterogeneity for two actual meta-analyses.

Removed studies	P-value of testing $H_0 : \tau^2 = 0$			Estimated τ (95% CI)			Heterogeneity measure (95% CI)		
	$Q^\dagger$	Q_r	Q_m	$\widehat{\tau}_{\mathrm{DL}}$	$\widehat{\tau}_r$	$\widehat{\tau}_m$	I^2	I_r^2	I_m^2
Meta-analysis in Ismail et al. [74]:									
None (Original)	< 0.001 (< 0.001)	0.013	0.006	0.39 (0, 0.62)	0.29 (0, 0.58)	0.30 (0, 0.56)	0.59 (0, 0.76)	0.44 (0, 0.73)	0.45 (0, 0.72)
1	< 0.001 (< 0.001)	0.047	0.030	0.35 (0, 0.58)	0.24 (0, 0.52)	0.24 (0, 0.51)	0.55 (0, 0.75)	0.36 (0, 0.69)	0.36 (0, 0.69)
19	< 0.001 (< 0.001)	0.048	0.031	0.34 (0, 0.58)	0.24 (0, 0.52)	0.24 (0, 0.51)	0.54 (0, 0.75)	0.36 (0, 0.69)	0.36 (0, 0.68)
29	0.008 (0.007)	0.100	0.070	0.28 (0, 0.46)	0.21 (0, 0.44)	0.21 (0, 0.43)	0.44 (0, 0.66)	0.29 (0, 0.63)	0.30 (0, 0.62)
1 and 19	0.003 (0.004)	0.154	0.121	0.29 (0, 0.54)	0.18 (0, 0.45)	0.18 (0, 0.44)	0.47 (0, 0.73)	0.25 (0, 0.64)	0.24 (0, 0.63)
1 and 29	0.052 (0.052)	0.272	0.223	0.22 (0, 0.40)	0.14 (0, 0.37)	0.13 (0, 0.36)	0.33 (0, 0.60)	0.16 (0, 0.56)	0.15 (0, 0.55)
19 and 29	0.057 (0.057)	0.278	0.232	0.21 (0, 0.40)	0.13 (0, 0.38)	0.13 (0, 0.37)	0.32 (0, 0.60)	0.15 (0, 0.56)	0.14 (0, 0.55)
1, 19 and 29	0.302 (0.298)	0.547	0.504	0.11 (0, 0.30)	0 (0, 0.29)	0 (0, 0.27)	0.11 (0, 0.47)	0 (0, 0.46)	0 (0, 0.42)
Meta-analysis in Haentjens et al. [75]:									
None (Original)	< 0.001 (< 0.001)	< 0.001	< 0.001	0.16 (0.02, 0.34)	0.15 (0, 0.37)	0.08 (0, 0.36)	0.74 (0.15, 0.86)	0.66 (0, 0.85)	0.63 (0, 0.85)
9	< 0.001 (< 0.001)	0.006	0.006	0.16 (0, 0.37)	0.13 (0, 0.42)	0.06 (0, 0.37)	0.68 (0, 0.84)	0.56 (0, 0.83)	0.52 (0, 0.81)
17	0.001 (0.001)	0.013	0.015	0.11 (0, 0.23)	0.11 (0, 0.27)	0.05 (0, 0.27)	0.60 (0, 0.76)	0.52 (0, 0.77)	0.47 (0, 0.76)
9 and 17	0.062 (0.059)	0.156	0.144	0.09 (0, 0.24)	0.07 (0, 0.27)	0.02 (0, 0.25)	0.39 (0, 0.65)	0.28 (0, 0.67)	0.23 (0, 0.65)

† The P-values outside the parentheses are produced by the resampling method; the P-values inside the parentheses are calculated using Q's theoretical distribution under the null hypothesis.

Chapter 3

Performance of Publication Bias Tests in the Cochrane Library

This chapter applies seven commonly-used publication bias tests to a large collection of published meta-analyses in the Cochrane Library, which is the leading resource for systematic reviews in health care. We investigate the proportion of meta-analyses that have statistically significant publication bias detected by each test. Its association with the size of the meta-analysis is also empirically assessed. In addition, we evaluate the agreement among various test results. These findings may guide researchers in properly assessing publication bias and interpreting test results in future systematic reviews.

3.1 Methods

We searched complete issues in the Cochrane Library that were available in January 2016; a total of 5677 systematic reviews were collected, containing more than 180,000 meta-analyses. We only considered meta-analyses with continuous or binary outcomes. For binary outcomes, the treatment effects were measured by the log odds ratio. When a study contained a zero data cell in one arm only, we added a continuity correction of 0.5 to all studies' data cells in the corresponding meta-analysis so that log odds ratios and their variances can be estimated [3, 84]. Studies with zero data cells in both treatment and control arms were removed from the meta-analyses because information provided by such studies is limited [3, 85, 86]. For continuous outcomes, some studies did not report the treatment effects' standard errors; they were also removed from the meta-analyses. After removing the ineligible studies, we focused on meta-analyses containing at least five studies. We finally obtained a total of 20,603 meta-analyses; among them, 6080 and 14,523 meta-analyses have continuous and binary outcomes, respectively.

For meta-analyses with continuous outcomes, we applied Begg's rank test, the trim and fill method, and Egger's and Tang's regression tests to assess publication bias; these approaches have been proposed for all types of outcomes [24–26, 29]. For meta-analyses with binary outcomes, we also considered Macaskill's, Deeks', and Peters' regression tests, which were originally designed for log odds ratios [31–33]. As suggested by many authors, the statistical significance level was set to 0.1 because the statistical power for testing publication bias is generally low [24, 25, 32]. Moreover, Cohen's κ, a coefficient upper bounded by 1, was used to measure pairwise agreement among publication bias tests [87]. The agreement was considered strong if κ was larger than 0.6, and weak if κ was smaller than 0.4; the agreement was moderate when κ is between 0.4 and 0.6 [88].

Within a systematic review, multiple meta-analyses may be performed for different outcomes, but using information from some common populations; therefore, these meta-analyses can be correlated [89]. To reduce the impact of such correlations, we also conducted the analysis using a restricted dataset. Specifically, the meta-analysis with the largest number of studies was chosen from each systematic review. If a systematic review contained at least two meta-analyses with the same largest number of studies, the one with the largest total sample size was selected; if the total sample sizes are still equal, one meta-analysis was randomly chosen from those having the largest number of studies and total sample size. Again, we focused on meta-analyses containing at least five studies. Using these criteria, 499 and 1380 meta-analyses with continuous and binary outcomes respectively were extracted from the entire set of 5677 systematic reviews.

3.2 Results

Figures 3.1 and 3.2 show the P-values produced by the various publication bias tests for the Cochrane meta-analyses with continuous and binary outcomes, respectively. The horizontal axis presents each meta-analysis sorted by its size (i.e., the number of studies); the meta-analyses with the same size are sorted by their IDs in the Cochrane Library. The vertical axis shows the P-values transformed by negative base 10 logarithm, and three statistical significance levels, 0.01, 0.05, and 0.1, are displayed. Both figures illustrate that the area representing small meta-analyses is much wider than that representing large meta-analyses, and most Cochrane meta-analyses contain less than 10 studies. Specifically, among the entire 20,603 meta-analyses with continuous or binary outcomes, 5338 meta-analyses contain 5 studies, while only 132 meta-analyses contain 20 studies. The median number of studies is 7, and the lower and upper quartiles are 5 and 10 respectively.

Overall, Table 3.1 shows that Begg's rank test and the trim and fill method detect statistically significant publication bias in far fewer meta-analyses than the regression tests. In particular, for small meta-analyses, Figures 3.1 and 3.2 indicate that the P-values produced by Begg's rank test and the trim and fill method are generally larger than the regression tests. For example, among the meta-analyses containing 5 studies, most P-values produced by Begg's rank test and all P-values produced by the trim and fill method are greater than 0.05, while the regression tests imply extreme publication bias with P-value < 0.01 in some meta-analyses. In addition, Begg's rank test and the trim and fill method are more likely to detect publication bias in large meta-analyses than in small ones. Furthermore, note that all P-values of the trim and fill method are discontinuous and massed at several specific values, because this method uses the negative binomial distribution, which is discrete, to calculate P-value [26]. Many P-values of Begg's rank test are also massed at several specific values. This is because the rank test calculates an exact P-value, taking certain discontinuous values, when the number of studies is small and the treatment effects have no ties; otherwise, the P-value is calculated using the normal approximation of the rank statistic's distribution.

Compared with Begg's rank test and the trim and fill method, the significance of publication bias assessed by the regression tests seems to be less dependent on the size of the meta-analysis. Table 3.1 shows that Egger's test detects statistically significant publication bias in 13.9% of meta-analyses with continuous outcomes and 16.9% of those with binary outcomes; these proportions are higher than the other regression tests. The numbers of meta-analyses with statistically significant publication bias detected by

Tang's, Deeks', and Peters' tests are similar for binary outcomes. Moreover, the P-value plots of Tang's and Deeks' tests in Figure 3.2 are fairly similar. However, the plots of the other regression tests are noticeably different: one test may not detect statistically significant publication bias for a meta-analysis, while another test could lead to an extremely small P-value for the same meta-analysis.

Table 3.2 quantifies the agreement among the tests using Cohen's κ coefficient. The upper table uses all extracted Cochrane meta-analyses, and the lower one is based on the restricted dataset, which consists of the largest meta-analysis from each systematic review. Most results in the upper and lower tables are similar. We may focus on the lower table, in which the meta-analyses are from different systematic reviews and may be deemed independent. Begg's rank test and the trim and fill method have a rather weak agreement ($\kappa \leq 0.40$), and their agreement with the regression tests is also weak. Egger's test has moderate agreement with Tang's, Deeks', and Peters' regression tests. Most Cohen's κ coefficients between Tang's, Macaskill's, Deeks', and Peters' tests are close to 0.60, which may imply moderately strong agreement. Note that the Cohen's κ coefficient between Tang's and Deeks' tests is close to 1, implying a nearly perfect agreement; this confirms our observation in Figure 3.2.

Categorized by the number of studies, Figure 3.3 describes the proportions of meta-analyses having statistically significant publication bias based on the various tests, and their Wald-type 95% confidence intervals. On the one hand, similarly to the patterns of the P-value plots in Figures 3.1 and 3.2, the proportion tends to be greater for larger meta-analyses, especially for binary outcomes. On the other hand, the proportions of the Cochrane meta-analyses having statistically significant publication bias are between approximately 10% and 30% for most sizes of meta-analyses. In addition, publication bias is detected by at least one test in more than 20% of meta-analyses with continuous outcomes and in more than 30% of meta-analyses with binary outcomes.

Figures A.2–A.4 in Appendix A.4 show the P-value plots and the plot of proportions of having publication bias based on the restricted dataset. The trends in these plots are similar to those in Figures 3.1–3.3, though the 95% confidence intervals in Figure A.4 are wider than those in Figure 3.3 because the restricted dataset contains far fewer meta-analyses.

3.3 Discussion

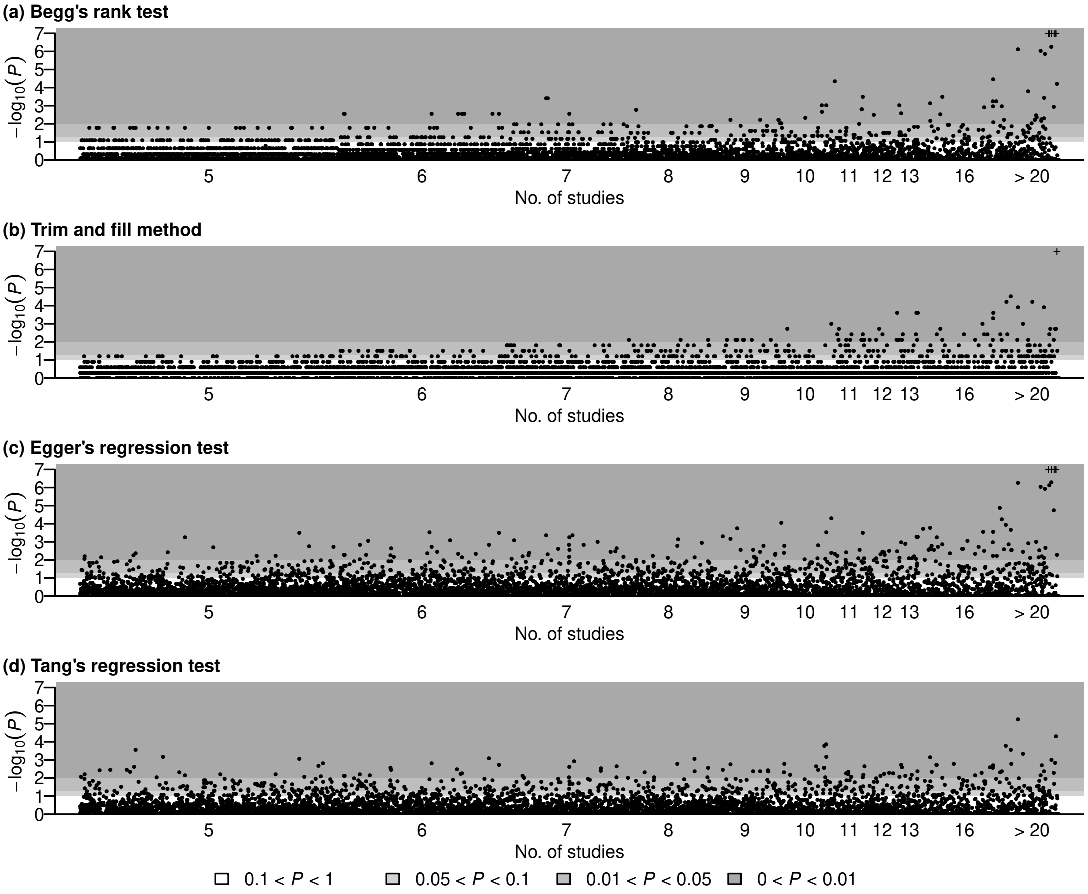
Using a large collection of meta-analyses, this chapter illustrated that publication bias frequently appears in the Cochrane systematic reviews, so it should be routinely assessed. Egger's regression test detects statistically significant publication bias in more meta-analyses than the others. However, this study has several limitations. For example, the Cochrane Library only contains meta-analyses in health care, so the results may not be generalizable to other research fields. Also, since we never know whether a Cochrane meta-analysis truly has publication bias, the results in Table 3.1 and Figures 3.1–3.3 may not directly imply statistical powers of the tests.

Since the agreement among most publication bias tests is weak or moderate, researchers need to carefully interpret the test results. Instead of reporting the result from a single test, researchers are encouraged to use a variety of methods: different tests make different assumptions about the association between the treatment effects and precision measures (e.g., treatment effects' standard error or sample size), so the tests that yield fairly small P-values may reveal some patterns for further investigation.

Tang's and Deeks' regression tests are shown to have almost identical performance. Tang's method is motivated by examining the asymmetry of the sample-size-based funnel plot for all types of outcomes, and the independent variable in the regression is the total sample size within each study [29]; Deeks' method was originally developed for meta-analysis of diagnostic tests, and the regression independent variable is the 'effective sample size' (Table 3.1) [31]. If the allocation ratio for the treatment and control groups is close to 1:1, which is common in randomized controlled trials, then the 'effective sample size' is close to the total sample size. Therefore, it is not surprising to obtain similar results using Tang's and Deeks' tests.

All seven tests considered in this chapter are motivated by the funnel plot; however, the funnel plot's asymmetry needs to be interpreted from various perspectives. For example, since small studies may be biased due to poor quality in design and they are likely targeted at high-risk groups that can produce positive treatment effects, some authors often view the funnel plot as an approach to checking for 'small study effects' in general, rather than publication bias in particular [30, 90, 91]. In addition, the P-value plots in Figures 3.1 and 3.2 indicate that some publication bias tests tend to detect more statistically significant publication bias in larger meta-analyses. As the number of studies increases, a meta-analysis likely collects more heterogeneous or outlying studies, which can cause a funnel plot's asymmetry for reasons other than publication bias. Outliers may be present in meta-analysis due to several reasons. For example, some study results could be outlying because of errors in the process of recording, analyzing, or reporting data. Also, if the quality of a systematic review is poor, the populations in certain studies may not meet strict inclusion and exclusion criteria, so such studies may be outlying compared with the other collected studies. Outliers may lead to a heavy tail on one side of the treatment effect's distribution, so the funnel plot may look asymmetric.

Between-study heterogeneity also seriously threatens proper interpretation of the funnel plot's asymmetry. It arises because the collected studies differ in their patient selection, baseline disease severity, study location, etc. [78, 92]. The random-effects meta-analysis is usually applied to deal with heterogeneity; a normal distribution is conventionally specified to model study-specific underlying treatment effects [6, 93]. This model is appropriate if heterogeneity permeates the entire collection of studies; however, it is also possible that heterogeneity is mostly limited to several subgroups of studies, while the studies within each subgroup share a common overall treatment effect. In the presence of multiple subgroups, even if the funnel plot within each subgroup is fairly symmetric, the funnel plot based on the entire collection of studies can be asymmetric; such asymmetry is induced by heterogeneity, but not publication bias [34, 94]. Performing separate analysis within each subgroup is more appropriate for such data than pooling the results of all studies. As heterogeneity is common in meta-analysis [78, 95], researchers need to carefully assess heterogeneity along with checking for publication bias. For example, Ioannidis and Trikalinos [89] advised that it may not be appropriate to use the publication bias tests if the I^2 statistic [13, 95] is greater than 50% or the Q statistic [9, 10] is significant with P-value < 0.1. Although these criteria may not be rigorous for determining whether the publication bias tests are appropriate, a fairly large heterogeneity measure alerts researchers to interpret the funnel plot's asymmetry with great caution.



(a) Begg's rank test

(b) Trim and fill method

(c) Egger's regression test

(d) Tang's regression test

Figure 3.1: The P-values produced by the various publication bias tests for the 6080 Cochrane meta-analyses with continuous outcomes. Plus signs indicate P-values $< 10^{-7}$.

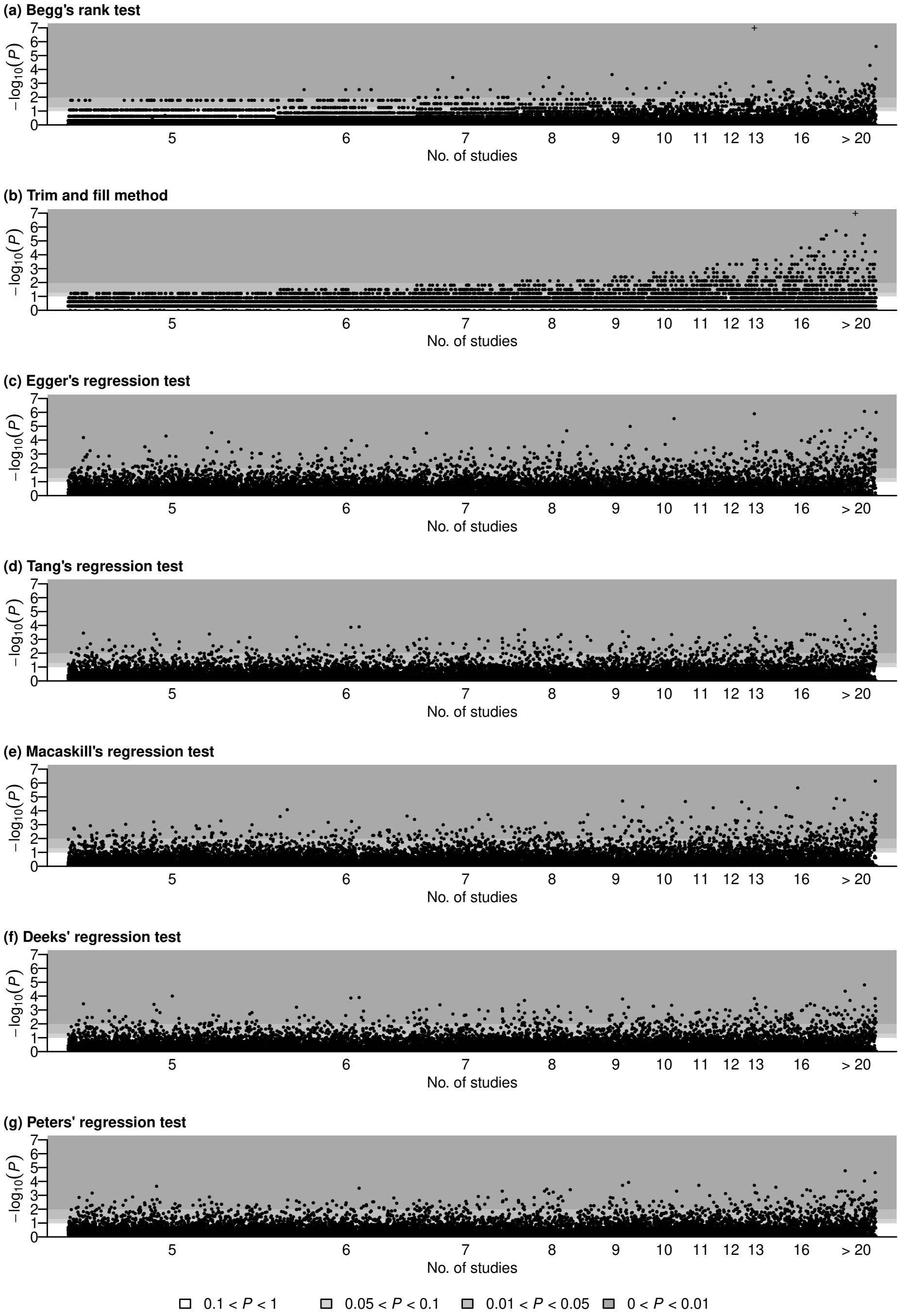


Figure 3.2: The P-values produced by the various publication bias tests for the 14,523 Cochrane meta-analyses with binary outcomes. Plus signs indicate P-values $< 10^{-7}$.

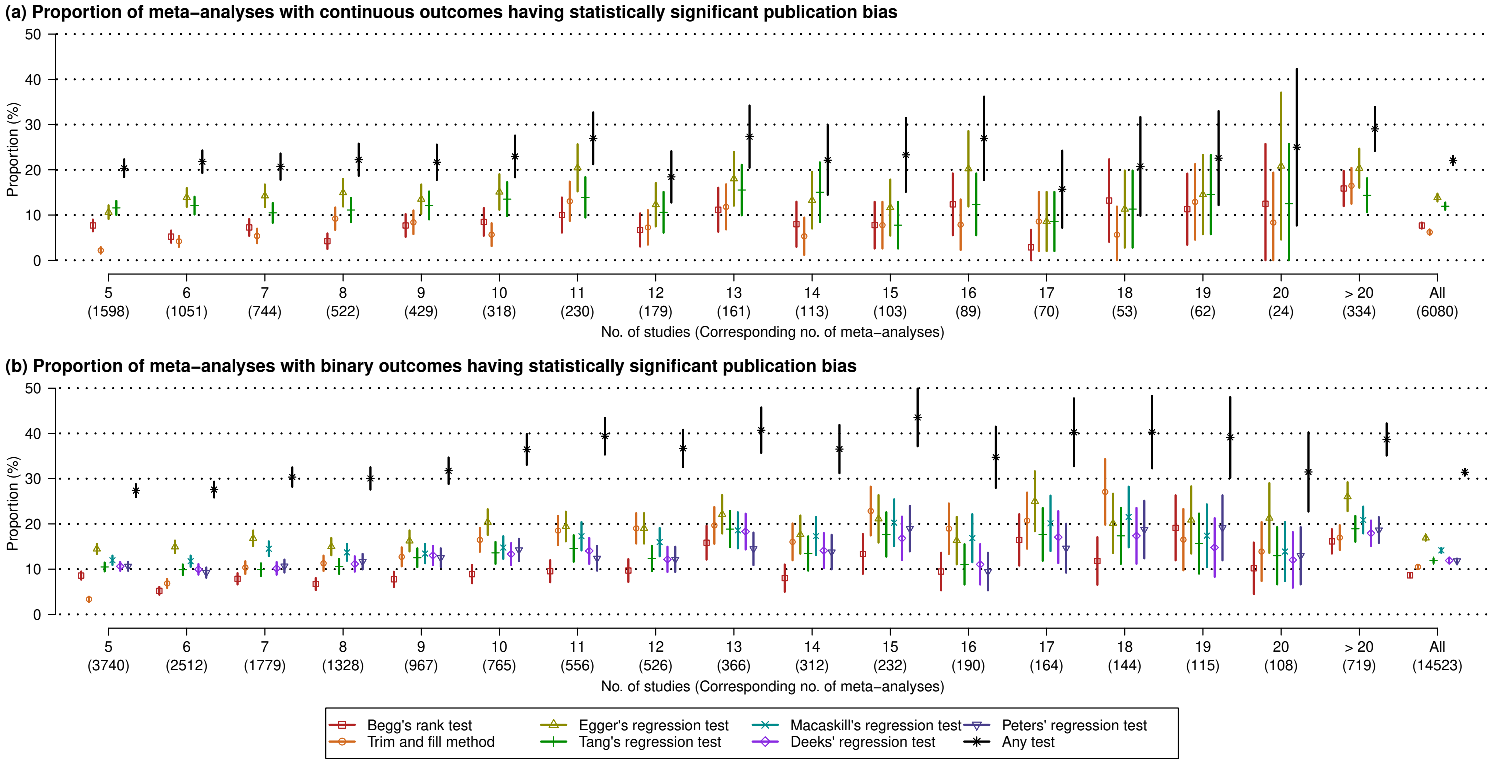


Figure 3.3: Proportions of the Cochrane meta-analyses having statistically significant publication bias (P-value < 0.1) based on the various tests and their 95% confidence intervals. ‘Any test’ implies the proportion of having statistically significant publication bias detected by at least one test. The label ‘All’ on the horizontal axis represents all extracted meta-analyses with continuous/binary outcomes.

Table 3.1: Brief descriptions for the various publication bias tests and summary of test results for the Cochrane meta-analyses.

			No. of meta-analyses with P-value < 0.1 (Proportion)			
Test	Designed for	Description	Based on all Cochrane meta-analyses		Based on the restricted dataset[a]	
			Continuous[b]	Binary[c]	Continuous[d]	Binary[e]
Begg's rank test	All outcomes	Use the rank correlation test to assess the association between standardized effect size and its standard error.	467 (7.7%)	1253 (8.6%)	43 (8.6%)	133 (9.6%)
Trim and fill method	All outcomes	Estimate the number of suppressed studies, and calculate P-value using its negative binomial distribution in the absence of publication bias.	378 (6.2%)	1523 (10.5%)	33 (6.6%)	177 (12.8%)
Egger's regression test	All outcomes	Weighted linear regression of y on s, with weights $1/s^2$.	843 (13.9%)	2455 (16.9%)	74 (14.8%)	264 (19.1%)
Tang's regression test	All outcomes	Weighted linear regression of y on $1/\sqrt{N}$, with weights N.	727 (12.0%)	1723 (11.9%)	67 (13.4%)	180 (13.0%)
Macaskill's regression test	Binary outcomes	Weighted linear regression of y on N, with weights $N_s \times N_f/N$.	N/A	2055 (14.1%)	N/A	200 (14.5%)
Deeks' regression test	Binary outcomes	Weighted linear regression of y on $1/\sqrt{N_e}$, with weights N_e.	N/A	1729 (11.9%)	N/A	182 (13.2%)
Peters' regression test	Binary outcomes	Weighted linear regression of y on $1/N$, with weights $N_s \times N_f/N$.	N/A	1717 (11.8%)	N/A	189 (13.7%)

Notation: y, effect size; s^2, within-study variance; N, total no. of patients; N_s and N_f, no. of patients with and without events for binary outcomes respectively; N_e, effective sample size, defined as $4N_0 \times N_1/N$, where N_0 and N_1 are sample sizes the control and treatment groups respectively; N/A, not applicable.

[a] The restricted dataset consists of the meta-analyses with the largest numbers of studies in the corresponding Cochrane systematic reviews.

[b] Among 6080 meta-analyses with continuous outcomes.

[c] Among 14,523 meta-analyses with binary outcomes.

[d] Among 499 meta-analyses with continuous outcomes in the restricted dataset.

[e] Among 1380 meta-analyses with binary outcomes in the restricted dataset.

Table 3.2: Cohen's κ coefficients for the agreement among the publication bias tests. Within each sub-table, the results in the upper and lower triangular are based on the Cochrane meta-analyses with continuous and binary outcomes, respectively.

Based on all Cochrane meta-analyses with at least five studies:						
Begg	0.23	0.48	0.33	N/A	N/A	N/A
0.25	**T & F**	0.35	0.20	N/A	N/A	N/A
0.46	0.43	**Egger**	0.51	N/A	N/A	N/A
0.26	0.30	0.43	**Tang**	N/A	N/A	N/A
0.14	0.24	0.35	0.55	**Macaskill**	N/A	N/A
0.27	0.30	0.43	**0.93**	0.53	**Deeks**	N/A
0.27	0.25	0.40	**0.67**	0.47	**0.66**	**Peters**
Based on the meta-analyses with the largest numbers of studies in the corresponding Cochrane systematic reviews:						
Begg	0.40	0.51	0.33	N/A	N/A	N/A
0.30	**T & F**	0.41	0.25	N/A	N/A	N/A
0.46	0.45	**Egger**	0.48	N/A	N/A	N/A
0.29	0.31	0.45	**Tang**	N/A	N/A	N/A
0.17	0.24	0.38	**0.60**	**Macaskill**	N/A	N/A
0.28	0.31	0.45	**0.95**	0.59	**Deeks**	N/A
0.27	0.28	0.46	**0.69**	0.55	**0.70**	**Peters**

Begg, the rank test; Egger, Tang, Macaskill, Deeks, and Peters, the regression tests; T & F, the trim and fill method; N/A, not applicable. Cohen's κ coefficients ≥ 0.60 are in bold.

Chapter 4

Quantifying Publication Bias in Meta-Analysis

This chapter introduces an alternative measure to quantify publication bias, the skewness of the standardized deviates. The new measure not only has an intuitive interpretation as the asymmetry of the collected study results but also can serve as a test statistic. The large sample properties of the new measure are studied. We also evaluate its performance using simulations and three actual meta-analyses published in the *Cochrane Database of Systematic Reviews.*

4.1 Notation and the regression test

Suppose a meta-analysis collects n studies; each study reports an effect size y_i (e.g., log odds ratio for binary outcomes) and its within-study variance s_i^2, due to sampling error ($i = 1, \ldots, n$). If the collected studies are deemed homogeneous, sharing a common underlying true effect size μ, then the fixed-effect model is customarily used, specified by $y_i \sim N(\mu, s_i^2)$. The studies are heterogeneous if they have different underlying effect sizes μ_i; the corresponding random-effects model assumes $y_i \sim N(\mu_i, s_i^2)$ and $\mu_i \sim N(\mu, \tau^2)$, where τ^2 is the between-study variance and μ is interpreted as the overall mean effect size [6]. The random-effects model reduces to the fixed-effect model by setting $\tau^2 = 0$.

To detect publication bias, Egger et al. [25] proposed a regression test, regressing the standardized effect sizes (y_i/s_i) on the corresponding precisions ($1/s_i$); that is,

$$y_i/s_i = \alpha + \mu \cdot 1/s_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Egger's regression test transforms the original null hypothesis, H_0: no publication bias, to testing H_0': the regression intercept is zero. Alternatively, in the presence of noticeable heterogeneity between studies, we may slightly modify Egger's test by using the marginal standard deviations to produce the regression predictors and responses under the random-effects model. Note that the random-effects model can be written marginally as $y_i = \mu + \delta_i + \xi_i$, where $\delta_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ is the random effect and $\xi_i \sim N(0, s_i^2)$ is the sampling error in study i. Dividing by the marginal standard deviation $(s_i^2 + \tau^2)^{1/2}$, we have the following modified regression test:

$$y_i(s_i^2 + \tau^2)^{-1/2} = \alpha + \mu(s_i^2 + \tau^2)^{-1/2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{4.1}$$

Like Egger's test, the intercept α is zero under the true model; in the presence of publication bias, it departs from zero. The overall mean effect size μ becomes the regression slope. Also, σ^2 allows potential under- or over-dispersion of the errors. In practice, heterogeneity is routinely assessed using the Q or I^2 statistic [6,10,13,95], and the between-study variance can be estimated as $\widehat{\tau}^2$ using the method of moments or the maximum restricted likelihood method [69,93]. If heterogeneity is not significant, then setting $\tau^2 = 0$ reduces Equation (4.1) to Egger's original test. Since the heterogeneity frequently appears in meta-analyses [78], this chapter will introduce publication bias measures based on the modified regression test.

Let the least squares estimates of the regression coefficients in model (4.1) be $\widehat{\alpha}$ and $\widehat{\mu}$. The estimated regression intercept is essential in the regression test; we denote this statistic as

$$T_I = \widehat{\alpha}.$$

Under the null hypothesis, T_I divided by its standard error follows the t-distribution with degrees of freedom $n - 2$, which gives the P-value of the regression test, denoted as P_I. Since the standardized effect sizes are unit-free, the estimated regression intercept T_I is also unit-free. Therefore, T_I can serve as a measure for quantifying publication bias [25]. However, the regression intercept T_I lacks an intuitive interpretation for the asymmetry of the collected study results. Meta-analysts usually report only the P-value of the regression test, not the magnitude of T_I, to describe the severity of publication bias.

4.2 Skewness and skewness-based test

The regression test does not fully describe the asymmetry of the collected study results. By linear regression theory, the estimated intercept can be expressed as $T_I = n^{-1}\sum_{i=1}^{n}\widehat{d}_i$, where

$$\widehat{d}_i = \frac{y_i - \widehat{\mu}}{\sqrt{s_i^2 + \widehat{\tau}^2}}$$

is an estimate of the study-specific standardized deviate $d_i = (y_i - \mu)(s_i^2 + \tau^2)^{-1/2}$. Therefore, the regression intercept T_I only reflects the *average* of the standardized deviates. To better test and quantify publication bias, we further consider the *shape* of the d_i's.

Note that $d_i = \alpha + \epsilon_i$, so the standardized deviates d_i are distributed with the same shape as the errors ϵ_i. To test the original H_0, we may alternatively test H_0'': $\alpha = 0$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ vs. H_1'': $\alpha \neq 0$ or ϵ_i's are iid from a skewed distribution with mean zero. Clearly, H_0'' is stronger than the null hypothesis H_0' of Egger's test, but it is still a necessary condition if the original null hypothesis H_0 holds. Hence, the statistical power should be enhanced by testing H_0'' compared to testing H_0'.

In the statistical literature, skewness has long been used as a descriptive quantity for the asymmetry of a distribution [96], but it is fairly novel in the literature of meta-analysis. To assess publication bias in meta-analysis, we may quantify the asymmetry of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$ by the skewness, calculated as $\mathrm{Skew}(\boldsymbol{\epsilon}) = m_3/s^3$, where $s = \left\{(n-1)^{-1}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2\right\}^{1/2}$ is the sample standard deviation, $m_3 = n^{-1}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^3$ is the sample third central moment, and $\bar{\epsilon} = n^{-1}\sum_{i=1}^{n}\epsilon_i$. In practice, we may replace the unknown errors $\boldsymbol{\epsilon}$ with the regression residuals $\widehat{\boldsymbol{\epsilon}} = (\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n)^{\mathrm{T}}$, where $\widehat{\epsilon}_i = \widehat{d}_i - T_I$. Denote the sample skewness of the errors as

$$T_S = \mathrm{Skew}(\widehat{\boldsymbol{\epsilon}}),$$

which we propose as an alternative measure of publication bias. We will show that T_S is a consistent estimate of the true skewness.

The sample skewness T_S can take any real value. A symmetric distribution (i.e., publication bias is not present) has zero skewness. A noticeably large positive skewness indicates that the right tail of standardized deviates' distribution is longer than its left tail. Therefore, some studies on the left side in the funnel plot (i.e., those with negative effect sizes) might be missing due to publication bias. In this situation, the regression intercept T_I is also expected to be positive. On the other hand, a large negative skewness implies that some studies may be missing on the right side. A common but rough rule of interpreting skewness is as follows. If the skewness is less than 0.5 in absolute magnitude, the distribution of the standardized deviates is approximately symmetric; the skewness is deemed considerable if it is between 0.5 and 1 in absolute magnitude, and it may be substantial if its absolute value is greater than 1. To interpret the skewness more rigorously, we study its large sample properties.

Denote $\beta_k = \mathrm{E}(\epsilon_1 - \beta)^k$ as the kth central moment of the errors ϵ_i, where $\beta = \mathrm{E}(\epsilon_1) = 0$, and the sample kth central moment is $m_k = n^{-1}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^k$. Then the true skewness of the errors is $\gamma = \beta_3/\beta_2^{3/2}$. In addition, let $\widehat{m}_k = n^{-1}\sum_{i=1}^{n}(\widehat{\epsilon}_i - \bar{\widehat{\epsilon}})^k$ be the sample kth central moment after plugging in the known residuals $\widehat{\epsilon}_i$; note that $\bar{\widehat{\epsilon}} = n^{-1}\sum_{i=1}^{n}\widehat{\epsilon}_i = 0$. Denote $\xrightarrow{D}$ as the convergence in distribution. We have the following proposition regarding the asymptotic distribution of the sample skewness T_S.

Proposition 4. *Assume that the study-specific errors ϵ_i have finite sixth central moment (i.e., $\beta_6 < \infty$) and the marginal precisions $(s_i^2+\tau^2)^{-1/2}$ have finite third moment. Then, $\sqrt{n}(T_S - \gamma)/\sqrt{\widehat{v}} \xrightarrow{D} N(0,1)$ as $n \to \infty$, where*

$$\widehat{v} = 9 + \frac{35}{4}\widehat{m}_2^{-3}\widehat{m}_3^2 - 6\widehat{m}_2^{-2}\widehat{m}_4 + \widehat{m}_2^{-3}\widehat{m}_6 + \frac{9}{4}\widehat{m}_2^{-5}\widehat{m}_3^2\widehat{m}_4 - 3\widehat{m}_2^{-4}\widehat{m}_3\widehat{m}_5.$$

Proposition 4 provides an approximate 95% confidence interval (CI) of the sample skewness T_S. Consequently, T_S not only quantifies publication bias but also serves as a test statistic. Under H_0'', we can simplify the asymptotic distribution of T_S as follows.

Corollary 1. *Under the null hypothesis H_0'', $\sqrt{n/6}T_S \xrightarrow{D} N(0,1)$ as $n \to \infty$.*

Appendix B.2 provides the proofs. The P-value of the skewness-based test is calculated using Corollary 1:

$$P_S = 2\left(1 - \Phi\left(\sqrt{n/6}|T_S|\right)\right).$$

The regression intercept T_I quantifies the departure of the average standardized deviate from zero; the skewness T_S quantifies the departure of the standardized deviates' distribution from symmetry. The regression test and the skewness-based test may differ in power in different situations. Therefore, we may combine the test results of T_I and T_S so that the combined test maintains high power across various settings. Under H_0'', note that T_I is the least squares estimate of the intercept and T_S depends only on the residuals $\widehat{\epsilon}_i$. Because the least squares estimates of regression coefficients are independent of the residuals if the errors ϵ_i are normally distributed, we immediately have the following proposition.

Proposition 5. *Under the null hypothesis H_0'', T_I and T_S are independent.*

Due to the independence of T_I and T_S, the adjusted P-value for combining T_I and T_S can be calculated as $P_C = 1 - (1 - P_{\min})^2$, where $P_{\min} = \min\{P_I, P_S\}$ [97]. The performance of the skewness-based test and the combined test will be studied using simulations and actual meta-analyses.

In practice, many meta-analyses only collect a small number of studies, and the large sample properties may apply poorly for them. Alternatively, a nonparametric bootstrap can be used to derive the 95% CI of the skewness: take samples of size n with replacement from the original data $\{(y_i, s_i^2)\}_{i=1}^n$ for B (say 1000) iterations and calculate 2.5% and 97.5% quantiles of the skewness over the B bootstrap samples. A parametric resampling method can also be used to produce a P-value for the skewness-based test. Specifically, first, estimate the overall mean effect size $\bar{\mu}$ under the null hypothesis that there is no publication bias. Second, draw n samples under the null hypothesis, i.e., $y_i^\star \sim N(\bar{\mu}, s_i^2 + \widehat{\tau}^2)$, and repeat this for B iterations. Third, based on the B sets of bootstrap samples, calculate the skewness as $T_S^{(b)}$ for $b = 1, \ldots, B$. Finally, the P-value of the skewness-based test is $P_S = \left[\sum_{b=1}^B \mathbb{I}(|T_S^{(b)}| \geq |T_S|) + 1\right] / (B + 1)$, where $\mathbb{I}(\cdot)$ is the indicator function. Similar procedures can also be used for the regression intercept T_I.

The code to implement the proposed methods will be included in our `R` package 'altmeta', available on the Comprehensive R Archive Network (CRAN).

4.3 Simulations

We performed simulations to evaluate the type I error rate and power of the modified regression test T_I, the proposed skewness-based test T_S, and the combined test based on the adjusted P-value P_C. The commonly-used Egger's regression test, Begg's rank test, and the trim and fill method (T & F) were also considered. In addition, we calculated the P-values of T_I and T_S using both their theoretical null distributions and the resampling methods. As suggested by many other authors (e.g., [32]), the nominal significance level was set to 10% for publication bias tests because the tests usually have low power. For each simulated meta-analysis, the true overall effect size was $\mu = 1$, the within-study standard errors were drawn from $s_i \sim U(1, 4)$, and the between-study standard deviation was set to $\tau = 0$ ($I^2 = 0\%$), 1 ($6\% \leq I^2 \leq 50\%$), and 4 ($50\% \leq I^2 \leq 94\%$). The study-specific effect sizes were then generated as $y_i \sim N(\mu_i, s_i^2)$ and $\mu_i \sim N(\mu, \tau^2)$. The number of studies collected in each meta-analysis was set to $n = 10$, 30, and 50. We considered the following three scenarios to induce publication bias.

I. (Suppressing non-significant findings) We used the above parameters to generate artificial studies, and suppose that they aimed at testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$. We assumed that studies with significant findings (i.e., P-value < 0.05 for treatment effect size) were published with probability 1. Also, studies with non-significant findings were published with probability π; the publication rate was set to $\pi = 0$, 0.02, 0.05, and 1. Note that $\pi = 1$ implies no publication bias. Studies were generated iteratively until we obtained n published studies to form a simulated meta-analysis.

II. (Suppressing small studies with non-significant findings) In many cases, small studies with non-significant findings are more likely to be suppressed than large studies; hence, some authors prefer to treat the funnel-plot-based methods as approaches to checking for 'small-study effects' [91]. We also simulated meta-analyses following this scenario. Studies with significant findings were published with probability 1. Large studies with non-significant findings and standard errors $s_i < 1.5$ were also published with probability 1; however, small studies with non-significant findings and standard errors $s_i \geq 1.5$ were published with probability π, where $\pi = 0$, 0.1, 0.2, and 1. Again, $\pi = 1$ implies no publication bias. The studies were generated iteratively until we obtained n published studies to form a simulated meta-analysis.

III. (Suppressing negative effect sizes) Publication bias can be also induced on the basis of study effect size [26, 33, 98]. For each simulated meta-analysis, $n + m$ studies were generated, and the m studies with the most negative effect sizes were suppressed. We set $m = 0$, $\lfloor n/3 \rfloor$, and $\lfloor 2n/3 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not greater than x. Note that $m = 0$ implies no publication bias.

For each setting, 10,000 meta-analyses were simulated. The Monte Carlo standard errors of all type I error rates and powers reported below were less than 1%.

Table 4.1 presents the type I error rates and powers for Scenario I. Type I error rates of most tests are controlled well, while that of Egger's test is a little inflated when the heterogeneity is substantial ($\tau = 4$). For weak or moderate heterogeneity ($\tau = 0$ or 1), Egger's regression test and the modified regression test T_I have similar power, and Begg's rank test seems to be more powerful than the regression test. Also, the trim and fill method performs poorly. Note that its power drops as π decreases from 0.05 to 0 when $n = 50$ and $\tau = 0$ or 1. Indeed, the trim and fill method is based on the assumption in Scenario III; that is, studies are suppressed if they have most negative (or positive) effect sizes, not according to their P-values. In Scenario I, the two-sided hypothesis testing for treatment effects $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ can produce significant findings with both negative and positive effect sizes, so the simulated meta-analyses can seriously violate the assumption of the trim and fill method.

For small meta-analysis with $n = 10$, using the asymptotic property in Corollary 1, the skewness-based test T_S is less powerful than the regression test and Begg's rank test when $\pi = 0.02$ or 0.05, and its type I error rate is much smaller than the nominal significance level 10%. This is possibly because T_S's asymptotic property is a poor approximation for small n. However, using the resampling method, the power of T_S is dramatically higher than the other tests when $\tau = 0$ and 1. Moreover, as the number of studies n increases to 30 and 50, the skewness-based test using either the asymptotic property or the resampling method still outperforms the other tests, and its power remains high as the heterogeneity becomes substantial ($\tau = 4$).

Table 4.2 shows the results for Scenario II. The regression test and Begg's rank test are more powerful than T_S when $\tau = 0$ and 1, while they are outperformed by T_S when $\tau = 4$. In this scenario, T_S seems to be less powerful than in Scenario I. For each simulated meta-analysis, because only small studies with non-significant findings were suppressed, large studies are still symmetric in the funnel plot. Consequently, the distribution of the n studies may have two modes: the large studies are centered around the true overall effect size μ, and the small studies have an overestimated mean due to the suppression. Since the interpretation of skewness is obscure for multi-modal distributions, T_S may lose power in this scenario.

Table 4.3 presents the type I error rates and powers for Scenario III. Since the trim and fill method's assumption is perfectly satisfied in this scenario, this method is generally more powerful than the other tests. In the absence of heterogeneity ($\tau = 0$), both the regression test and Begg's rank test are more powerful than the skewness-based test T_S; as the heterogeneity increases, they are outperformed by T_S, especially when n is large.

In summary, the skewness-based test T_S can be much more powerful than the existing tests in some settings, while no test can uniformly outperform the others. Although T_S suffers from low power when the heterogeneity is weak or moderate in Scenarios II and III, the combined test of T_I and T_S maintains high power in most settings by borrowing strengths from each of the separate test.

4.4 Case studies

We illustrate the performance of the skewness measure and test by three actual meta-analyses published in the *Cochrane Database of Systematic Reviews.* The first meta-analysis was performed by Stead et al. [99] to investigate the effect of nicotine gum for smoking cessation; it contains 56 studies and the effect size is the log risk ratio. The second meta-analysis, performed by Hróbjartsson and Gøtzsche [100], investigates the effect of placebo interventions for all clinical conditions regarding patient-reported outcomes; it contains 109 studies and the effect size is standardized mean difference. The third meta-analysis reported in Liu and Latham [101] compares the effect of the progressive resistance strength training exercise vs. control; it contains 33 studies and the effect size is also standardized mean difference. Figure 4.2 presents their contour-enhanced funnel plots; the shaded regions represent different significance levels [102].

The proposed methods and the commonly-used tests were applied to the three meta-analyses, and both the theoretical null distributions and the resampling methods were used to calculate the 95% CIs and P-values for T_I and T_S. We also calculated the P-values for the combined test. Table 4.4 presents the results. Since the size of each example n is large (for meta-analyses), the 95% CIs and P-values based on the theoretical null distributions are similar to those based on the resampling methods.

For the meta-analysis in Stead et al. [99], the three commonly-used tests yield P-values > 0.10, indicating non-significant publication bias; the P-value of the modified regression test T_I is also large. However, the proposed skewness T_S is 0.91 with 95% CI $(0.14, 1.68)$ and P-value 0.005 using the resampling methods; it implies substantial publication bias. Since T_S is significantly greater than zero, some studies with negative effect sizes may be missing. Indeed, the funnel plot in Figure 4.2(a) shows that most studies are massed on the right side, tending to have significant positive results; some studies are potentially missing on the left side. Moreover, benefiting from the high power of the skewness-based test, the combined test also indicates significant publication bias.

For the meta-analysis in Hróbjartsson and Gøtzsche [100], all tests imply significant publication bias; the P-values of Begg's rank test, the trim and fill method, and the skewness-based test are fairly small (< 0.01). Both the regression intercept T_I and the skewness T_S are significantly negative, indicating that some studies are missing on the right side in the funnel plot; Figure 4.2(b) confirms this. For the meta-analyses in Liu and Latham [101], Figure 4.2(c) shows that its funnel plot is approximately symmetric, so there appears to be no publication bias. Indeed, all tests yield P-values much greater than 0.1, and the publication bias measures T_I and T_S are close to zero.

4.5 Discussion

This chapter proposed a new measure, the skewness of the standardized deviates, for quantifying potential publication bias in meta-analysis. The intuitive interpretation of the asymmetry of the collected study results makes this measure appealing; its performance was illustrated by three actual meta-analyses. Also, the skewness can serve as a test statistic and its large sample properties have been studied. The simulations showed that the skewness-based test has high power in many cases. The large-sample properties of the skewness did not perform well for small n, but this can be remedied by using resampling methods. In addition, we proposed a combined test that depends on the P-values of both the regression and skewness-based tests; it is shown to be powerful in most simulation settings.

The proposed skewness has some limitations. First, for small meta-analyses, the variation of the sample skewness can be large. Researchers should always use skewness along with its 95% confidence interval. Second, although a symmetric distribution has zero skewness, zero skewness does not necessarily imply a symmetric distribution; for example, an asymmetric distribution may have zero skewness if it has a long but thin tail on one side and a short but fat tail on the other side. Also, the skewness generally describes publication bias well when the effect sizes are unimodal, but its interpretation for multi-modal distributions is obscure. Therefore, the regression intercept is preferred when the studies appear to have multiple modes, which may be identified by visual examining the funnel plot. Third, like many other approaches to assessing publication bias, the skewness is based on checking the funnel plot's asymmetry. However, such asymmetry can be caused by sources other than publication bias [90], such as reference bias [103, 104], studies with poor quality in design [105, 106], the existence of multiple subgroups [34], etc. When applying the methods in this chapter to detect or quantify the asymmetry of study results, researchers may need to examine carefully whether the asymmetry is caused by publication bias or other sources of bias. In addition, in the simulations and actual meta-analyses, different methods for publication bias can lead to fairly different conclusions. Therefore, we are allowed to use a wealth of methods to detect any potential publication bias.

Like the routinely-used I^2 statistic for assessing heterogeneity, the skewness may be a good characteristic of meta-analysis for quantifying publication bias. In the statistical literature, the skewness is a conventional descriptive quantity for asymmetry, but it may not be optimal to serve as a test statistic; more sophisticated tests for a continuous distribution have been extensively discussed (e.g., [107–109]). Exploring more powerful tests based on the standardized deviates warrants future study.

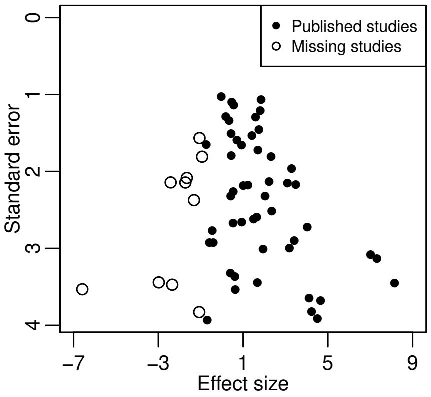


Figure 4.1: The funnel plot of a simulated meta-analysis containing 60 studies. The 10 studies with the most negative effect sizes were suppressed due to publication bias, and the remaining 50 studies were ‘published’.

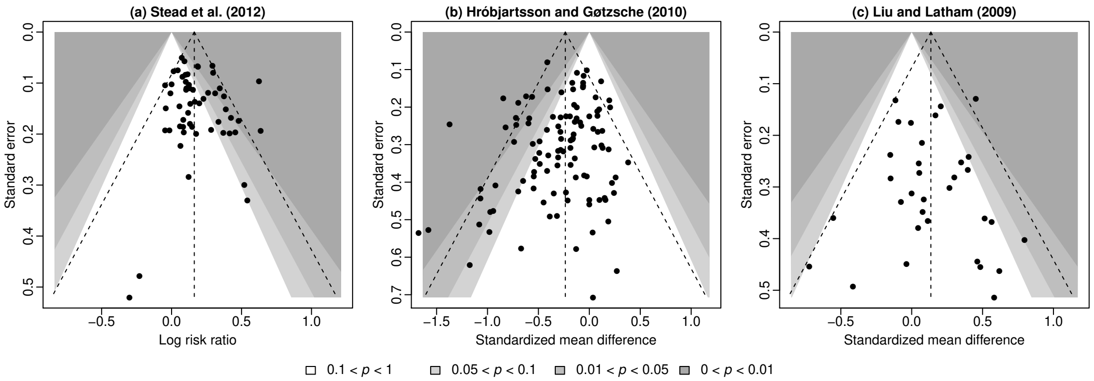


Figure 4.2: Contour-enhanced funnel plots of the three actual meta-analyses. The vertical and diagonal dashed lines represent the overall estimated effect size and its 95% confidence limits, respectively, based on the fixed-effect model. The shaded regions represent different significance levels for the effect size.

Table 4.1: Type I error rates ($\pi = 1$) and powers ($\pi < 1$) expressed as percentage, for various tests for publication bias due to suppressing non-significant findings (Scenario I).

Test	$\tau = 0$ ($I^2 = 0\%$)				$\tau = 1$ ($6\% \le I^2 \le 50\%$)				$\tau = 4$ ($50\% \le I^2 \le 94\%$)			
	$\pi = 1$	$\pi = 0.05$	$\pi = 0.02$	$\pi = 0$	$\pi = 1$	$\pi = 0.05$	$\pi = 0.02$	$\pi = 0$	$\pi = 1$	$\pi = 0.05$	$\pi = 0.02$	$\pi = 0$
$n = 10$:												
Egger	10	15	23	35	11	14	20	31	13	10	10	11
Begg	7	13	28	57	5	12	23	44	5	4	4	4
T & F	11	8	12	30	11	7	10	21	5	8	9	9
T_I	10	17	26	40	10	17	25	39	10	14	15	16
${T_I}^*$	[9]	[21]	[29]	[41]	[11]	[19]	[27]	[39]	[9]	[15]	[17]	[17]
T_S	1	7	20	37	1	8	18	32	1	3	3	4
${T_S}^*$	[10]	[27]	[48]	[59]	[10]	[29]	[46]	[58]	[10]	[15]	[17]	[19]
Combined	6	14	29	61	6	14	27	52	5	9	10	11
Combined*	[10]	[26]	[50]	[75]	[10]	[27]	[47]	[68]	[8]	[15]	[17]	[18]
$n = 30$:												
Egger	10	17	27	45	10	14	23	35	14	11	12	12
Begg	7	28	64	97	7	24	55	89	5	4	5	6
T & F	12	16	18	17	13	19	20	18	9	21	21	20
T_I	10	18	27	42	10	17	25	36	10	15	16	18
${T_I}^*$	[9]	[22]	[33]	[49]	[11]	[21]	[31]	[43]	[10]	[18]	[20]	[22]
T_S	6	50	83	94	6	59	83	92	5	16	20	24
${T_S}^*$	[10]	[61]	[88]	[96]	[10]	[70]	[88]	[94]	[10]	[26]	[30]	[34]
Combined	8	42	77	93	8	48	76	90	8	16	19	23
Combined*	[10]	[53]	[85]	[96]	[11]	[61]	[84]	[94]	[9]	[23]	[28]	[32]
$n = 50$:												
Egger	9	20	35	58	11	17	28	46	14	12	13	14
Begg	7	38	83	100	7	33	75	98	5	5	7	9
T & F	12	20	17	10	12	23	19	13	9	18	18	18
T_I	9	19	31	49	10	18	28	43	10	16	18	20
${T_I}^*$	[9]	[24]	[38]	[57]	[11]	[23]	[34]	[51]	[10]	[19]	[21]	[24]
T_S	7	77	96	99	7	84	96	98	7	30	36	41
${T_S}^*$	[10]	[82]	[97]	[99]	[10]	[87]	[97]	[99]	[10]	[37]	[44]	[49]
Combined	8	67	94	99	9	75	93	98	8	25	30	35
Combined*	[9]	[74]	[96]	[100]	[11]	[81]	[96]	[99]	[9]	[31]	[36]	[42]

The nominal significance level is 10%.

* The results in square brackets are based on the parametric resampling method.

Table 4.2: Type I error rates ($\pi = 1$) and powers ($\pi < 1$) expressed as percentage, for various tests for publication bias due to suppressing small studies with non-significant findings (Scenario II).

Test	$\tau = 0$ ($I^2 = 0\%$)				$\tau = 1$ ($6\% \leq I^2 \leq 50\%$)				$\tau = 4$ ($50\% \leq I^2 \leq 94\%$)			
	$\pi = 1$	$\pi = 0.2$	$\pi = 0.1$	$\pi = 0$	$\pi = 1$	$\pi = 0.2$	$\pi = 0.1$	$\pi = 0$	$\pi = 1$	$\pi = 0.2$	$\pi = 0.1$	$\pi = 0$
$n = 10$:												
Egger	10	14	22	51	11	13	19	43	13	9	10	12
Begg	7	8	13	30	5	7	12	30	5	4	5	7
T & F	11	10	11	15	11	9	10	13	5	4	5	5
T_I	10	15	23	56	10	14	23	54	10	13	16	21
${T_I}^*$	[9]	[19]	[29]	[61]	[11]	[19]	[28]	[59]	[9]	[15]	[18]	[25]
T_S	1	1	1	5	1	1	1	5	1	1	1	3
${T_S}^*$	[10]	[10]	[11]	[19]	[10]	[9]	[11]	[22]	[10]	[9]	[11]	[17]
Combined	6	9	16	48	6	8	15	46	5	7	9	14
Combined*	[10]	[15]	[23]	[58]	[10]	[14]	[22]	[55]	[8]	[10]	[14]	[21]
$n = 30$:												
Egger	10	20	34	69	10	18	30	62	14	10	12	16
Begg	7	16	30	68	7	14	28	66	5	5	7	13
T & F	12	18	23	32	13	15	17	21	9	13	14	13
T_I	10	21	36	70	10	20	33	66	10	14	18	25
${T_I}^*$	[9]	[24]	[40]	[74]	[11]	[23]	[37]	[71]	[10]	[17]	[22]	[32]
T_S	6	5	12	54	6	6	14	58	5	6	10	21
${T_S}^*$	[10]	[10]	[18]	[59]	[10]	[10]	[21]	[64]	[10]	[11]	[17]	[31]
Combined	8	16	30	80	8	14	28	75	8	10	14	24
Combined*	[10]	[20]	[36]	[83]	[11]	[18]	[33]	[81]	[9]	[13]	[20]	[33]
$n = 50$:												
Egger	9	26	46	82	11	24	41	78	14	12	14	20
Begg	7	21	43	85	7	19	41	84	5	5	9	19
T & F	12	17	19	21	12	14	15	13	9	12	12	10
T_I	9	26	46	82	10	25	42	79	10	15	19	29
${T_I}^*$	[9]	[29]	[50]	[85]	[11]	[27]	[46]	[82]	[10]	[19]	[24]	[36]
T_S	7	7	20	79	7	9	24	83	7	10	18	36
${T_S}^*$	[10]	[10]	[25]	[81]	[10]	[11]	[30]	[85]	[10]	[14]	[24]	[43]
Combined	8	20	41	92	9	19	39	89	8	12	18	34
Combined*	[9]	[23]	[46]	[93]	[11]	[22]	[44]	[91]	[9]	[16]	[24]	[41]

The nominal significance level is 10%.

* The results in square brackets are based on the parametric resampling method.

Table 4.3: Type I error rates ($m = 0$) and powers ($m > 0$) expressed as percentage, for various tests for publication bias due to suppressing the m most negative effect sizes out of a total of $n + m$ studies (Scenario III).

Test	$\tau = 0$ ($I^2 = 0\%$)			$\tau = 1$ ($20\% \leq I^2 \leq 50\%$)			$\tau = 3$ ($70\% \leq I^2 \leq 90\%$)		
	$m = 0$	$\lfloor n/3 \rfloor$	$\lfloor 2n/3 \rfloor$	$m = 0$	$\lfloor n/3 \rfloor$	$\lfloor 2n/3 \rfloor$	$m = 0$	$\lfloor n/3 \rfloor$	$\lfloor 2n/3 \rfloor$
$n = 10$:									
Egger	10	21	31	10	19	25	13	15	14
Begg	6	12	18	6	10	14	4	5	6
T & F	11	27	38	11	25	33	5	13	18
T_I	10	21	31	10	18	25	10	11	13
${T_I}^*$	[9]	[12]	[13]	[11]	[13]	[12]	[9]	[12]	[13]
T_S	2	2	4	2	2	3	1	2	3
${T_S}^*$	[10]	[13]	[17]	[10]	[13]	[16]	[10]	[14]	[16]
Combined	6	14	20	6	12	16	6	7	8
Combined*	[9]	[13]	[15]	[11]	[13]	[15]	[8]	[13]	[16]
$n = 30$:									
Egger	10	57	77	11	44	60	14	18	20
Begg	8	46	67	7	35	54	5	12	17
T & F	13	87	97	13	81	92	9	51	63
T_I	10	57	77	10	44	60	10	14	16
${T_I}^*$	[10]	[38]	[46]	[12]	[33]	[39]	[10]	[13]	[17]
T_S	6	25	40	6	25	39	6	26	40
${T_S}^*$	[10]	[34]	[51]	[11]	[34]	[51]	[10]	[37]	[52]
Combined	8	54	76	8	43	64	8	23	35
Combined*	[10]	[42]	[56]	[12]	[39]	[53]	[9]	[30]	[44]
$n = 50$:									
Egger	10	77	93	11	61	80	14	19	22
Begg	8	69	89	8	56	76	5	18	26
T & F	12	98	100	13	95	99	9	69	75
T_I	10	77	93	11	61	80	10	16	20
${T_I}^*$	[10]	[59]	[74]	[12]	[52]	[61]	[10]	[15]	[19]
T_S	8	46	69	7	47	68	7	51	69
${T_S}^*$	[10]	[53]	[75]	[10]	[54]	[75]	[10]	[58]	[76]
Combined	9	77	95	10	67	87	8	44	62
Combined*	[10]	[65]	[85]	[12]	[62]	[79]	[9]	[49]	[67]

The nominal significance level is 10%.

* The results in square brackets are based on the parametric resampling method.

Table 4.4: Results of assessing publication bias for three actual meta-analyses.

Meta-analysis	No. of	I^2 (%)	P-value			Intercept T_I			Skewness T_S			P-value of the
	studies		Egger	Begg	T & F	Measure	95% CI	P-value	Measure	95% CI	P-value	combined test
Stead et al.	56	39	0.173	0.136	0.500	0.47	$(-0.47, 1.41)$	0.323	0.91	$(0.14, 1.68)$	0.005	0.011
							$[-0.43, 1.42]$	[0.317]		$[0.06, 1.50]$	[0.005]	[0.010]
Hróbjartsson and Gøtzsche	109	42	0.049	0.009	<0.001	-0.81	$(-1.54, -0.09)$	0.028	-0.74	$(-1.23, -0.24)$	0.002	0.003
							$[-1.56, -0.10]$	[0.030]		$[-1.17, -0.25]$	[0.002]	[0.004]
Liu and Latham	33	11	0.905	0.469	0.500	0.06	$(-0.91, 1.02)$	0.905	0.01	$(-0.63, 0.64)$	0.989	0.991
							$[-1.09, 1.25]$	[0.894]		$[-0.73, 0.68]$	[0.987]	[0.989]

The results in square brackets are based on the parametric resampling method.

Chapter 5

Bayesian Multivariate Meta-Analysis of Multiple Factors

This chapter proposes *multivariate meta-analysis of multiple factors* (MVMA-MF) to jointly synthesize all risk and protective factors in a field-wide systematic review. Using the information across multiple factors, this method can produce better estimates of association measures between the factors and the disease condition, compared with separate meta-analyses. Multivariate meta-analysis methods have gained much attention in the recent literature [110–113]. They improve effect estimates by borrowing information on the correlations between multiple endpoints [114]. Multivariate meta-analysis methods have been applied to several areas, such as meta-analysis of diagnostic tests [115–117], meta-analysis of multiple outcomes [118,119], and network meta-analysis of mixed treatment comparisons [51, 53, 56, 67]. Mixed treatment comparisons use both direct and indirect evidence of treatment contrasts to synthesize the comparisons between multiple treatments; its focus is different from MVMA-MF, because MVMA-MF is concerned with estimating the effect of multiple factors, but not the contrasts between them.

A multivariate random-effects model generally requires estimates of correlations within each collected study. In some situations, within-study correlations are known to be zero. For example, in meta-analysis of diagnostic tests, the study-specific sensitivity and specificity are statistically independent within studies because they are calculated from the true negative and true positive patients, respectively [111]. However, in MVMA-MF, the factors can be correlated within each study because they may be measured on the same patients. Such within-study correlations are unknown unless individual patient data are available. Ignoring within-study correlations in the standard multivariate random-effects model may have a great impact on the estimated overall effect sizes [120].

To deal with unknown within-study correlations, this chapter considers an alternative Bayesian model for MVMA-MF. The conventional multivariate model partitions the overall covariance matrix into two parts: the within-study level that is due to sampling error, and the between-study level that is due to heterogeneity between the collected studies. Instead of partitioning the overall correlations into the two levels, the alternative model directly specifies one single overall correlation matrix; hence, it may be viewed as a *hybrid* approach. This model is the Bayesian version of the model introduced by Riley et al. [121]. Currently, Riley's model is implemented in a frequentist way, such as using the restricted maximum likelihood method [122]. However, the data for a MVMA-MF are usually fairly sparse (e.g., Table 5.1), and our simulation study in Appendix A.6 shows that the frequentist method generates poor 95% confidence intervals for sparse data; also, the algorithm for maximizing the (restricted) likelihood does not converge for many simulated data. Instead of using the frequentist method, a fully Bayesian approach is applied to perform MVMA-MF. Both the simulations and the case study demonstrate the benefit of joint modeling.

5.1 The motivating pterygium data

Instead of reporting only one risk factor at a time, Serghiou et al. [49] collected the odds ratios of all putative risk factors for pterygium, an eye disease. Specifically, they identified 60 eligible studies reporting on a total of 65 risk factors. Since most risk factors were only reported in less than 3 studies, we focus on the following 8 risk factors, each of which was reported in at least 4 studies: (1) occupation type (outdoor vs. indoor); (2) smoking status (yes vs. no); (3) education attainment (low vs. high); (4) use of hat (yes vs. no); (5) use of spectacles (yes vs. no); (6) area of residence (rural vs. urban); (7) use of sunglasses (yes vs. no); and (8) latitude of residence (low vs. high). These risk factors are sorted from high to low according to their frequencies reported in the collected studies. Also, we cleaned the data by removing the log odds ratios that were obtained using a multivariate regression model, because they were adjusted for different risk factors in different studies. Table 5.1 presents the cleaned data and Figure 5.1 shows the network plot of the 8 risk factors. The network indicates that most pairs of risk factors are simultaneously reported in some studies, but several pairs, such as 'area of residence' and 'use of hat', are not. From Table 5.1, the risk factor 'latitude of residence' was reported in only 4 studies, while 23 studies reported 'occupation type'. Most studies reported different subsets of the 8 risk factors, and many entries in Table 5.1 are missing. The estimated overall effect sizes produced by univariate models may be poor because some risk factors have data in few studies. Also, many factors (e.g., 'area of residence' and 'education attainment') are expected to be correlated, so a multivariate model may be more appropriate for this dataset than univariate models.

5.2 Conventional meta-analysis models

This section reviews some existing models for general multivariate meta-analysis. Suppose that n independent studies are collected; each study reports a p-dimensional vector of effect sizes, denoted as $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$. Denote its within-study covariance matrix as $\mathbf{S}_i$ $(i = 1, \ldots, n)$. The conventional univariate meta-analysis pools the results for each $j = 1, \ldots, p$ separately; a fixed- or random-effects model is applied to the data $\{(y_{ij}, v_{ij})\}_{i=1}^{n}$, where v_{ij} is the within-study variance, i.e., the jth diagonal element in $\mathbf{S}_i$ [6,7]. Since most studies were conducted by different research teams in different places using different methods, the studies are usually expected to be heterogeneous [78]. Also, the random-effects model may produce more conservative results than the fixed-effects model [123,124], so this chapter focuses on the random-effects setting that accounts for the heterogeneity between studies. We denote the univariate model as *Model U*, which ignores both within- and between-study correlations.

Multivariate meta-analysis has recently gained much attention for simultaneously synthesizing the p-dimensional effect sizes [110–112]. Given that the within-study covariance matrices $\mathbf{S}_i$ are known, the commonly used random-effects model is specified as follows to analyze the multivariate data $\{(\boldsymbol{y}_i, \mathbf{S}_i)\}_{i=1}^n$:

$$\begin{aligned} \boldsymbol{y}_i &\sim N(\boldsymbol{\mu}_i, \mathbf{S}_i); \\ \boldsymbol{\mu}_i &\sim N(\boldsymbol{\mu}, \mathbf{T}), \end{aligned} \tag{5.1}$$

where $\boldsymbol{\mu}_i$ represents study i's underlying true effect sizes, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ contains the overall effect sizes, and $\mathbf{T} = (\tau_{ij})$ is the $p \times p$ between-study covariance matrix. We denote this multivariate model as *Model M*. Note that within-study variances are routinely reported in published articles, but within-study correlations are usually unavailable. Let $\mathbf{D}_i = \mathrm{diag}(\mathbf{S}_i)$ be the diagonal matrix consisting of the within-study variances.

To analyze the data $\{(\boldsymbol{y}_i, \mathbf{D}_i)\}_{i=1}^n$ when the within-study correlations are unknown, a naïve multivariate method is to simply ignore these correlations by setting them to 0 but still account for the between-study correlations; we denote this model as *Model M_0*. Nevertheless, ignoring within-study correlations could lead to poor estimated effect sizes, especially when the within-study correlations are comparable to or greater than the between-study correlations [120]. The following section introduces an alternative model that can incorporate both within- and between-study correlations for MVMA-MF.

5.3 Multivariate meta-analysis of multiple factors

5.3.1 Multivariate hybrid meta-analysis model

Model M may be deemed ideal to perform MVMA-MF: it uses the factors' within-study correlations that are usually unknown, and it provides a benchmark for the performance of other potential models. Note that study i's marginal covariance matrix in Model M is $\mathbf{M}_i = \mathbf{S}_i + \mathbf{T}$, which can be written as $\mathbf{M}_i = (\mathbf{D}_i + \boldsymbol{\Delta})^{1/2}\mathbf{R}_i(\mathbf{D}_i + \boldsymbol{\Delta})^{1/2}$, where $\boldsymbol{\Delta} = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$ contains the between-study variances (i.e., diagonal elements in $\mathbf{T}$), and $\mathbf{D}_i$ is a diagonal matrix containing the within-study variances (i.e., diagonal elements in $\mathbf{S}_i$). The marginal correlation matrix of study i, $\mathbf{R}_i$, is determined by both $\mathbf{S}_i$ and $\mathbf{T}$, and thus needs to be estimated if the within-study correlations are unknown.

It may be inefficient to use the data merely from the n studies to simultaneously estimate all the $\mathbf{R}_i$'s, which involve too many parameters.

Alternatively, extending the bivariate model in Riley et al. [121], we consider a multivariate model that does not require within-study correlations to perform MVMA-MF:

$$\boldsymbol{y}_i \sim N\left(\boldsymbol{\mu}, (\mathbf{D}_i + \boldsymbol{\Psi})^{1/2}\mathbf{R}(\mathbf{D}_i + \boldsymbol{\Psi})^{1/2}\right),$$

where $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1^2, \ldots, \psi_p^2)$ is a diagonal matrix that consists of additional variances beyond sampling error due to between-study heterogeneity for the p effect sizes. In this model, all collected studies are assumed to share a common marginal correlation matrix $\mathbf{R}$. This assumption effectively reduces the number of parameters to be estimated and accounts for both within- and between-study correlations. The simulation study in Appendix A.6 generates data with different study-specific correlation matrices; it shows that the alternative model still performs well even if its assumption that $\mathbf{R}_i \equiv \mathbf{R}$ does not hold. Like Model M, the alternative model partitions the marginal variances of $\boldsymbol{y}_i$, $\mathbf{D}_i + \boldsymbol{\Psi}$, into the within- and between-study levels; however, it directly uses the matrix $\mathbf{R}$ to model the overall correlations, instead of partitioning the correlations into the previous two levels. Therefore, the alternative model may be deemed hybrid, and we denote it as *Model H*.

5.3.2 Missing data

So far, only models for complete data have been discussed. In MVMA-MF, each collected study often reports a small subset of the complete set of factors, and many factors are missing, as in Table 5.1. It is straightforward to extend the four methods (Models U, M, M$_0$, and H) to deal with missing data; Model H for missing data will be detailed here. Suppose that $\widetilde{\boldsymbol{y}}_i = (\widetilde{y}_{i1}, \ldots, \widetilde{y}_{ip})^{\mathrm{T}}$ contains the complete p factors in study i; however, we only observe t_i factors and their within-study variances ($1 \leq t_i \leq p$). Denote the effect sizes of the t_i factors as $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{it_i})^{\mathrm{T}}$, which is a t_i-dimensional sub-vector of $\widetilde{\boldsymbol{y}}_i$, and let $\mathbf{D}_i$ be the $t_i \times t_i$ diagonal matrix containing the within-study variances. We write $\boldsymbol{y}_i = \mathbf{X}_i\widetilde{\boldsymbol{y}}_i$, where $\mathbf{X}_i = (\boldsymbol{e}_{i1}, \ldots, \boldsymbol{e}_{it_i})^{\mathrm{T}}$ is a $t_i \times p$ matrix indicating missingness. Specifically, for each $j = 1, \ldots, t_i$, we define $\boldsymbol{e}_{ij} = (e_{ij1}, \ldots, e_{ijp})^{\mathrm{T}}$ with $e_{ijk} = 1$ if the observed y_{ij} is the effect size of factor k, and $e_{ijk} = 0$ otherwise. For example, for study 3 in the pterygium data (Table 5.1),

$$\mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Recall that $\boldsymbol{\mu}$, $\boldsymbol{\Psi}$, and $\mathbf{R}$ represent the overall mean effect sizes, the between-study variances, and the marginal correlation matrix for the complete p factors, respectively. Hence, for the observed t_i-dimensional vector $\boldsymbol{y}_i$, its overall mean is $\mathbf{X}_i\boldsymbol{\mu}$, its marginal variances form the diagonal matrix $\mathbf{D}_i + \mathbf{X}_i\boldsymbol{\Psi}\mathbf{X}_i^{\mathrm{T}}$, and its marginal correlation matrix is $\mathbf{X}_i\mathbf{R}\mathbf{X}_i^{\mathrm{T}}$. Consequently, the hybrid model for missing data can be specified as

$$\boldsymbol{y}_i \sim N\left(\mathbf{X}_i\boldsymbol{\mu}, \boldsymbol{\Phi}_i\right), \text{where } \boldsymbol{\Phi}_i = (\mathbf{D}_i + \mathbf{X}_i\boldsymbol{\Psi}\mathbf{X}_i^{\mathrm{T}})^{1/2}\mathbf{X}_i\mathbf{R}\mathbf{X}_i^{\mathrm{T}}(\mathbf{D}_i + \mathbf{X}_i\boldsymbol{\Psi}\mathbf{X}_i^{\mathrm{T}})^{1/2}.$$

The simulation study in Appendix A.6 compares the performance of Model H with Models M, M_0, and U when some factors are missing under various mechanisms. The performance of Model H is shown to be close to the ideal Model M that requires unknown within-study correlations. When factors are missing not at random (e.g., in the presence of publication bias), Model H produces estimated overall effect sizes with smaller biases and mean squared errors and larger 95% credible interval (CrI) coverage probabilities, compared with Models M_0 and U.

5.3.3 Bayesian hybrid model

Currently existing statistical software, such as the `Stata` command 'mvmeta', can only implement Model H in a frequentist way [121,122]. However, when the dimension of factors p is large compared with the number of collected studies, the estimated covariance matrix using the frequentist method may be quite inconsistent [125], leading to poor interval estimates. Indeed, the simulation study in Appendix A.6 shows that the frequentist method produces poor 95% confidence intervals when the data for MVMA-MF are sparse, and the algorithm for maximizing the (restricted) likelihood does not converge for many simulated datasets. Therefore, we use a fully Bayesian approach to estimating the overall multivariate effect size $\boldsymbol{\mu}$ and its covariance matrix, which are of interest in Model H. Vague priors are assigned to both the mean and variance-covariance structures. Appendix A.5 provides the details of the implementation. The

R code for the Bayesian MVMA-MF will be provided in our package 'altmeta', freely available at `http://cran.r-project.org/package=altmeta`. The simulation study in Appendix A.6 indicates that the 95% CrIs obtained using the Bayesian method generally have higher coverage probabilities than those obtained using the frequentist method for sparse data, which are common in MVMA-MF.

5.4 Real data analysis

Section 5.1 introduced the pterygium dataset in detail. Since the within-study correlations are unknown, we used Models H, M_0, and U but not Model M to estimate the overall log odds ratios of the 8 risk factors. Due to the sparsity of the dataset, the Bayesian method may be preferred. The Markov chain Monte Carlo (MCMC) algorithm was used to implement the Bayesian analysis with three chains; each chain contained a run of 100,000 updates after a 100,000-run burn-in period. The convergence of each chain was checked using trace plots. Table 5.2 presents the median overall log odds ratios with 95% CrIs. Figure 5.2 shows the posterior density plots of the 8 risk factors; each plot contains three density curves corresponding to the three models.

For risk factors that are reported in a relatively large number of studies (e.g., occupation type, smoking status, and education attainment), the three models yield similar estimated overall log odds ratios; their density curves are also fairly similar. For risk factors that are only reported in a few studies (e.g., use of spectacles, area of residence, and latitude of residence), the peaks of the posterior densities produced by Model H are narrower and higher compared with those produced by Models M_0 and U, indicating that Model H produces narrower 95% CrIs. Also, for the risk factor use of sunglasses, the location of its posterior density produced by Model H is noticeably different from those produced by the other two modes. Figure 5.3 depicts the estimated overall correlations between the 8 risk factors produced by Model H. It shows that many factors are correlated and some correlations are fairly high. Hence, ignoring the within-study correlations could lead to fairly different estimated log odds ratios; the unknown within-study correlations need to be carefully considered in MVMA-MF.

Furthermore, we performed a sensitivity analysis to investigate the impact of risk factor selection on the estimated overall log odds ratios. We considered two scenarios for the set of risk factors to be included in MVMA-MF: (i) the sub-dataset consists of the k most frequently reported risk factors; and (ii) the sub-dataset consists of the k least frequently reported risk factors ($k = 2, \dots, 8$). For example, when $k = 2$, the sub-dataset under scenario (i) contains the risk factors (1) occupation type and (2) smoking status; the sub-dataset under scenario (ii) contains the risk factors (7) use of sunglasses and (8) latitude of residence. The proposed hybrid model was implemented using the Bayesian method to analyze these sub-datasets.

Figure 5.4 shows the 95% CrIs of the overall log odds ratios under both scenarios; the labels of risk factors used in the figure can be found in Table 5.2's first column. In scenario (i), starting from the two most frequently reported risk factors, infrequently reported risk factors are iteratively added to the multivariate meta-analysis. Figure 5.4(i) shows that the estimated overall log odds ratios of risk factors 4 and 5 have some changes as new factors were added to the sub-datasets. The 95% CrIs of the three most frequently reported risk factors 1–3 change little. This might be explained by two reasons. First, the correlations between these three factors are weak (Figure 5.3), so the addition of risk factor 3 has little impact on estimating the effect sizes of factors 1 and 2. Also, the later added factors 4–8 have much smaller sample sizes (less than six studies) compared with factors 1–3, which are reported in more than ten studies, so the correlations may contribute little to the estimated effect sizes of factors 1–3.

Compared with scenario (i), Figure 5.4(ii) shows larger changes of estimated overall log odds ratios in scenario (ii). Under this scenario, starting from the two least frequently reported risk factors, more frequently reported risks are iteratively added to the multivariate meta-analysis. The 95% CrIs of infrequently reported risk factors, such as 5 and 6, become narrower as more reported risk factors are included in the MVMA-MF. This illustrates the benefit of jointly modeling multiple risk factors: the inference on infrequently reported risk factors can be strengthened by borrowing information from frequently reported risk factors through the correlations between them.

5.5 Discussion

This chapter proposed MVMA-MF with application to the pterygium data. In contrast to the tradition of meta-analyzing each single factor separately, we encourage researchers to collect all possible factors and analyze them jointly to enhance the estimation of overall effect sizes. A multivariate hybrid model was introduced to implement MVMA-MF in which within-study correlations are usually unknown. The simulation study in Appendix A.6 shows that the proposed method performs better than the univariate model and the model that ignores within-study correlations, especially when some factors are missing not at random.

An important issue of MVMA-MF is to incorporate the effect size of a certain factor that has been adjusted for other factors. For example, in the original pterygium data presented in Serghiou et al. [49], many collected studies report only log odds ratios that are obtained using multivariate regression after adjusting for different factors (e.g., age and gender), while log odds ratios without any adjustments are unavailable from these studies. We do not include such data in Table 5.1 due to the inconsistent adjustments. How to incorporate such data with different adjustments is of great interest to enrich the data for MVMA-MF and enhance the robustness and precision of MVMA-MF. We leave this to future studies.

Another interesting but challenging problem is to robustly impute the missing factors when the missingness is not at random; this missingness mechanism is closely related to publication bias [28]. Although the simulation study in Appendix A.6 shows that the proposed hybrid model performs better than Models M_0 and U, its performance (assessed by bias and 95% CrI coverage probability) is expected to be further improved in the presence of publication bias. Approaches to correcting publication bias have been introduced and widely used in univariate meta-analysis [26]; similar methods are highly needed for multivariate meta-analysis.

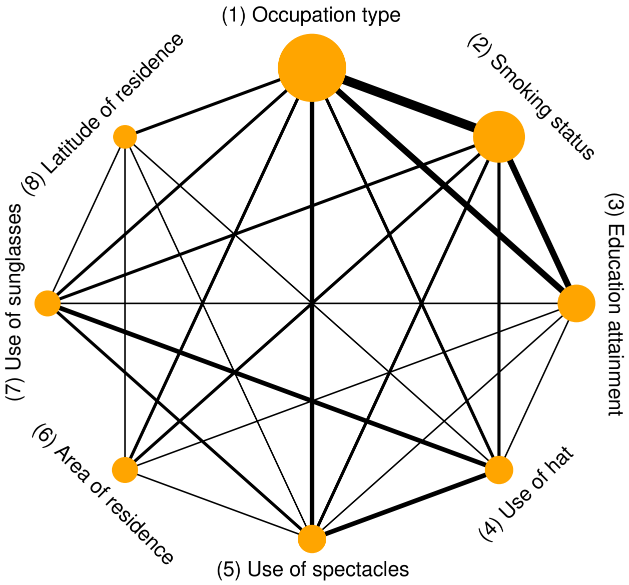


Figure 5.1: Network plot of the pterygium data. The nodes represent the risk factors, and the edge between two nodes indicate that these nodes are simultaneously reported in common studies. The node size is proportional to the number of studies that report the corresponding risk factor, and the edge thickness is proportional to the number of studies that simultaneously report the corresponding two risk factors.

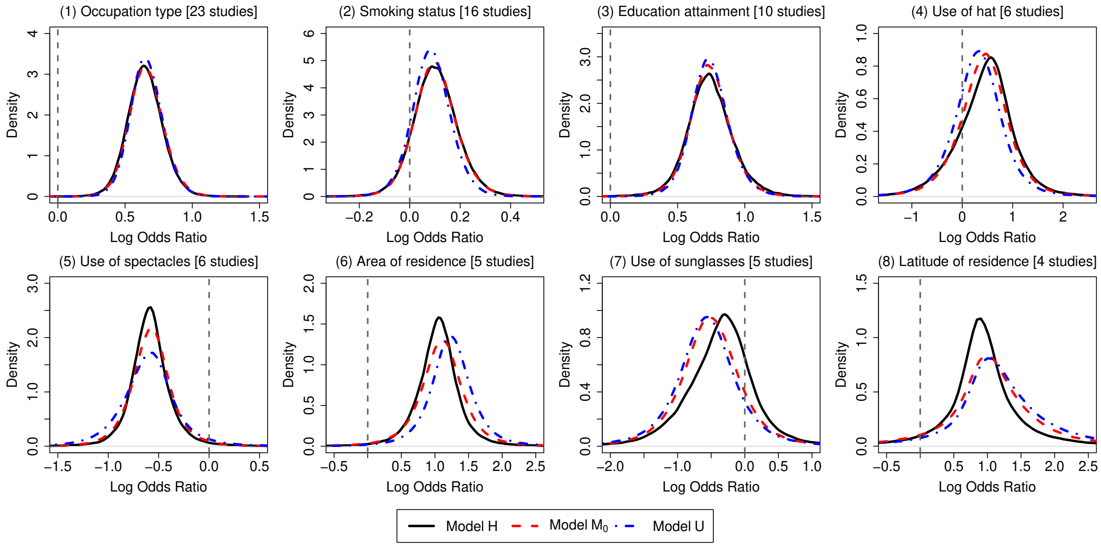


Figure 5.2: Posterior density plots produced by Models H (accounting for both between- and within-study correlations), Model M$_0$ (only accounting for between-study correlations), and Model U (ignoring both between- and within-study correlations) for the log odds ratios of the 8 risk factors in the pterygium data.

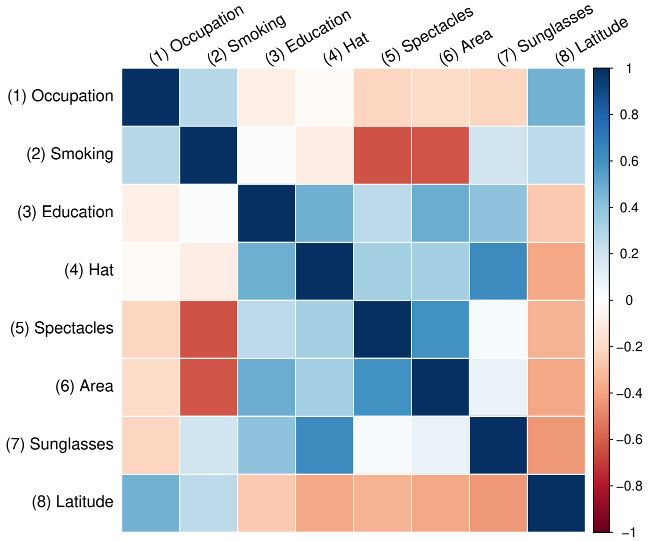


Figure 5.3: Plot of the estimated overall correlations between the 8 risk factors produced by Model H in the pterygium data. Darker color implies higher correlation.

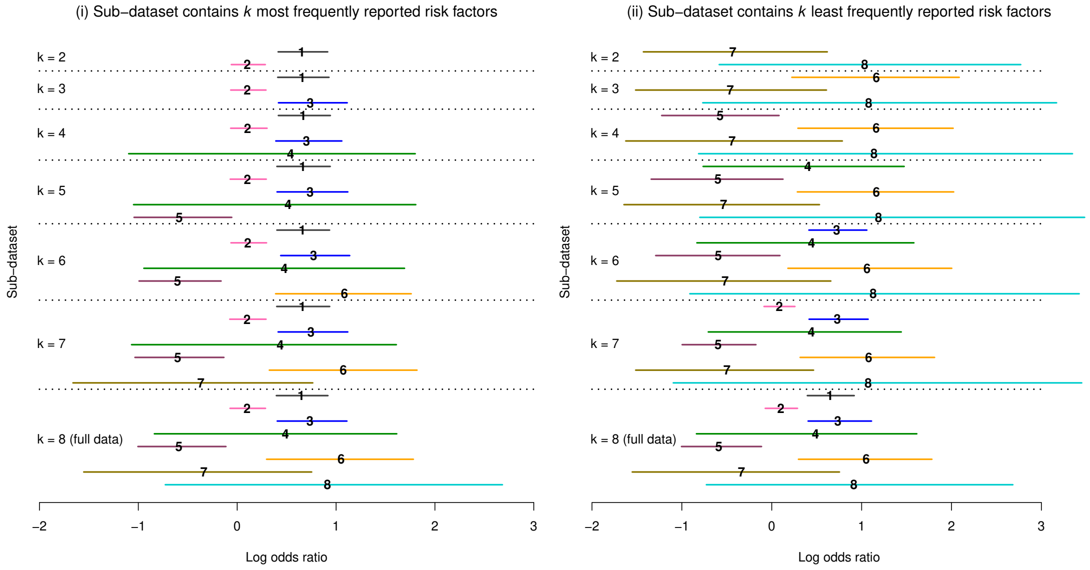


Figure 5.4: Bayesian estimates of log odds ratios produced by Model H based on subsets of the pterygium data. Each horizontal solid line represents 95% CrI of log odds ratio. The number placed at the median log odds ratio within each 95% CrI represents the corresponding risk factor's label. The results of the sub-datasets that contain $k = 2, \ldots, 8$ risk factors are accordingly listed from upper to lower, separated by the dotted lines.

Table 5.1: The pterygium data containing 29 studies with 8 risk factors. The effect size is log odds ratio with within-study standard error in parentheses. The blank entries indicate that the risk factors are unavailable from the corresponding studies.

Study	Risk factor							
	(1) Occupation	(2) Smoking	(3) Education	(4) Hat	(5) Spectacles	(6) Area	(7) Sunglasses	(8) Latitude
1	−0.08 (0.34)							
2						1.54 (0.10)		
3	0.28 (0.17)	0.53 (0.10)	0.53 (0.13)					
4	0.45 (0.11)					0.41 (0.10)		0.97 (0.23)
5	0.30 (0.40)							
6				0.12 (0.40)	0.48 (0.70)			
7	1.40 (0.23)							
8	0.39 (0.13)	0.05 (0.13)						
9	0.55 (0.22)	−0.04 (0.26)						
10	3.04 (1.03)							
11	1.95 (0.40)	1.21 (0.42)		0.67 (0.36)	−1.34 (0.34)		0.12 (0.31)	
12	1.10 (0.30)							
13	2.03 (0.39)			−0.69 (0.21)	−1.14 (0.38)		−1.64 (0.21)	2.99 (0.74)
14	0.83 (0.09)	0.11 (0.09)	0.91 (0.09)		−0.58 (0.12)	1.14 (0.13)		
15	0.41 (0.22)							
16		−0.20 (0.24)				1.73 (0.18)		
17	0.42 (0.15)	−0.11 (0.25)	0.38 (0.21)	0.22 (0.15)	−0.64 (0.15)		−0.52 (0.37)	
18	0.63 (0.12)	0.03 (0.05)	1.24 (0.24)					
19	0.89 (0.08)	−0.08 (0.07)		1.70 (0.08)	−0.17 (0.08)		−0.06 (0.14)	
20	0.39 (0.22)	−0.13 (0.30)						
21	0.90 (0.30)							0.85 (0.49)
22	−0.48 (0.25)	0.31 (0.25)	0.87 (0.20)					
23								0.66 (0.26)
24				−0.46 (0.72)			−0.73 (0.52)	
25	0.76 (0.34)	−0.48 (0.55)	1.05 (0.33)					
26		0.03 (0.11)	0.19 (0.09)					
27	1.16 (0.36)	0.53 (0.23)	0.83 (0.25)					
28	0.12 (0.03)	0.14 (0.20)	1.43 (0.28)			1.43 (0.22)		
29	0.41 (0.09)	−0.26 (0.10)	0.46 (0.09)					

Risk factors: (1) occupation type; (2) smoking status; (3) education attainment; (4) use of hat; (5) use of spectacles; (6) area of residence; (7) use of sunglasses; and (8) latitude of residence.

Table 5.2: The estimated overall log odds ratios (95% CrI) of the 8 risk factors in the pterygium data obtained by Models H (accounting for both between- and within-study correlations), Model M_0 (only accounting for between-study correlations), and Model U (univariate model ignoring both between- and within-study correlations) using the Bayesian method.

Risk factor	No. of Studies	Estimated overall log odds ratio		
		Model H	Model M_0	Model U
(1) Occupation type	23	0.65 (0.40, 0.92)	0.65 (0.41, 0.93)	0.66 (0.42, 0.91)
(2) Smoking status	16	0.10 (−0.07, 0.29)	0.10 (−0.07, 0.29)	0.08 (−0.07, 0.25)
(3) Education attainment	10	0.74 (0.40, 1.11)	0.74 (0.42, 1.07)	0.74 (0.46, 1.06)
(4) Use of hat	6	0.49 (−0.84, 1.61)	0.44 (−0.77, 1.55)	0.32 (−0.84, 1.41)
(5) Use of spectacles	6	−0.59 (−1.00, −0.11)	−0.58 (−1.06, −0.06)	−0.60 (−1.26, 0.05)
(6) Area of residence	5	1.05 (0.30, 1.78)	1.10 (0.22, 1.97)	1.24 (0.39, 2.11)
(7) Use of sunglasses	5	−0.34 (−1.55, 0.75)	−0.51 (−1.64, 0.62)	−0.57 (−1.76, 0.63)
(8) Latitude of residence	4	0.91 (−0.73, 2.68)	1.04 (−0.98, 3.10)	1.14 (−0.70,3.37)

Chapter 6

Sensitivity to Excluding Treatments in Network Meta-Analysis

This chapter examines the sensitivity to treatment exclusion of an alternative approach to network meta-analysis, namely the arm-based approach, recently developed from the perspective of missing data analysis [67]. The detailed model is briefly reviewed in Appendix A.8. This model assumes: 1) each study is independently chosen from a conceptual urn containing a large number of studies, and thus we can assign a joint distribution on the arm parameters independently across different studies; 2) each study hypothetically compares all treatments, many of which are missing at random. The arm-based model does not estimate the population-averaged absolute risk of each arm independently; instead, it respects the study randomization by accounting for the correlations between treatments within each study, which allows for 'borrowing information' across treatment arms. This point is illustrated by an example in Appendix A.7, in which absolute risk estimates from the arm-based model differ from estimates from a simple logit random effects model using only studies with a specific treatment arm. In addition, simulation results and real data analyses have shown that in some cases the effect size estimates given by this arm-based method are less biased than those given by the contrast-based model [67].

Besides reporting changes due to treatment exclusion in the population-averaged absolute risk estimates from the arm-based model, we compare changes in relative effects (i.e., log odds ratio change) with those obtained from the contrast-based model. In this regard, the arm-based and contrast-based methods have a key difference: If a study only has two treatment arms and one of these arms is omitted from the network meta-analysis, a contrast-based analysis must omit the entire study, while an arm-based analysis can retain the single remaining arm. Note that single-arm studies do contribute information to estimation of relative effects from the perspective of missing data analysis, which is somewhat counter-intuitive. To give a simple illustration, consider paired bivariate normally distributed random variables X and Y with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, e.g., the probit-transformed absolute risks in the arm-based model. The expected value of Y given X is $\mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x)$. Once we observe a value $X = x$ in a particular pair (with Y unobserved), the expected difference between Y and X for this pair becomes $\mu_y - x + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$, which does not equal $\mu_y - \mu_x$. Also, the variance of Y given X is $(1 - \rho^2)\sigma_y^2 \leq \sigma_y^2$. Therefore, even if Y is unobserved, modeling X and Y jointly, as in the arm-based model, helps reduce the standard error of a comparison. This point is illustrated by an example in Appendix A.7.

This chapter is organized as follows. Section 6.1 describes the specific network meta-analysis models being compared and the datasets to which we applied them. Section 6.2 presents results describing sensitivity of the network meta-analysis models to treatment exclusion. Section 6.3 closes with some suggestions on network meta-analysis and several limitations in our study.

6.1 Statistical analysis methods

6.1.1 Dataset selection

We reviewed forty network meta-analyses studied by Veroniki et al. [126] and selected fourteen networks containing 567 randomized controlled trials with a total of 389,361 participants. Our selection criteria were that every treatment in the network should be evaluated in at least three studies; otherwise, the networks are poorly connected at that treatment node. We denote the fourteen networks as *Ara 2009* [127], *Ballesteros 2005* [128], *Bucher 1997* [129], *Cipriani 2009* [61], *Eisenberg 2008* [130], *Elliott 2007* [131], *Lu 2006* [52,132], *Lu 2009* [54,133,134], *Middleton 2010* [135], *Mills 2009* [136], *Picard 2000* [137], *Puhan 2009* [138], *Thijs 2008* [139], and *Trikalinos 2009* [140]. Tables 6.1 and 6.2 lists their characteristics including the outcomes, the investigated treatments with their weighted node degrees, and the total number of studies, participants, and events. For each node (treatment) in a network, the weighted degree is defined as the sum of weights on all edges incident to that node. In a network meta-analysis, the edge weight equals the number of pairwise comparisons between two treatments, so the weighted degree represents the frequency with which a particular treatment is investigated in all of the network's studies. The node with the greatest weighted degree can be considered the most well-connected. Figure 6.1 shows network plots for the 14 datasets.

6.1.2 Performing network meta-analysis and removing treatments

We fit the arm-based and contrast-based network meta-analysis Bayesian hierarchical models separately to each of the 14 network datasets. Appendix A.8 gives details about these models. We used Markov chain Monte Carlo (MCMC) to compute posteriors for the effect sizes of interest, implemented using `JAGS` via the `R` package 'rjags'.

Analyses with a treatment removed were performed as follows. Suppose a network includes K treatments. We first applied both the arm-based and contrast-based models to the complete network dataset (the full network) to estimate log odds ratios comparing each pair from the K treatments; we also estimated the population-averaged treatment-specific absolute risks using the arm-based model. Next, for each treatment, we excluded it from the network and applied the analyses to the remaining dataset (the reduced network) consisting of $K - 1$ treatments. The key difference between the arm-based and contrast-based models becomes pertinent at this point. If a treatment was removed from a network, then for any two-arm studies that included that treatment, only one treatment arm remained. For an analysis using the arm-based model, we could keep the single-arm studies as they still contribute to the likelihood function from the perspective of an analysis with missing data. However, for an analysis using the contrast-based model, because it uses information about contrasts, the single remaining arm no longer provides any information for estimation in the reduced network, so the whole two-arm study must be deleted if one of the treatments is excluded. Multi-arm studies

– those comparing more than two treatments – that included the removed treatment were retained for analyses under both the arm-based and contrast-based models. For the present study, we did not consider any exclusion that creates a disconnected or poorly connected network, i.e., that resulted in at least one treatment in a network being evaluated in fewer than three studies. Table 6.2's 'ineligible trt removal' column shows treatment exclusions that produce such ineligible reduced networks under analyses with the arm-based and contrast-based models. When comparing the arm-based and contrast-based models, we only considered treatment removals that were eligible under both models. Appendix A.9 gives an example of which treatments were considered for exclusion.

6.1.3 Fold changes of estimated absolute risks in the arm-based model

For analysis using the arm-based model, we used fold changes of estimated population-averaged treatment-specific absolute risks to assess the impact of the treatment exclusion. Assume that the population-averaged absolute risk for a particular treatment is estimated as $\widehat{\pi}_f$ using the full network and $\widehat{\pi}_r$ using the reduced network. Then, the fold change for this treatment-specific absolute risk is defined as the maximum of $\widehat{\pi}_f/\widehat{\pi}_r$ and $\widehat{\pi}_r/\widehat{\pi}_f$. Thus, the fold change is never less than 1. Mills et al. [66] judged that a relative change not exceeding 1.03-fold is minor while a change greater than 1.10-fold is large, and over 1.20-fold is substantial, though such categorization is subjective and may need to be adapted to specific situations.

6.1.4 Comparison between arm-based and contrast-based methods

Without either external data or a separate model to estimate a reference treatment's absolute risk, the contrast-based method can only estimate odds ratios or their logarithms [58,65,67]. We focused on the changes of log odds ratio (LOR) when comparing the arm-based and contrast-based methods according to their sensitivity to treatment exclusion in the fourteen networks. For each network and treatment exclusion, we applied both models to the full and reduced networks. Then we calculated the LOR change (LORC) as the difference between the LOR estimates using the full and reduced networks: $\text{LORC}_{ij}^{(-k)} = \widehat{\text{LOR}}_{ij}^{(-k)} - \widehat{\text{LOR}}_{ij}$, where i, j index the treatments compared by this LOR and k indexes the treatment removed in the reduced network, $i \neq j \neq k$. $\widehat{\mathrm{LOR}}_{ij}$ and $\widehat{\mathrm{LOR}}_{ij}^{(-k)}$ are point estimates from the Bayesian analysis, which can be either the posterior means or medians; this chapter presents results for the posterior means.

Because of $\mathrm{LORC}_{ij}^{(-k)} + \mathrm{LORC}_{ji}^{(-k)} = 0$ by the symmetry of LOR, when we used statistical tests to compare the arm-based and contrast-based models according to sensitivity to treatment exclusion, we considered the absolute LOR change, i.e., $|\mathrm{LORC}_{ij}^{(-k)}|$. Further, the average absolute LOR change is calculated for each model by averaging every $|\mathrm{LORC}_{ij}^{(-k)}|$ comparing all possible pairs of treatments based on all eligible treatment exclusions across all networks; a smaller average absolute change indicates a more robust network meta-analysis model with respect to treatment exclusion. To demonstrate that LOR changes resulting from an individual treatment exclusion in a network may be in opposite directions for the arm-based and contrast-based models, we present $\mathrm{LORC}_{ij}^{(-k)}$ with their directions in Figure 6.2, rather than their absolute values. To preserve any correlation structure between treatments in a network, when testing the difference between the arm-based and contrast-based methods, we used bootstrap resampling [141] at the network level (using 10,000 bootstrap samples); that is, each bootstrap sample consisted of fourteen resampled networks, drawn with replacement from the original fourteen networks. Based on the bootstrap samples, we calculated 95% confidence intervals (CIs) and P-values for each model's mean absolute LOR change and their difference.

6.2 Results

6.2.1 Fold changes of estimated population-averaged absolute risks by the arm-based model

For the arm-based model, Table 6.3 reports the average and maximal fold changes of estimated population-averaged absolute risks and connectivity information about nodes associated with the maximal change for each network. The average fold changes in all networks are 1.05 or less; in 13 of 14 networks, the maximal change is below 1.10-fold. These small changes indicate the arm-based model's robustness to treatment exclusion. Although the arm-based model is robust in most cases, treatment exclusion can still produce significant changes (larger than 1.20-fold) for certain networks. For example, in the network *Trikalinos 2009*, excluding the treatment PTCA results in a 1.39-fold change in the absolute risk estimate of the treatment MT. Referring to Tables 6.1–6.3, several potential factors affecting the fold changes are listed below.

1. Removing treatments with larger weighted degree tends to cause larger fold changes, while the most affected treatment tends to have small weighted degree. In 6/14 networks, the maximal fold change is caused by the removal of treatment with the largest weighted degree; in 7/14 networks, the most affected treatment has the smallest weighted degree.

2. Including more studies and increasing network connectivity may help to reduce the impact of treatment exclusion. For example, *Cipriani 2009* examined 111 studies on 12 treatments, and each treatment is connected to at least three other treatments. All changes are smaller than 1.04-fold, and the average change is less than 1.01-fold. On the other hand, *Ara 2009* summarized only 11 studies on 5 treatments, and treatment exclusion caused changes as large as 1.09-fold.

3. Network meta-analyses with low event rates may produce large fold changes. For example, the naïve absolute risk in *Trikalinos 2009* is only 3.1%, and 22/124 treatment groups reported zero events (Table 6.2). Although this network includes 62 studies, the maximal fold change is 1.39.

We should note that the factors above are not sufficient or necessary conditions when judging whether a network is robust to treatment exclusion. For example, *Ballesteros 2005* has only 9 studies, but its average and maximal fold changes are smallest among the fourteen networks. The changes may be small in this network because it has a high naïve absolute risk.

6.2.2 Comparing the arm-based and contrast-based models

Figure 6.2 presents all LOR changes ($\text{LORC}_{ij}^{(-k)}$) due to treatment exclusions in the fourteen networks under the arm-based and contrast-based models. In Figure 6.2's upper panels, the LOR changes under the arm-based model are estimated including single-arm studies. The scatter plot indicates that LOR changes for the arm-based model tend to be smaller in magnitude than those for the contrast-based model. The empirical cumulative distribution function (ECDF) in top right panel of Figure 6.2 supports this observation. Because $\text{LORC}_{ij}^{(-k)} + \text{LORC}_{ji}^{(-k)} = 0$, the LOR changes between treatments i and j appear symmetrically in both the scatter plot and the ECDF graph. In addition, as each symmetric pair has the same absolute LOR change, we may only keep one value when we statistically test the difference between the arm-based and contrast-based models. (The resulting P-value remains the same if we include both values, as we use a nonparametric bootstrap resampling technique at the network level.)

Let μ_{AB} and μ_{CB} denote the true mean absolute LOR change (i.e., the expected value of $|\text{LORC}_{ij}^{(-k)}|$ across all treatment exclusions in all networks) under the arm-based and contrast-based models, respectively. Based on 10,000 bootstrap samples, μ_{AB} is estimated as 0.020 with 95% CI $(0.015, 0.031)$, and μ_{CB} is estimated as 0.047 with 95% CI $(0.029, 0.100)$; $\mu_{\text{CB}} - \mu_{\text{AB}}$ is estimated as 0.028 with 95% CI $(0.011, 0.071)$ and two-sided P-value 0.005 for testing $H_0 : \mu_{\text{AB}} = \mu_{\text{CB}}$ vs. $H_A : \mu_{\text{AB}} \neq \mu_{\text{CB}}$. Therefore, at 0.05 significance level, the absolute change under the contrast-based model is significantly larger than the change under the arm-based model, which suggests that the arm-based model is more robust than the contrast-based model to treatment exclusion.

To see whether the smaller average absolute LOR change caused by the arm-based model is due to the additional information it uses (that is, the retained single-arm studies), we applied the arm-based model to the same reduced networks that were used by the contrast-based model, in which single-arm studies were excluded. Figure 6.2's lower panels show the resulting LOR changes: the scatter plot and ECDF graph suggest that the arm-based and contrast-based models perform nearly the same when they use the same information. Let $\widetilde{\mu}_{\text{AB}}$ denote the true mean absolute LOR change when applying the arm-based model to the data used by the contrast-based model. Using the same bootstrap approach as above, the 95% CI for $\widetilde{\mu}_{\text{AB}}$ is $(0.044, 0.089)$ with point estimate 0.054. The point estimate is comparable to that of the contrast-based model, but the 95% CI is slightly narrower; $\mu_{\text{CB}} - \widetilde{\mu}_{\text{AB}}$ is estimated as -0.006 with 95% CI $(-0.049, 0.027)$, with two-sided P-value 0.59 for testing $H_0 : \widetilde{\mu}_{\text{AB}} = \mu_{\text{CB}}$ vs. $H_A : \widetilde{\mu}_{\text{AB}} \neq \mu_{\text{CB}}$. These findings indicate that single-arm studies – which the arm-based model can use – provide valuable information.

The above conclusions are based on using posterior means as Bayesian point estimates. We also considered posterior medians as point estimates, with results similar to those presented here.

6.3 Discussion

This chapter examined the sensitivity of arm-based network meta-analysis to treatment exclusion, and compared that to the sensitivity of the contrast-based approach. For the arm-based model, we investigated the fold changes of estimated population-averaged absolute risks and found that the arm-based model is fairly robust for most networks. Because the changes of estimated population-averaged absolute risks were mostly less than 1.05-fold, relative effect sizes based on the marginal absolute risks, such as the odds ratio or relative risk, would also have small changes. Although in general the changes were minor, removing specific treatments can be influential, as in, e.g., *Trikalinos 2009*. An influential treatment is typically investigated in many studies [66], while infrequently studied treatments are most likely to be affected by exclusion of other treatments to which they were compared. This suggests that when performing a network meta-analysis, researchers should be cautious if they only want to assess new treatments or if they want to exclude placebo arms or well-established treatments [66].

When comparing log odds ratio changes, the arm-based model generally outperformed the contrast-based model. Using bootstrap resampling, the difference between the arm-based and contrast-based models was statistically significant when single-arm studies were included in analyses using the arm-based model. However, when we dropped single-arm studies from reduced networks, the arm-based model performed almost the same as the contrast-based model. This implies that the arm-based model's greater robustness arises mainly from retaining single-arm studies. Some traditional pairwise meta-analyses have considered incorporating single-arm studies [142–145]; when single-arm studies are available for network meta-analysis, the arm-based model can be an attractive alternative approach.

One might wonder why the arm-based and contrast-based models did not give identical results when the arm-based model was restricted by excluding single-arm studies in Section 6.2.2. The reason is that the two models involve different random-effect assumptions. Specifically, Shuster et al. [146] described two types of assumptions about random effects in meta-analysis. The first type of random effects, called 'studies at random' (SR), assumes that the studies are independently chosen from a conceptual urn containing a large number of studies. The second type assumes that the relative effects in each study are randomly drawn from a conceptual urn while the studies are fixed; this is called 'effects at random' (ER), which makes assumptions over and above SR, namely that the distribution of the random relative effects is independent of the study design. Arguably, the arm-based model requires the SR assumption, while the contrast-based model requires ER.

Our study has several limitations. First, we did not check evidence consistency in the investigated networks; detecting inconsistency in network meta-analysis is still an open question, which is partly discussed by Lu and Ades [52]. For the contrast-based model, this study assumes that the pairwise comparisons among any trio of treatments, say A, B, and C, are inter-related as $\theta_{BC} = \theta_{AC} - \theta_{AB}$. If this consistency does not hold, we could use approaches based on an inconsistency model such as $\theta_{BC} = \theta_{AC} - \theta_{AB} + \phi$, which is discussed in Salanti et al. [53]. Here, ϕ represents the inconsistency between the direct evidence for treatment B vs. C and the indirect evidence from pairwise comparisons of A vs. B and A vs. C. For the arm-based model, one may consider detecting inconsistency between two treatments by comparing their absolute risk differences in direct comparisons vs. indirect comparisons [147]. A large discrepancy implies potential inconsistency between these two treatments. The second limitation of our study is that we used a selection criterion requiring each treatment to be studied in at least three studies, mainly due to the need for an adequate number of studies to estimate parameters for the distribution of random effects. The literature has no well-established criterion serving this purpose.

In conclusion, arm-based methods can be an attractive alternative when data from some single-arm studies are available. For example, if we are interested in comparing treatments A, B, and C in a network meta-analysis, 'single-arm' study data on A can come from two-arm studies comparing A vs. D or other treatments. Furthermore, although the arm-based model is generally more robust than the contrast-based model, for some network meta-analyses, the contrast-based methods seem to be more robust to some treatment exclusions. For example, the LOR changes under the arm-based model can be fairly large, while the corresponding changes under the contrast-based model can be nearly zero (Figure 6.2). Therefore, analysts are advised to consider both the arm-based and contrast-based models for network meta-analysis, especially when making inference for a small or poorly connected network.

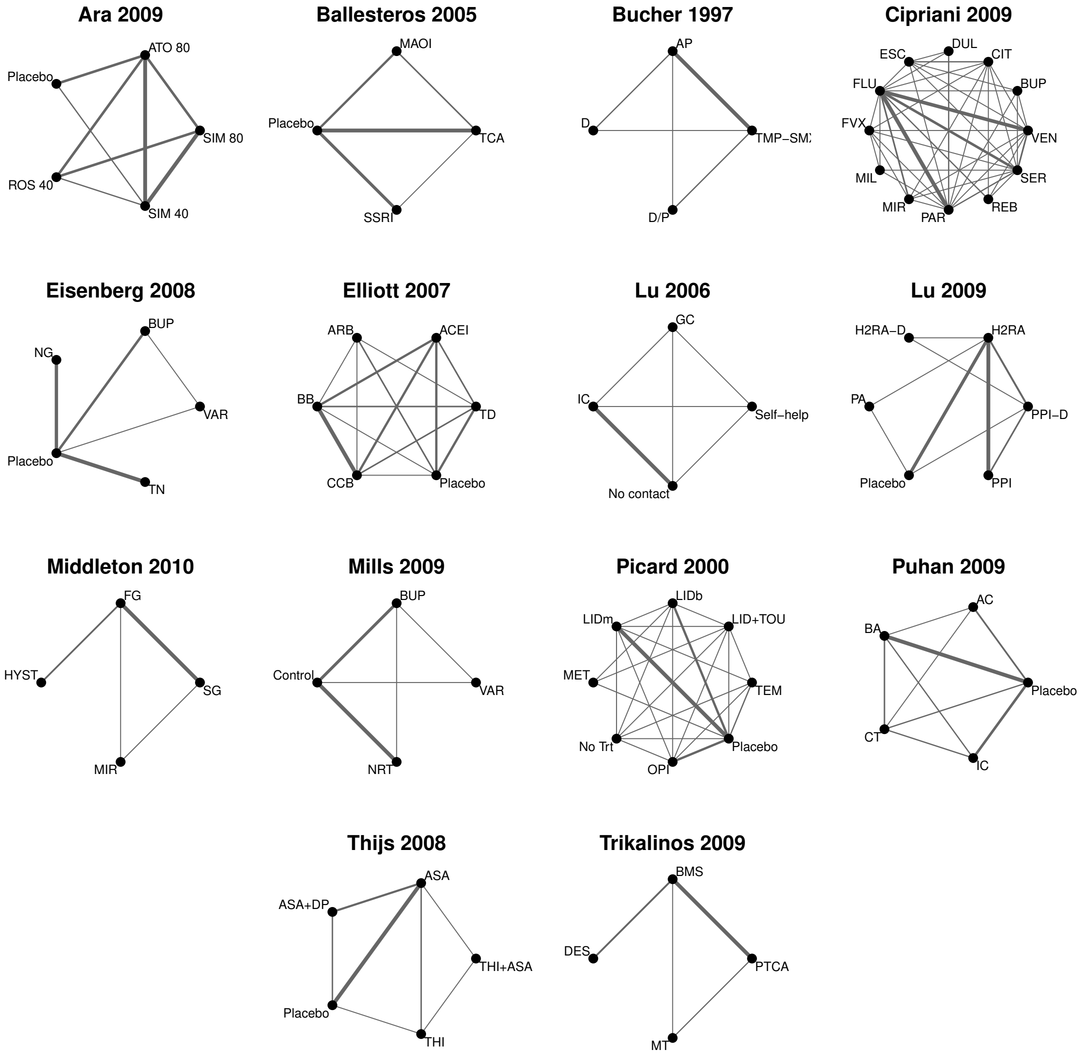


Figure 6.1: Network plots of the fourteen networks. A thicker edge indicates more comparisons between the treatments (nodes) of the edge.

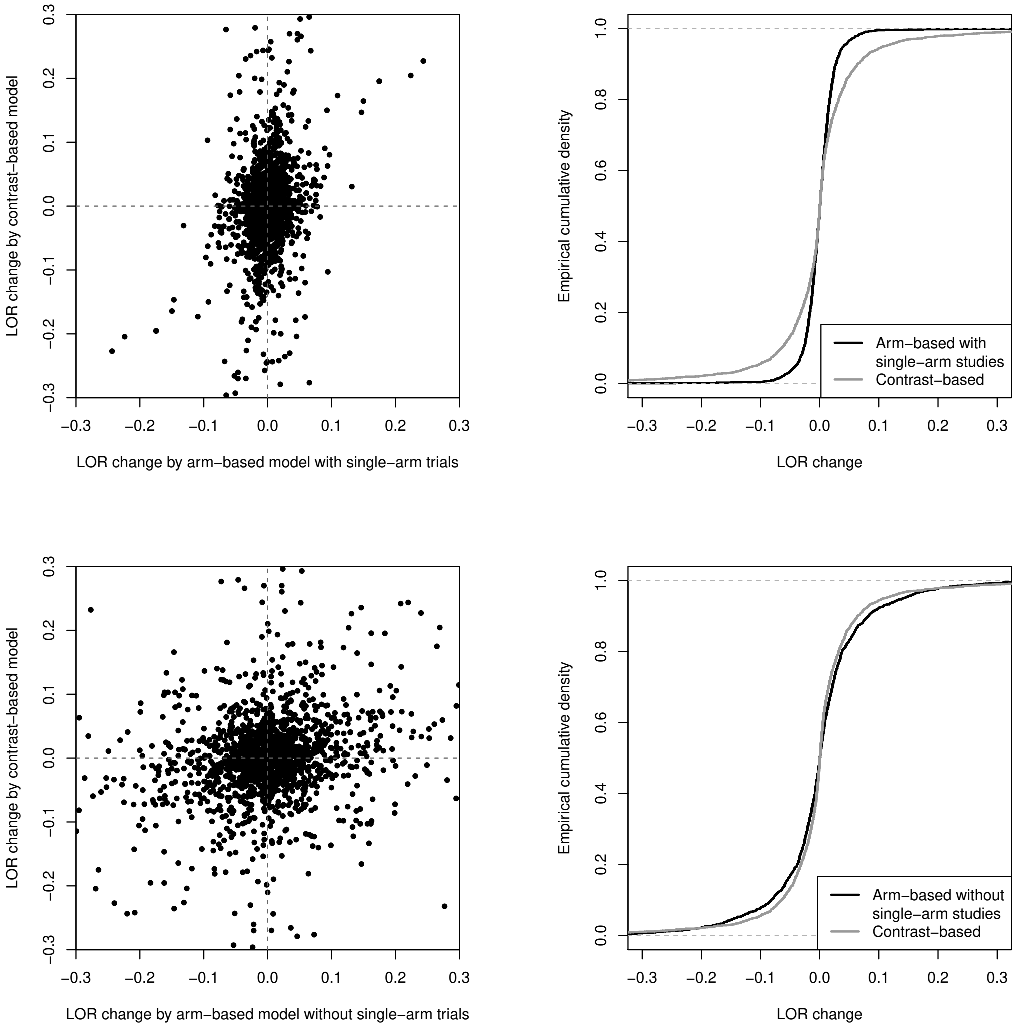


Figure 6.2: Comparing the arm-based and contrast-based models according to log odds ratio changes. In the upper two panels, single-arm studies are kept in the reduced networks for the arm-based model; in the lower panels, the arm-based model is applied only to studies that can also be used by the contrast-based model, i.e., single-arm studies are excluded. Left panels are scatter plots of log odds ratio changes under the contrast-based model (vertical axis) vs. those under the arm-based model (horizontal axis). Right panels show the empirical cumulative distribution function of log odds ratio changes under the two models.

Table 6.1: Characteristics of the fourteen network meta-analyses.

Network	Outcome	No. of studies	No. of treatments	Treatment names (abbreviations) [weighted degree], sorted by weighted degree (largest to smallest)
Ara 2009	Adverse event leading to drug discontinuation	11	5	Atorvastatin 80 mg/day (ATO 80) [9]; Simvastatin 40 mg/day (SIM 40) [8]; Simvastatin 80 mg/day (SIM 80) [7]; Rosuvastatin 40 mg/day (ROS 40) [5]; Placebo [3].
Ballesteros 2005	Efficacy of antidepressants in dysthymia	9	4	Placebo [12]; Tricyclic antidepressant (TCA) [8]; Monoamine oxidase inhibitor (MAOI) [5]; Selective serotonin reuptake inhibitor (SSRI) [5].
Bucher 1997	Number of Pneumocystis carinii pneumonia	18	4	Aerosolized pentamidine (AP) [14]; Trimethoprim-sulphamethoxazole (TMP-SMX) [13]; Dapsone/pyrimethamine (D/P) [5]; Dapsone (D) [4].
Cipriani 2009	Unipolar major depression	111	12	Fluoxetine (FLU) [54]; Paroxetine (PAR) [32]; Sertraline (SER) [28]; Venlafaxine (VEN) [27]; Escitalopram (ESC) [17]; Citalopram (CIT) [14]; Mirtazapine (MIR) [13]; Bupropion (BUP) [12]; Fluvoxamine (FVX) [11]; Duloxetine (DUL) [8]; Reboxetine (REB) [8]; Milnacipran (MIL) [6].
Eisenberg 2008	Smoking abstinence	61	5	Placebo [64]; Transdermal nicotine (TN) [23]; Nicotine gum (NG) [20]; Bupropion (BUP) [18]; Varenicline (VAR) [9].
Elliott 2007	The proportion of patients who developed diabetes	22 22	6 6	β blocker (BB) [12]; Calcium-channel blocker (CCB) [12]; Angiotensin-converting enzyme inhibitor (ACEI) [11]; Placebo [10]; Thiazide diuretic (TD) [10]; Angiotensin-receptor blocker (ARB) [5].
Lu 2006	Smoking cessation	24	4	Individual counselling (IC) [21]; No contact [20]; Group counselling (GC) [8]; Self-help [7].
Lu 2009	Gastroesophageal reflux disease	40	6	H_2 receptor antagonist (H2RA) [34]; Proton pump inhibitor (PPI) [17]; Placebo [14]; PPI double dose (PPI-D) [13]; Prokinetic agent (PA) [6]; H2RA double dose (H2RA-D) [4].
Middleton 2010	Patients' dissatisfaction	20	4	'First generation' endometrial destruction techniques (FG) [17]; 'Second generation' endometrial destruction techniques (SG) [14]; Hysterectomy (HYST) [5]; Mirena (MIR) [4].
Mills 2009	Smoking abstinence at at-least 4 weeks post-target quit data	89	4	Control [92]; Nicotine replacement therapy (NRT) [49]; Bupropion (BUP) [39]; Varenicline (VAR) [10].
Picard 2000	Pain on injection with propofol	43	8	Placebo [48]; Lidocaine (mg) mixed with propofol 200 mg (LIDm) [26]; Lidocaine (mg) given before the injection of propofol (LIDb) [19]; No treatment (No Trt) [19]; Opioids (OPI) [19]; Lidocaine (mg) with tourniquet (LID+TOU) [13]; Temperature (TEM) [13]; Metoclopramide (MET) [7].
Puhan 2009	Exacerbation in patients with chronic obstructive pulmonary disease	34	5	Placebo [44]; Long-acting beta-agonists (BA) [33]; Inhaled corticosteroids (IC) [24]; Combined treatment with a long-acting beta-agonist and an inhaled corticosteroid (CT) [20]; Long-acting anticholinergics (AC) [11].
Thijs 2008	Efficacy of antiplatelet	23	5	Aspirin (ASA) [22]; Placebo [16]; Aspirin and dipyridamole (ASA+DP) [10]; Thienopyridines (ticlopidin or clopidogrel, THI) [7]; THI+ASA [3].
Trikalinos 2009	Non-acute coronary artery disease	62	4	Bare-metal stents (BMS) [52]; Percutaneous transluminal balloon coronary angioplasty (PTCA) [43]; Drug-eluting stents (DES) [16]; Medical therapy (MT) [13].

Table 6.2: (Continued) Characteristics of the fourteen network meta-analyses.

Network	Total no. of participants	Total no. of no. of events	Naïve absolute risk[†]	Total no. of treatment groups	Total no. of treatment groups with zero events	Ineligible treatment removal		Smallest weighted degree[‡]	Largest weighted degree
						Arm-based model	Contrast-based model		
Ara 2009	24,793	1155	0.047	24	2		SIM 40; ATO 80; SIM 80	3	9
Ballesteros 2005	1386	663	0.478	21	0		Placebo	5	12
Bucher 1997	3416	248	0.073	36	4		AP; TMP-SMX	4	14
Cipriani 2009	24,595	13,951	0.567	224	0			6	54
Eisenberg 2008	26,750	3908	0.146	125	0	Placebo	Placebo	9(9)[*]	64(23)[*]
Elliott 2007	154,176	10,962	0.071	48	0			5	12
Lu 2006	16,737	2072	0.124	50	2			7	21
Lu 2009	4626	2273	0.491	82	4		H2RA	4	34
Middleton 2010	2886	342	0.119	40	0	FG	FG; SG	4(4)[*]	17(14)[*]
Mills 2009	29,525	10,847	0.367	181	1		Control	10	92
Picard 2000	4495	2400	0.534	104	2			7	48
Puhan 2009	26,789	7200	0.269	81	1			11	44
Thijs 2008	42,666	6830	0.160	49	0		ASA; THI	3	22
Trikalinos 2009	26,521	821	0.031	124	22	BMS	BMS	13(13)[*]	52(43)[*]

[†] Naïve absolute risk is calculated as the ratio of the total no. of Events compared to the total no. of Participants.

[‡] Weighted degree of a node (treatment) is the sum of weights (the number of pairwise comparisons between two treatments) on all edges incident to that node.

[*] In each of these three networks, one particular treatment is not removed to remain network connectivity; the numbers in parentheses are given without accounting for these treatments.

Table 6.3: Summary of fold changes of estimated population-averaged absolute risks using the arm-based model.

Network	Fold change		Removed treatment (weighted degree[§]) causing maximal fold change [Rank[†]/No. of eligible treatment removals]	Maximally affected treatment (weighted degree[§]) by the removal [Rank[‡]/No. of treatments]
	Average	Maximal		
Ara 2009	1.030	1.087	ATO 80 (9) [1/5]	ROS 40 (5) [2/5]
Ballesteros 2005	1.003	1.007	MAOI (5) [3/4]	Placebo (12) [4/4]
Bucher 1997	1.019	1.058	TMP-SMX (13) [2/4]	D (4) [1/4]
Cipriani 2009	1.005	1.033	SER (28) [3/12]	MIL (6) [1/12]
Eisenberg 2008	1.003	1.008	VAR (9) [4/4*]	BUP (18) [2/5]
Elliott 2007	1.015	1.056	Placebo (10) [4/6]	ACEI (11) [4/6]
Lu 2006	1.012	1.028	No contact (20) [2/4]	Self-help (7) [1/4]
Lu 2009	1.006	1.036	H2RA (34) [1/6]	Placebo (14) [4/6]
Middleton 2010	1.013	1.037	SG (14) [1/3*]	FG (17) [4/4]
Mills 2009	1.011	1.045	Control (92) [1/4]	VAR (10) [1/4]
Picard 2000	1.009	1.050	LIDb (19) [3/8]	MET (7) [1/8]
Puhan 2009	1.017	1.055	Placebo (44) [1/5]	BA (33) [4/5]
Thijs 2008	1.019	1.084	ASA+DP (10) [3/5]	THI+ASA (3) [1/5]
Trikalinos 2009	1.055	1.390	PTCA (43) [1/3*]	MT (13) [1/4]

[†] Rank from largest to smallest according to the weighted degrees within the corresponding network.

[‡] Rank from smallest to largest according to the weighted degrees within the corresponding network.

[§] The weighted degrees refer to the corresponding full network.

[*] In each of these three networks, one particular treatment is not removed to remain network connectivity (See Table 6.2).

Chapter 7

On Network Meta-Analysis Without Evidence Cycles

Although a variety of methods are available for performing network meta-analysis [67, 148–151], currently the most widely used approach is the Bayesian hierarchical model proposed by Lu and Ades [51], which this chapter calls the Lu–Ades model. In a recent survey by Nikolakopoulou et al. [152], 111 out of 186 network meta-analyses used the Lu–Ades model.

By combining information from both direct and indirect comparisons, network meta-analysis is generally considered more powerful than conventional pairwise meta-analysis, which compares each pair of treatments separately and thus can only use direct comparisons [50]. In network meta-analysis, treatments A and B can be compared via a common comparator, say treatment C, and the information from A vs. C and B vs. C provides indirect evidence, while a trial including both A and C provides direct evidence. Hence, provided that the studies are consistent with each other, network meta-analysis may be expected to produce more accurate effect estimates with narrower confidence/credible intervals, compared to pairwise meta-analysis; also, it can provide a coherent ranking of treatments and thus guide decision making [153]. Because of these attractive features, many researchers focus on collecting as many treatments as possible in a network meta-analysis, but pay little attention to the network's geometry, taking for granted the benefit from synthesizing direct and indirect evidence.

If each pair of treatments A, B, and C is directly compared in at least one study, then the three treatments form a so-called evidence cycle [52]. This chapter shows that evidence cycles in the treatment network play a critical role in the improvement of effect estimates produced by Lu–Ades network meta-analysis compared to separate pairwise meta-analyses. Specifically, Lu–Ades network meta-analysis yields posterior distributions identical to separate pairwise meta-analyses for all treatment comparisons when a treatment network does not contain any evidence cycles. Networks without evidence cycles frequently appear in systematic reviews. A special case is the star-shaped network, that is, all collected studies share a common treatment, which is usually placebo or a well-established standard treatment. For example, among the 186 network meta-analyses investigated by Nikolakopoulou et al. [152], 35 networks are star-shaped. We also extend our conclusion to networks with general shapes, which are common in real applications: treatment comparisons that are not in any evidence cycles cannot benefit from Lu–Ades network meta-analysis. Instead of discouraging researchers from performing network meta-analysis, we seek to raise awareness of the power of network meta-analysis compared to pairwise meta-analysis when using the Lu–Ades model in certain situations.

The remaining of this chapter is organized as follows. After reviewing the development of the Lu–Ades model in Section 7.1, Section 7.2 shows theoretically that the joint posterior distributions of effect estimates produced by Lu–Ades and by separate pairwise meta-analyses are identical for networks without evidence cycles. The proofs are in Appendix B.3 unless given in the main text. Simulations and a case study to illustrate the equivalence relationship are presented in Section 7.3, and Section 7.4 concludes with some discussion.

7.1 Methods for general network meta-analysis: a review

7.1.1 Smith model for pairwise meta-analysis

First, we introduce the Bayesian hierarchical model for the conventional pairwise meta-analysis proposed by Smith et al. [154], which lays the foundation for Lu–Ades network meta-analysis of multiple treatment comparisons. Suppose that a pairwise meta-analysis collects N studies and each study compares the same two treatments, such as an active treatment and a control. Let y_{i1} and y_{i2} be the observed aggregated outcome measures in study i's treatment groups 1 and 2, respectively. The overall relative effect comparing the two treatments is usually of interest. The Smith random-effects model can be generalized as follows to estimate the overall relative effect [154]:

$$\begin{aligned} &y_{ik} \sim f(y \mid \Delta_{ik}, \xi_{ik}), \quad i = 1, \ldots, N, k = 1, 2; \\ &g(\Delta_{i1}) = \mu_i, \quad g(\Delta_{i2}) = \mu_i + \delta_i; \\ &\delta_i \sim N(d, \sigma^2). \end{aligned} \tag{7.1}$$

Here, μ_i is commonly called the baseline effect of study i, and the study-specific relative effects δ_i are assumed to be exchangeable across studies with mean d, which is interpreted as the overall relative effect. The variance parameter σ^2 reflects heterogeneity between studies. The link function is $g(\cdot)$, and $f(\cdot \mid \cdot, \cdot)$ is the outcome measure's density function, depending on an unknown location parameter Δ_{ik} and a nuisance parameter ξ_{ik}, which is assumed to be known. For example, if the outcome is continuous, y_{ik} is usually assumed to be normally distributed with unknown mean Δ_{ik} and known standard error ξ_{ik}, and $g(\cdot)$ is the identity link. If the outcome is binary, such as the condition of having a certain event, then y_{ik} is the number of events, which follows a binomial density with unknown event rate Δ_{ik} and known sample size ξ_{ik}. When the logit link function $\text{logit}(t) = \log\{t/(1-t)\}$ is used for binary outcomes, the fixed effect d represents the overall log odds ratio of treatment 2 compared to treatment 1.

7.1.2 Lu–Ades model for network meta-analysis

Lu and Ades [51, 54] extended the Smith model to multiple treatment comparisons. Instead of comparing merely two treatments, N studies are included comparing a total of K treatments in a network meta-analysis ($K > 2$). Specifically, each study compares a subset of the K treatments; denote the treatment subset of study i as $\mathcal{T}_i$. A study is called a two-arm study if it compares two treatments, while a multi-arm study investigates more than two treatments. Again, assume that the observed aggregated outcome measure y_{ik} in study i's treatment group k follows the distribution $f(\cdot \mid \Delta_{ik}, \xi_{ik})$. To use the Lu–Ades model, a baseline treatment b_i needs to be specified for each study i. Different studies can have different baseline treatments in the Lu–Ades model because the treatment subsets $\mathcal{T}_i$ need not intersect. We denote b_i simply as b when it does not lead to confusion. The Lu–Ades random-effects model for network meta-analysis is specified as follows:

$$\begin{aligned}
&y_{ik} \sim f(y \mid \Delta_{ik}, \xi_{ik}), \quad i = 1, \ldots, N, k \in \mathcal{T}_i; \\
&g(\Delta_{ik}) = \mu_i + X_{ik}\delta_{ibk}; \\
&\delta_{ibk} \sim N(d_{bk}, \sigma^2_{bk}), \quad \mathrm{Corr}(\delta_{ibh}, \delta_{ibk}) = \gamma_{bhk}, \quad h, k \in \mathcal{T}_i.
\end{aligned} \tag{7.2}$$

Here, X_{ik} is a dummy variable; $X_{ik} = 0$ if $k = b$ and $X_{ik} = 1$ if $k \in \mathcal{T}_i \backslash \{b\}$. Within a multi-arm study, the correlation between the treatment contrasts δ_{ibh} and δ_{ibk} is assumed to be γ_{bhk}. Again, μ_i represents the baseline effect of study i, the study-specific relative effects are assumed to be exchangeable, and we focus on estimating the relative effects of all treatment contrasts d_{hk} $(1 \leq h \neq k \leq K)$.

A critical assumption in Lu–Ades network meta-analysis is the consistency equation for an evidence cycle, which relates the contrasts for a trio of treatments as

$$d_{hk} = d_{\ell k} - d_{\ell h}, \quad \text{for all } 1 \leq h \neq k \neq \ell \leq K. \tag{7.3}$$

If a treatment network contains evidence cycles, this equation synthesizes both direct and indirect evidence for the treatment comparisons in the cycles, so that the network meta-analysis uses more information than a conventional pairwise meta-analysis, which uses only direct evidence.

The consistency assumption may not hold even approximately in many cases, and alternative approaches have been proposed to deal with evidence inconsistency; see, e.g., [52, 53, 58, 155–157]. A popular method is to add inconsistency factors w to Equation (7.3), that is, $d_{hk} = d_{\ell k} - d_{\ell h} + w_{hk\ell}$. This method is closely related to the number of independent cycles in the network, which is quantified by the inconsistency degrees of freedom df_{IC} [52]. If all studies are two-armed, then $df_{\mathrm{IC}} = T - K + 1$, where T is the number of all treatment comparisons, i.e., the edges in the network. However, when multi-arm studies are present, the definition of inconsistency degrees of freedom is fairly complex and needs to be considered case by case.

Besides random-effects models, fixed-effects models are also frequently used in meta-analysis. These models assume that the collected studies are *homogeneous*, that is, that the relative effects for each treatment comparison share a common mean across studies, and their variation is entirely due to sampling error within studies. To be specific, the

Smith fixed-effects model for pairwise meta-analysis is

$$\begin{aligned} &y_{ik} \sim f(y \mid \Delta_{ik}, \xi_{ik}), \quad i = 1, \dots, N, k = 1, 2; \\ &g(\Delta_{i1}) = \mu_i, \quad g(\Delta_{i2}) = \mu_i + d, \end{aligned} \tag{7.4}$$

while the Lu–Ades fixed-effects model for network meta-analysis is

$$\begin{aligned} &y_{ik} \sim f(y \mid \Delta_{ik}, \xi_{ik}), \quad i = 1, \dots, N, k \in \mathcal{T}_i; \\ &g(\Delta_{ik}) = \mu_i + X_{ik} d_{bk}. \end{aligned} \tag{7.5}$$

Implementation is easier for the fixed-effects model than the random-effects model because the latter involves complex specification of heterogeneity variances, which will be detailed in Section 7.2.3. However, the homogeneity assumption may be unrealistic in many cases [78], and the credible intervals produced by the fixed-effects model may have low coverage probabilities if heterogeneity is present in some treatment comparisons [158].

7.2 Network meta-analysis without evidence cycles

7.2.1 Direct and indirect evidence

The treatment network is assumed to be connected throughout this chapter; if the network consists of several disjoint sub-networks, then a separate analysis can be applied to each sub-network. For a treatment network without cycles, all collected studies must be two-armed because multi-arm studies create evidence cycles. Consequently, we no longer need to account for the correlations between treatment contrasts within studies in the Lu–Ades random-effects model (7.2).

To investigate the performance of the Lu–Ades model for a network without cycles, we explore the posterior distributions of all treatment contrasts. The $(K-1)K/2$ treatment contrasts are denoted as a vector $\boldsymbol{e} = (d_{hk}; 1 \leq h < k \leq K)^{\mathrm{T}}$. In graph theory, a connected network without cycles is a spanning tree and contains exactly $K-1$ edges; denote the set of these edges as a $(K-1)$-dimensional vector $\boldsymbol{e}_{\mathrm{b}} = (e_1, \dots, e_{K-1})^{\mathrm{T}}$, where $e_j = d_{hk}$ for some $h < k$ and each e_j provides direct evidence. Thus, the set of all treatment contrasts $\boldsymbol{e}$ can be split into two subsets: $\boldsymbol{e}_{\mathrm{b}}$, each contrast in which is directly compared in the network, and a $(K-2)(K-1)/2$-dimensional vector $\boldsymbol{e}_{\mathrm{f}} = (d_{hk}; d_{hk} \notin \boldsymbol{e}_{\mathrm{b}})^{\mathrm{T}}$ that can only be imputed from indirect evidence. By the definition of Lu and Ades [52], the treatment contrasts in $\boldsymbol{e}_{\mathrm{b}}$ are basic parameters, which involve all K treatments but do not form cycles; those in $\boldsymbol{e}_{\mathrm{f}}$ are referred to as functional parameters because they can be represented as functions of the basic parameters. The evidence consistency equation (7.3) necessarily holds for networks without cycles because evidence inconsistency only occurs within evidence cycles; indeed, these networks have zero inconsistency degrees of freedom. Therefore, $\boldsymbol{e}_{\mathrm{f}}$ is entirely determined by $\boldsymbol{e}_{\mathrm{b}}$; that is, we may write $\boldsymbol{e}_{\mathrm{f}} = \mathbf{A}\boldsymbol{e}_{\mathrm{b}}$, where $\mathbf{A}$ is a known $(K-2)(K-1)/2 \times (K-1)$ transformation matrix. We have the following proposition regarding the transformation matrix $\mathbf{A}$.

Proposition 6. *The transformation matrix* $\mathbf{A}$ *is unique for each set of basic parameters, and each entry of* $\mathbf{A}$ *is 0 or* ± 1*.*

Proposition 6 holds for any type of connected network, including those containing cycles, under the assumption of evidence consistency. In networks without cycles, there is only one set of basic parameters $\boldsymbol{e}_{\mathrm{b}}$, so the transformation matrix $\mathbf{A}$ is uniquely defined.

7.2.2 Equivalence of the Lu–Ades model and separate Smith models

In a network without cycles, suppose that study i, which must be two-armed, compares treatments k_i vs. h_i ($h_i < k_i$); that is, the corresponding treatment contrast is $d_{h_i k_i}$. For $j = 1, \ldots, K-1$, let $\mathcal{S}_j = \{i : d_{h_i k_i} = e_j\}$ be the set of studies that give the direct treatment comparison e_j. Consequently, the N studies $\mathcal{S} = \{1, \ldots, N\}$ in the network can be partitioned into $K-1$ subsets according to their treatment contrasts: $\mathcal{S} = \bigcup_{j=1}^{K-1} \mathcal{S}_j$. Moreover, let $\mathcal{D}_j = \{(y_{ik}, \xi_{ik}); i \in \mathcal{S}_j, k \in \mathcal{T}_i\}$ be the data (aggregated outcome measures and nuisance parameters) provided by the studies in $\mathcal{S}_j$, and let $\mathcal{D} = \bigcup_{j=1}^{K-1} \mathcal{D}_j$ be the full data in the whole network. The Smith model for pairwise meta-analysis uses the data $\mathcal{D}_j$ for each j separately to estimate the corresponding treatment contrast e_j, and we denote the resulting posterior distribution as $p(e_j \mid \mathcal{D}_j)$. The Lu–Ades model for network meta-analysis uses the full data $\mathcal{D}$ to simultaneously compare all treatments, and we denote the joint posterior distribution of the direct treatment contrasts as $p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D})$. We have the following theorem.

Theorem 1. *For a treatment network without evidence cycles, given the same set of priors, the Lu–Ades fixed-effects model (7.5) gives posterior distributions of direct treatment contrasts identical to those from separate Smith fixed-effects model (7.4), that is,*

$$p(\boldsymbol{e}_b \mid \mathcal{D}) = \prod_{j=1}^{K-1} p(e_j \mid \mathcal{D}_j). \tag{7.6}$$

This equation also holds for the Smith and Lu–Ades random-effects models (7.1) and (7.2), if the Lu–Ades model uses different heterogeneity variances for different treatment contrasts.

Equation (7.6) implies that the posterior estimate of e_j produced by the Lu–Ades model is only informed by the data in studies $\mathcal{S}_j$; thus, the posterior distributions of the e_j's are mutually independent.

Proof of Theorem 1. In the Smith and Lu–Ades fixed-effects models, we denote $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^{\mathrm{T}}$ as the vector of all studies' baseline effects, and let $\widetilde{\boldsymbol{\mu}}_j = (\mu_i; i \in \mathcal{S}_j)^{\mathrm{T}}$ be the vector of those baseline effects in studies $\mathcal{S}_j$. For $j = 1, \ldots, K-1$, denote $\boldsymbol{y}_j = (y_{ik}; i \in \mathcal{S}_j, k \in \mathcal{T}_i)^{\mathrm{T}}$ and $\boldsymbol{\xi}_j = (\xi_{ik}; i \in \mathcal{S}_j, k \in \mathcal{T}_i)^{\mathrm{T}}$. Also, let $\boldsymbol{y} = (y_{ik}; i \in \mathcal{S}, k \in \mathcal{T}_i)^{\mathrm{T}}$ and $\boldsymbol{\xi} = (\xi_{ik}; i \in \mathcal{S}, k \in \mathcal{T}_i)^{\mathrm{T}}$; thus, $\mathcal{D} = \{(\boldsymbol{y}, \boldsymbol{\xi})\}$ and $\mathcal{D}_j = \{(\boldsymbol{y}_j, \boldsymbol{\xi}_j)\}$. By the properties of conditional probability, the joint posterior of the direct treatment contrasts produced by the Lu–Ades fixed-effects model is

$$p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) = \int p(\boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\mu} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\mu} \propto \int f(\boldsymbol{y} \mid \boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\mu}, \boldsymbol{\xi}) p(\boldsymbol{e}_{\mathrm{b}}) p(\boldsymbol{\mu})\, \mathrm{d}\boldsymbol{\mu}.$$

Here, $f(\cdot \mid \cdot)$ is the probability density function of the observed outcome measures conditional on the pertinent parameters, and $p(\boldsymbol{e}_{\mathrm{b}}) = \prod_{j=1}^{K-1} p(e_j)$ and $p(\boldsymbol{\mu}) = \prod_{j=1}^{K-1} p(\widetilde{\boldsymbol{\mu}}_j)$ are priors for the treatment contrasts and baseline effects, respectively. Since conditional on $\boldsymbol{\mu}$ and $\boldsymbol{e}_{\mathrm{b}}$, the outcome measure y_j in studies $\mathcal{S}_j$ depends on $\widetilde{\boldsymbol{\mu}}_j$ and e_j but not the other basic parameters in $\boldsymbol{e}_{\mathrm{b}}$, we have $f(\boldsymbol{y} \mid \boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\mu}, \boldsymbol{\xi}) = \prod_{j=1}^{K-1} f(\boldsymbol{y}_j \mid e_j, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j)$. Consequently,

$$p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) \propto \prod_{j=1}^{K-1} \int f(\boldsymbol{y}_j \mid e_j, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\widetilde{\boldsymbol{\mu}}_j)\, \mathrm{d}\widetilde{\boldsymbol{\mu}}_j \propto \prod_{j=1}^{K-1} p(e_j \mid \mathcal{D}_j).$$

In the random-effects models, we further denote σ_j as the heterogeneity standard deviation of the treatment contrast e_j. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{K-1})^{\mathrm{T}}$. Similarly, we have

$$p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) = \iint p(\boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\sigma}, \boldsymbol{\mu} \mid \mathcal{D}) \, \mathrm{d}\boldsymbol{\sigma} \, \mathrm{d}\boldsymbol{\mu} \propto \iint f(\boldsymbol{y} \mid \boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{\xi}) p(\boldsymbol{e}_{\mathrm{b}}) p(\boldsymbol{\sigma}) p(\boldsymbol{\mu}) \, \mathrm{d}\boldsymbol{\sigma} \, \mathrm{d}\boldsymbol{\mu},$$

where $p(\boldsymbol{\sigma}) = \prod_{j=1}^{K-1} p(\sigma_j)$ is the prior for the heterogeneity standard deviations. Again, because $f(\boldsymbol{y} \mid \boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{\xi}) = \prod_{j=1}^{K-1} f(\boldsymbol{y}_j \mid e_j, \sigma_j, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j)$, the joint posterior distribution of direct treatment contrasts produced by the Lu–Ades random-effects model is

$$p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) = \prod_{j=1}^{K-1} \iint f(\boldsymbol{y}_j \mid e_j, \sigma_j, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\sigma_j) p(\widetilde{\boldsymbol{\mu}}_j) \, \mathrm{d}\sigma_j \, \mathrm{d}\widetilde{\boldsymbol{\mu}}_j \propto \prod_{j=1}^{K-1} p(e_j \mid \mathcal{D}_j).$$

This completes the proof. □

Unlike the e_j's in $\boldsymbol{e}_{\mathrm{b}}$ that are directly compared in the network, the estimates of $\boldsymbol{e}_{\mathrm{f}}$ are entirely informed by indirect evidence. The network meta-analysis seems to be an efficient approach to simultaneously estimating all treatment contrasts, including the indirect ones. However, the following theorem shows that separate Smith models also produce posterior distributions of indirect treatment contrasts identical to those given by the Lu–Ades model.

Theorem 2. *Under the model settings in Theorem 1 and using the evidence consistency equation (7.3), the joint posterior distributions of the indirect treatment contrasts $\boldsymbol{e}_f$ produced by the Lu–Ades model and by separate Smith models are identical for a network without evidence cycles. Specifically, under some regularity assumptions given in Appendix B.3, the joint posterior distribution of $\boldsymbol{e}_f$ is*

$$p(\boldsymbol{e}_f \mid \mathcal{D}) = \frac{1}{(2\pi)^P} \int_{\mathbb{R}^P} e^{-i\boldsymbol{t}^{\mathrm{T}}\boldsymbol{e}_f} \varphi_f(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t},$$

where $P = (K-2)(K-1)/2$ and $i^2 = -1$. The characteristic function of $\boldsymbol{e}_f$ is $\varphi_f(\boldsymbol{t}) = \varphi_b(\mathbf{A}^{\mathrm{T}}\boldsymbol{t})$ for $\boldsymbol{t} \in \mathbb{R}^P$, where $\varphi_b(\boldsymbol{s}) = \prod_{j=1}^{K-1} \int_{\mathbb{R}} e^{is_j e_j} p(e_j \mid \mathcal{D}_j) \, \mathrm{d}e_j$ is the characteristic function of the direct treatment contrasts $\boldsymbol{e}_b$ for $\boldsymbol{s} = (s_1, \ldots, s_{K-1})^{\mathrm{T}} \in \mathbb{R}^{K-1}$.

Theorems 1 and 2 imply that for a network without cycles Lu–Ades network meta-analysis does not change the posterior distributions (thus, point estimates and credible intervals) of any treatment contrasts produced by separate Smith pairwise meta-analyses.

7.2.3 Lu–Ades random-effects model with equal heterogeneity variances

Besides potential evidence inconsistency, modeling the heterogeneity variances and covariances is another important issue in the Lu–Ades random-effects model (7.2). The difficulty arises from the fundamental relationship of the relative effects, $\delta_{ihk} = \delta_{i\ell k} - \delta_{i\ell h}$, so the heterogeneity standard deviations are constrained by the triangular inequality

$$|\sigma_{\ell h} - \sigma_{\ell k}| \leq \sigma_{hk} \leq |\sigma_{\ell h} + \sigma_{\ell k}|. \tag{7.7}$$

Lu and Ades [54] introduced a reparameterization of the σ_{hk}'s that permits specification of unstructured variance and correlation components for network meta-analysis. However, for conceptual and technical simplicity, the heterogeneity variances σ^2_{bk} are often assumed to be equal to a common variance σ^2 and the between-contrast correlations γ_{bkl} are set to 1/2 [50, 51]. This assumption is widely used in applications (e.g., [61, 159, 160]), though it imposes a possibly quite strong constraint on the treatment comparisons, which may be unrealistic for many cases [54].

Under the assumption of equal heterogeneity variances, we have the following theorem.

Theorem 3. *For a treatment network without evidence cycles, the Lu–Ades random-effects model (7.2) with equal heterogeneity variances $\sigma^2_{bk} = \sigma^2$ is equivalent to simultaneously using the Smith random-effects models (7.1) for studies $\mathcal{S}_j$, conditional on the common heterogeneity variance σ^2.*

The Smith models in Theorem 3 may not be deemed separate, because each model uses the common heterogeneity variance σ^2, which is informed by all studies $\mathcal{S}$ instead of the study set $\mathcal{S}_j$ for a specific treatment contrast. Although the Lu–Ades random-effects model can therefore produce different results from separate Smith random-effects models that have no constraints on the heterogeneity variances σ^2_{bk}, Theorem 3 implies that these differences are caused entirely by the specification of heterogeneity variances if the treatment network does not contain evidence cycles. Thus, under this model setting, the Lu–Ades model still provides no gain from synthesizing direct and indirect evidence apart from the strong assumption that $\sigma^2_{bk} = \sigma^2$.

7.2.4 Acyclic treatment comparisons in general networks

In a general treatment network that may contain evidence cycles, it commonly occurs that some treatment comparisons are not in any cycles [161]; we refer to such treatment comparisons as acyclic comparisons. Theorem 1 can be extended to the posterior distributions of acyclic comparisons in networks with general shapes. Specifically, suppose that a network with K treatments contains J acyclic comparisons, denoted as $\boldsymbol{e}_{\mathrm{a}} = (e_1, \ldots, e_J)^{\mathrm{T}}$.

Proposition 7. *For a network with K treatments, the number of acyclic comparisons J does not exceed $K - 1$.*

Studies that report the acyclic comparison e_j $(j = 1, \ldots, J)$ must be two-armed; otherwise, multi-arm studies create evidence cycles containing e_j, contradicting the definition of an acyclic comparison. As in Section 7.2.2, let $\mathcal{S}_j$ be the set of studies that report the acyclic comparison e_j, and $\mathcal{S}^\star = \mathcal{S} \backslash \bigcup_{j=1}^{J} \mathcal{S}_j$ be the remaining studies in the network. The studies in $\mathcal{S}^\star$ produce a sub-network that does not have any acyclic comparisons and thus must contain evidence cycles if the set $\mathcal{S}^\star$ is not empty. Suppose that $\boldsymbol{e}_{\mathrm{b}}^\star$ is a set of basic parameters in the sub-network consisting of $\mathcal{S}^\star$; then $\boldsymbol{e}_{\mathrm{b}} = (\boldsymbol{e}_{\mathrm{a}}^{\mathrm{T}}, \boldsymbol{e}_{\mathrm{b}}^{\star\mathrm{T}})^{\mathrm{T}}$ is a set of basic parameters for the full network $\mathcal{S}$. Also, denote the data provided by $\mathcal{S}_j$ as $\mathcal{D}_j$ and the data provided by $\mathcal{S}^\star$ as $\mathcal{D}^\star$. Then we have the following theorem.

Theorem 4. *For acyclic treatment comparisons in a general network, the Lu–Ades model does not improve their posterior distributions compared to separate Smith models under the model settings in Theorem 1. Specifically, using the same set of priors in the two models, the joint posterior distribution of the basic parameters produced by the Lu–Ades model is*

$$p(\boldsymbol{e}_b \mid \mathcal{D}) = p(\boldsymbol{e}_b^\star \mid \mathcal{D}^\star) \prod_{j=1}^{J} p(e_j \mid \mathcal{D}_j). \tag{7.8}$$

Here, $p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D})$ is produced by the Lu–Ades network meta-analysis on the full network $\mathcal{S}$, while $p(\boldsymbol{e}_{\mathrm{b}}^\star \mid \mathcal{D}^\star)$ is the posterior based on the sub-network consisting of $\mathcal{S}^\star$. The study set $\mathcal{S}^\star$ does not exist in a network without cycles so that $p(\boldsymbol{e}_{\mathrm{b}}^\star \mid \mathcal{D}^\star)$ drops out of Equation (7.8), which is thus reduced to Equation (7.6). Theorem 4 can therefore be viewed as a generalization of Theorem 1.

Since the acyclic comparisons $\boldsymbol{e}_{\mathrm{a}}$ are not contained in any evidence cycles, they are not subject to the risk of evidence inconsistency. The sub-network consisting of $\mathcal{S}^{\star}$ contains evidence cycles, so the evidence may be inconsistent; however, Theorem 4 still applies for this situation.

7.3 Numerical studies

7.3.1 Simulations

We conducted simulations to illustrate the equivalence of the Lu–Ades and Smith models' performance when the treatment network does not contain any cycles. The outcome was assumed to be continuous and normally distributed, and each treatment's outcome measure y_{ik} and its within-study standard error ξ_{ik} were observed. The situation of a binary outcome will be explored in the real data analysis in Section 7.3.2. We simulated data containing five treatments with three network shapes, shown in Figure 7.1. Each network does not contain cycles: Shape 1 is a star-shaped network with its center at treatment 1; Shape 2 is a chain-shaped network with treatment contrasts from 2 vs. 1 to 5 vs. 4; and Shape 3 is more general than the star and chain shapes. Also, in each network four treatment contrasts are observed and form a set of basic parameters $\boldsymbol{e}_{\mathrm{b}}$. These treatment contrasts are reported in 5, 10, 15, or 20 studies, as described in Figure 7.1. Thus, each simulated network contained a total of 50 studies.

To simulate the outcome measures, we first generated samples for all five treatments in each study, and then omitted certain treatment arms to create networks with the shapes in Figure 7.1; the omitted data were assumed to be missing completely at random. Specifically, the five treatments' within-study standard errors were drawn from $\xi_{ik} \sim U(0.1, 1)$ $(i = 1, \ldots, 50,\ k = 1, \ldots, 5)$. The observed treatment-specific outcome measure was generated from $y_{ik} \sim N(\mu_{ik}, \xi_{ik}^2)$, where μ_{ik} represents the underlying true measure of treatment k in study i. The study-specific true measures were drawn from $(\mu_{i1}, \ldots, \mu_{i5})^{\mathrm{T}} \sim N((\mu_1, \ldots, \mu_5)^{\mathrm{T}}, \boldsymbol{\Psi})$, where μ_k represents the overall mean of treatment k $(k = 1, \ldots, 5)$, and $\boldsymbol{\Psi}$ represents the between-study covariance matrix. We set $\mu_k = k$; hence, the true relative effect of treatments k vs. h was $d_{hk} = \mu_k - \mu_h = k - h$. Also, $\boldsymbol{\Psi} = \mathbf{DRD}$, where $\mathbf{R} = (\rho_{hk})$ is the correlation matrix with $\rho_{kk} = 1$ and $\rho_{hk} = 0.4$ $(1 \leq h \neq k \leq 5)$, and the between-study standard deviations

$\mathbf{D} = \text{diag}(\tau_1, \ldots, \tau_5)$ were sampled for three cases: (i) all studies were homogeneous with $\tau_k = 0$; (ii) all treatments had a common heterogeneity standard deviation $\tau_k = \tau$ with $\tau \sim U(1, 1.5)$; and (iii) the five treatments had different heterogeneity standard deviations with $\tau_k \sim U(0.4k - 0.4, 0.4k)$ for $k = 1, \ldots, 5$. Finally, certain treatments in certain studies were randomly omitted to produce networks with Shapes 1–3. For example, in the network with Shape 1, treatments 3–5 were omitted in five studies, so these five studies compared treatments 2 vs. 1. For each network shape, 1000 replicates of network data were generated; for each replicate, the Markov chain Monte Carlo algorithm was applied to implement the Smith and Lu–Ades models using one chain, which contained a run of 50,000 updates after a 20,000-run burn-in period. For both the Smith and Lu–Ades models, three model settings were considered: a fixed-effects model, a random-effects model with different heterogeneity variances, and a random-effects model with a common heterogeneity variance. Vague priors were used for the study-specific baseline effects and the basic parameters; $U(0, 10)$ priors were used for the heterogeneity standard deviations in the random-effects models. The functional parameters, such as d_{23} in network with Shape 1, were estimated using the evidence consistency equation (7.3). The models' performance was evaluated according to bias and mean squared error of the estimated relative effects and coverage probability of the 95% credible intervals.

Table 7.1 presents the results of some treatment contrasts for Case (iii) of the between-study standard deviation; the simulation results for Cases (i) and (ii) are in Appendix A.10. Since the treatments were missing completely at random in all cases, each model produced nearly unbiased point estimates for each treatment contrast. In Case (i), where the treatment effects were homogeneous across studies, using either pairwise or network meta-analysis for all three networks in Figure 7.1, both the fixed- and random-effects models produced estimated relative effects with similar mean squared errors. Also, the fixed-effects model led to credible interval coverage probabilities that are fairly close to the nominal level 95%, while the two random-effects models produced slightly inflated coverage probabilities, indicating that their 95% credible intervals were wider than the fixed-effects model. However, in Cases (ii) and (iii), due to the high heterogeneity, the fixed-effects model led to very poor credible interval coverage probabilities, while those produced by the random-effects models were generally satisfactory;

the mean squared errors produced by the fixed-effects model were also much larger than those of the random-effects models. Moreover, in Case (iii), the true heterogeneity variances τ_k^2 differed across treatments, while the second random-effects model incorrectly assumed the τ_k^2's were equal. Interestingly, the results produced by this random-effects model were fairly similar to those produced by the correct random-effects model assuming different heterogeneity variances, although the incorrect model had slightly low credible interval coverage for the treatment contrast d_{15} in network with Shape 1 and d_{45} in networks with Shapes 2 and 3. Most importantly, for all three network shapes, the Smith model for pairwise meta-analysis produced effect estimates with biases, mean squared errors, and credible interval coverage probabilities almost identical to those produced by the Lu–Ades model for network meta-analysis; some slight differences are due to Monte Carlo error. Therefore, the Lu–Ades network meta-analysis did not improve the effect estimates compared to Smith pairwise meta-analysis, as suggested in Theorems 1–3.

7.3.2 Real data analysis

We applied the Smith and Lu–Ades models to the data collected by Trikalinos et al. [140], consisting of 63 studies of four treatments for non-acute coronary artery disease. All studies are two-armed. We indexed the treatments as (1) medical therapy; (2) percutaneous transluminal balloon coronary angioplasty; (3) bare-metal stents; and (4) drug-eluting stents. The outcome is the number of deaths due to the disease in each treatment group, which follows a binomial distribution. The complete data are available in Appendix A.11. We used the logit link function for the Smith and Lu–Ades models, so the overall relative effects produced by these models are log odds ratios comparing pairs among the four treatments. Also, in the Lu–Ades model, the treatment with the smallest index was used as the baseline in each study.

Figure 7.2 presents the treatment network; we refer to this as the full network. The full network has one evidence cycle, while the treatment comparison 4 vs. 3 is acyclic as it is not contained in any cycles. To illustrate the performance of the Lu–Ades model in a network without evidence cycles, we removed the four studies that directly compare treatments 3 vs. 1 from the complete data; the remaining studies lead to a chain-shaped network without cycles, which we call the reduced network. The

Smith and Lu–Ades models were applied to both the full and reduced networks. In the Lu–Ades model, $\boldsymbol{e}_{\mathrm{b}} = (d_{12}, d_{23}, d_{34})^{\mathrm{T}}$ was chosen as the set of basic parameters; thus, $\boldsymbol{e}_{\mathrm{f}} = (d_{13}, d_{14}, d_{24})^{\mathrm{T}}$ was the set of functional parameters. The three model settings in Section 7.3.1's simulations were considered, and vague priors were assigned to the study-specific baseline effects and the basic parameters. In the random-effects models, $U(0, 10)$ priors were used for the heterogeneity standard deviations σ_{12}, σ_{23}, and σ_{34}. When the Lu–Ades random-effects model with different heterogeneity variances was applied to the full network, due to the triangle inequality constraint (7.7) in the evidence cycle, the prior of σ_{13} was set to $U(|\sigma_{12} - \sigma_{23}|, \sigma_{12} + \sigma_{23})$ as suggested by Lu and Ades [52]. Three chains were used to implement the Smith and Lu–Ades models via Markov chain Monte Carlo; each chain contained a run of 100,000 updates after a 100,000-run burn-in period.

Table 7.2 presents the median overall log odds ratios of all treatment contrasts with their 95% credible intervals. When pairwise meta-analysis was applied to the full network, the estimation of the indirect comparisons d_{14} and d_{24} was not applicable due to unknown correlations between the separate estimated effects of d_{12}, d_{13}, d_{23}, and d_{34}; however, this difficulty does not exist in the reduced network without cycles, as shown in Theorem 1. The potential scale reduction factors [162] of all traced parameters were much smaller than 1.05, indicating that the Markov chains sampled from the posterior distributions have stabilized; also, the convergence of the chains was visually checked using trace plots. In addition, we assessed the Monte Carlo standard errors of the point and interval estimates using the `R` package 'mcmcse'. Most results have Monte Carlo standard errors much less than 0.01; those with standard errors greater than 0.01 are noted in Table 7.2.

For the reduced chain-shaped network, under each model setting, the Smith and Lu–Ades models produced nearly the same estimates of log odds ratios for all six treatment contrasts. Most differences between the two models are no more than 0.01 in absolute magnitude for point estimates and lower/upper bounds of 95% credible intervals, and are due entirely to Monte Carlo error. These results are consistent with Theorems 1–3. When the Lu–Ades model was applied to the full network, Table 7.2 shows that the estimated overall log odds ratios of the basic parameters d_{12} and d_{23} differ from those using the reduced network; thus, d_{13}, d_{14}, and d_{24}, which are based on the basic parameters d_{12} and d_{23}, also differ from their results using the reduced network. Recall that the reduced network only removed four studies that compare treatments 3 vs. 1. However, two of the four studies enrolled more than 1000 patients in each of their treatment groups, and they are the largest two among all 63 studies in the full network; see Table A.10 in Appendix A.11. Thus, the removal of these large studies caused the large differences noted above. Nevertheless, since the treatment contrast 4 vs. 3 is not contained in any evidence cycles, the estimated overall log odds ratio of d_{34} differs by no more than Monte Carlo error when the Lu–Ades model was used for the full and reduced networks under both the fixed-effects setting and the random-effects setting with different heterogeneity variances. This is consistent with Theorem 4. Furthermore, when all treatment contrasts were assumed to have a common heterogeneity variance, the 95% credible interval of the log odds ratio for d_{34} using the reduced network noticeably differs from that using the full network. This change arises because the estimate of d_{34} partly depends on the estimated heterogeneity variance, which is influenced by the removal of the four studies that compare treatments 3 vs. 1 from the full network.

7.4 Discussion

In applications, the equivalence of Lu–Ades network meta-analysis and Smith pairwise meta-analysis for acyclic comparisons is rarely noticed, even if the results from both types of meta-analyses are reported. This may be due to two reasons: inconsistent model assumptions and model specifications. First, most articles implement the Lu–Ades model using a common heterogeneity variance for all treatment comparisons, while performing separate pairwise meta-analyses using different heterogeneity variances for each treatment comparison. Due to these inconsistent model assumptions, the effect estimates produced by network and pairwise meta-analyses for acyclic comparisons are different, as suggested by Theorem 3. As noted, the assumption of a common heterogeneity variance was used in Lu and Ades [51] for conceptual and technical simplicity, and it may not be realistic in many cases [54]. Second, compared to the Smith Bayesian hierarchical model, the frequentist inverse-variance fixed-effects model or DerSimonian–Laird random-effects model [69] currently dominates pairwise meta-analysis, possibly because the frequentist models can be easily implemented by various statistical software packages (e.g., [163, 164]). These frequentist methods usually produce effect estimates noticeably different from the Smith Bayesian model. Hence, when reporting results from both pairwise and network meta-analyses, researchers are encouraged to use consistent model specifications, such as the Lu–Ades model combined with the Smith model, so that the benefit of network meta-analysis can be accurately reflected by the differences between the results from pairwise and network meta-analyses.

This chapter showed that evidence cycles are necessary to improve effect estimates when using Lu–Ades network meta-analysis. Such improvement depends highly on the evidence consistency assumption (7.3) for each cycle, which effectively reduces the degrees of freedom of the total of $(K-1)K/2$ treatment comparisons d_{hk} $(1 \leq h < k \leq K)$. However, each cycle potentially suffers from evidence inconsistency [52, 58], which is caused by a discrepancy among the trio of treatment comparisons within evidence cycles. By allowing inconsistency factors w for evidence cycles to deal with this problem, the degrees of freedom of the treatment contrasts increases, and the power of Lu–Ades network meta-analysis is accordingly reduced. In other words, when using the Lu–Ades model to gain more power from network meta-analysis, researchers must accept a greater risk of evidence inconsistency.

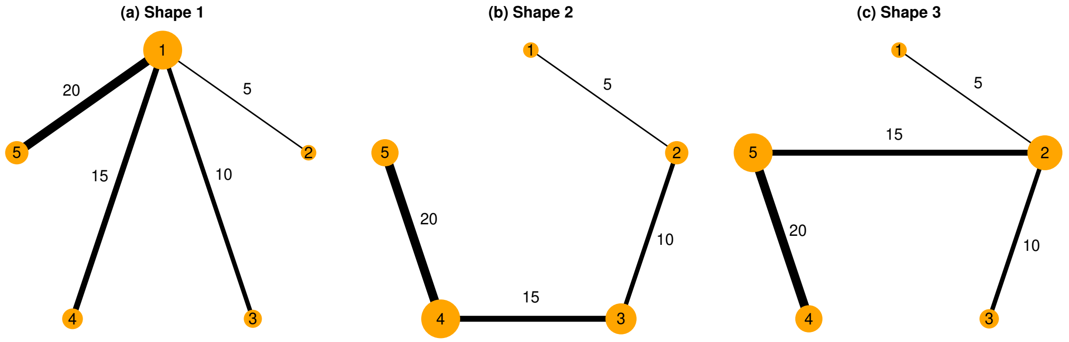


Figure 7.1: Simulated treatment networks with three shapes. Vertices represent treatments; edges represent direct comparisons. Edge width is proportional to the number of studies that report the corresponding direct comparison; vertex size is proportional to the number studies that include the corresponding treatment.

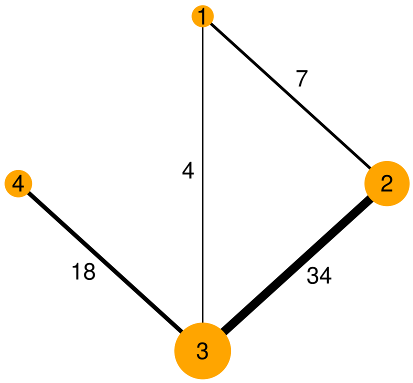


Figure 7.2: Network of four treatments on non-acute coronary artery disease. Treatment IDs: (1) medical therapy; (2) percutaneous transluminal balloon coronary angioplasty; (3) bare-metal stents; and (4) drug-eluting stents.

Table 7.1: Biases (outside brackets), mean squared errors (inside parentheses), and 95% credible interval coverage probabilities (%, inside square brackets) of the estimated relative effects produced by the Smith model (pairwise meta-analysis) and the Lu–Ades model (network meta-analysis) in simulations. The data were simulated using different heterogeneity standard deviations for different treatments.

Network	Treatment	Network meta-analysis			Pairwise meta-analysis		
shape	contrast	FE	RE1	RE2	FE	RE1	RE2
Shape 1	d_{12}	−0.03	−0.03	−0.03	−0.03	−0.03	−0.02
		(0.21)	(0.20)	(0.20)	(0.21)	(0.20)	(0.20)
		[81]	[99]	[100]	[81]	[99]	[100]
	d_{15}	0.03	0.03	0.03	0.03	0.03	0.03
		(0.35)	(0.18)	(0.18)	(0.35)	(0.18)	(0.18)
		[41]	[96]	[92]	[40]	[95]	[92]
	d_{23}	0.02^{a}	0.02	0.02	0.02^{a}	0.02	0.02
		(0.43)	(0.35)	(0.35)	(0.43)	(0.35)	(0.35)
		[73]	[99]	[100]	[73]	[99]	[100]
	d_{45}	0.04^{a}	0.03	0.03	0.04^{a}	0.03	0.03
		(0.59^{c})	(0.31)	(0.31)	(0.59^{c})	(0.31)	(0.31)
		[44]	[97]	[95]	[44]	[97]	[95]
Shape 2	d_{12}	−0.02	−0.03	−0.02	−0.03	−0.03	−0.02
		(0.21)	(0.20)	(0.21)	(0.21)	(0.20)	(0.21)
		[82]	[99]	[100]	[81]	[99]	[100]
	d_{13}	-0.01^{a}	−0.02	−0.02	-0.03^{a}	−0.02	−0.02
		(0.42)	(0.35)	(0.37)	(0.42)	(0.35)	(0.36)
		[76]	[99]	[99]	[75]	[99]	[100]
	d_{15}	-0.01^{b}	-0.03^{a}	-0.02^{a}	-0.03^{b}	-0.02^{a}	-0.02^{a}
		(1.09^{d})	(0.73^{d})	(0.74^{d})	(1.09^{d})	(0.73^{d})	(0.74^{d})
		[63]	[99]	[98]	[62]	[99]	[98]
	d_{45}	0.02	0.02	0.02	0.02	0.02	0.02
		(0.37)	(0.18)	(0.18)	(0.37)	(0.18)	(0.18)
		[37]	[96]	[92]	[36]	[96]	[92]
Shape 3	d_{12}	−0.02	−0.03	−0.02	−0.03	−0.03	−0.02
		(0.21)	(0.20)	(0.21)	(0.21)	(0.20)	(0.21)
		[82]	[99]	[100]	[81]	[99]	[100]
	d_{13}	-0.02^{a}	−0.02	−0.02	-0.03^{a}	−0.02	−0.02
		(0.42)	(0.35)	(0.37)	(0.42)	(0.35)	(0.37)
		[75]	[99]	[100]	[75]	[99]	[100]
	d_{15}	-0.01^{a}	-0.02^{a}	-0.02^{a}	-0.02^{a}	-0.02^{a}	-0.02^{a}
		(0.63^{c})	(0.44)	(0.44)	(0.63^{c})	(0.43)	(0.44)
		[62]	[99]	[99]	[61]	[99]	[99]
	d_{45}	0.02	0.02	0.02	0.02	0.02	0.02
		(0.37)	(0.18)	(0.18)	(0.37)	(0.18)	(0.18)
		[37]	[96]	[92]	[36]	[96]	[92]

FE: fixed-effects model; RE1: random-effects model with different heterogeneity variances for different treatment contrasts; RE2: random-effects model with a common heterogeneity variance.

d_{hk}: treatment k compared to h.

Monte Carlo standard error of bias: a, 0.02–0.03; b, 0.03–0.04; otherwise, less than 0.02. Monte Carlo standard error of mean squared error: c, 0.02–0.03; d, 0.03–0.05; otherwise, less than 0.02. Monte Carlo standard errors of all coverage probabilities are less than 2%.

Table 7.2: Log odds ratios (95% credible intervals) between the four treatments on non-acute coronary artery disease.

	Network meta-analysis			Pairwise meta-analysis		
LOR	FE	RE1	RE2	FE	RE1	RE2
Full network:						
d_{12}	−0.07	−0.16	−0.12	−0.21	−0.29	−0.29
	(−0.31, 0.17)	(−0.65, 0.32)	(−0.58, 0.28)	(−0.52, 0.09)	(−1.06, 0.30)	(−0.84, 0.20)
d_{13}	−0.11	−0.24	−0.22	−0.04	0.00	−0.01
	(−0.31, 0.08)	(−0.91[a], 0.25)	(−0.73, 0.20)	(−0.26, 0.18)	(−2.06[c], 2.64[c])	(−0.65, 0.73)
d_{14}	−0.03	−0.22	−0.19	N/A	N/A	N/A
	(−0.49, 0.42)	(−1.10[a], 0.50[a])	(−0.98, 0.46)			
d_{23}	−0.05	−0.10	−0.10	−0.21	−0.22	−0.21
	(−0.29, 0.20)	(−0.58, 0.34)	(−0.47, 0.25)	(−0.53, 0.11)	(−0.81, 0.34)	(−0.62, 0.19)
d_{24}	0.03	−0.07	−0.07	N/A	N/A	N/A
	(−0.45, 0.52)	(−0.83[a], 0.60)	(−0.75, 0.54)			
d_{34}	0.08	0.04	0.03	0.08	0.04	0.03
	(−0.33, 0.49)	(−0.56[a], 0.53)	(−0.52, 0.54)	(−0.33, 0.50)	(−0.55, 0.53)	(−0.53, 0.55)
Reduced chain-shaped network:						
d_{12}	−0.21	−0.29	−0.31	−0.21	−0.29	−0.30
	(−0.51, 0.09)	(−1.03[b], 0.31[a])	(−0.91, 0.24)	(−0.52, 0.09)	(−1.06, 0.30)	(−0.91, 0.24)
d_{13}	−0.42	−0.52	−0.52	−0.42	−0.53	−0.51
	(−0.86, 0.03)	(−1.41[b], 0.31[a])	(−1.26, 0.17)	(−0.86, 0.02)	(−1.47, 0.30)	(−1.26, 0.17)
d_{14}	−0.34	−0.49[a]	−0.51	−0.34	−0.49	−0.50
	(−0.95, 0.27)	(−1.55[b], 0.46[a])	(−1.48[a], 0.36)	(−0.95, 0.27)	(−1.60, 0.48)	(−1.48, 0.35)
d_{23}	−0.21	−0.22	−0.21	−0.21	−0.22	−0.21
	(−0.53, 0.11)	(−0.80, 0.34)	(−0.64, 0.21)	(−0.53, 0.11)	(−0.81, 0.34)	(−0.65, 0.21)
d_{24}	−0.13	−0.19	−0.20	−0.13	−0.18	−0.20
	(−0.66, 0.40)	(−1.00[a], 0.56)	(−0.94, 0.49)	(−0.66, 0.40)	(−1.02, 0.56)	(−0.94, 0.48)
d_{34}	0.08	0.04	0.01	0.08	0.04	0.01
	(−0.34, 0.50)	(−0.55[a], 0.53)	(−0.58, 0.55)	(−0.33, 0.50)	(−0.55, 0.53)	(−0.58, 0.55)

LOR: log odds ratio; FE: fixed-effects model; RE1: random-effects model with different heterogeneity variances for different treatment contrasts; RE2: random-effects model with a common heterogeneity variance; N/A: not applicable.

d_{hk}: treatment k compared to h.

Monte Carlo standard error: a, 0.01–0.02; b, 0.02–0.03; c, 0.06–0.07; otherwise, less than 0.01.

Chapter 8

Conclusion

8.1 Summary of major findings

This thesis introduced several innovative statistical methods and ideas for both univariate and multivariate meta-analyses. Outlying studies are common in meta-analyses and have great impact on conventional heterogeneity measures; however, no widely accepted guidelines exist for handling outliers. Chapter 2 proposed new heterogeneity measures that are less affected by outliers than the conventional ones. Assessing publication bias is another critical problem in meta-analysis. Chapter 3 empirically compared the performance of seven popular methods for publication bias using a large collection of real meta-analyses from the Cochrane Library. We found that Egger's regression test detected publication bias in more meta-analyses than the other methods, while the agreement among the seven tests was generally low or moderate, indicating potential limitations of current methods. Chapter 4 proposed an intuitive measure to quantify publication bias so that the severity of publication bias can be compared across meta-analyses. Specifically, the intercept from Egger's regression test can serve as a measure of publication bias, though it does not fully reflect the collected studies' asymmetry due to publication bias. We introduced a new measure, the skewness of the regression residuals, which has a more intuitive interpretation than the regression intercept. It can be also used as a test statistic for publication bias; simulations showed that it is powerful in many settings.

The data available for meta-analysis have been greatly enriched due to recent trends of data sharing, and methods for multivariate meta-analysis are being increasingly developed to synthesize the effects of multiple outcomes, multiple treatments, etc. A disease condition is typically associated with multiple risk and protective factors in medical sciences. Many studies report associations for multiple factors, but so far nearly all published meta-analyses separately synthesize the association between each factor and the disease condition. As the collected studies usually report different subsets of factors and use different subpopulations, results from separate meta-analyses may not be comparable. Chapter 5 introduced an innovative concept, multivariate meta-analysis of multiple factors, which synthesizes the factors simultaneously and thus improves statistical efficiency and reduces potential biases compared with separate analyses. The difficulty in multivariate meta-analysis of multiple factors is that the factors are likely correlated within studies but such correlations are usually unavailable from published articles. We used a Bayesian hierarchical model to handle this problem.

Network meta-analysis has also become very popular in the last decade. We have released a user-friendly `R` package 'pcnetmeta' to implement an arm-based network meta-analysis model, which focuses on estimating treatment-specific effects and is based on a missing data perspective. For binary outcomes, the arm-based method reports comprehensive summary results, including event rates, risk ratios, risk differences, and odds ratios; thus, it is more flexible than the contrast-based network meta-analysis method, which focuses only on estimating relative effects (e.g., odds ratios). The arm-based method can use single-arm studies, while the contrast-based method cannot. Chapter 6 showed that single-arm studies provide valuable information for treatment comparisons and enhance the robustness of a network meta-analysis.

Although network meta-analysis is generally considered more powerful than conventional pairwise meta-analyses of pairs of treatments, the improvement of effect estimates produced by network meta-analysis has never been studied theoretically. Chapter 7 proved that a pairwise comparison that is not part of an evidence cycle in a contrast-based network meta-analysis has posterior distribution identical to that produced by a simple pairwise meta-analysis. Many network meta-analyses do not contain evidence cycles, such as star-shaped treatment networks, in which several active treatments are compared with the control but the active treatments are not mutually compared. In such settings, the results of Lu–Ades network meta-analysis model are therefore equivalent to performing separate pairwise meta-analyses on each treatment comparison. We also illustrated this equivalence using simulation studies and a real data analysis. We hope that the findings in this thesis will provide other researchers with valuable insights into future systematic reviews and meta-analyses.

8.2 Future research

Data sharing will necessarily lead to fast development of multivariate meta-analysis so that all available information can be used effectively. In this connection, we have several aims for future research on meta-analysis.

(i) *Methods for assessing publication bias in multivariate meta-analysis.* Although many methods are available for univariate meta-analysis, assessing publication bias in multivariate meta-analysis (including network meta-analysis) is largely untouched. As the dimension increases, assessing publication bias becomes challenging because some studies may be completely suppressed while some may selectively report subsets of their outcomes. Approaches to adjusting for publication bias will greatly improve the precision of conclusions from multivariate meta-analyses. We plan to extend the work of quantifying publication bias in univariate meta-analysis to multivariate settings. Based on the multivariate version of Egger's regression test, the skewness of study-specific multivariate residuals can be used as an overall publication bias measure for all endpoints. We will derive test statistics and their theoretical properties based on this overall measure.

(ii) *Meta-analysis methods based on penalized likelihood.* As discussed in Chapter 2, conventionally, heterogeneity between studies needs to be assessed separately for each endpoint, and either fixed or random effects are accordingly used to estimate the effect of that endpoint. This process is inefficient when the meta-analysis dataset contains many endpoints. Also, it is generally recognized that both the fixed- and random-effects models have several limitations. In addition, heterogeneity is overestimated in the presence of outliers, which are common in meta-analyses. The new penalized-likelihood-based method will use a set of tuning parameters to control the strength of penalties on the likelihood of the random-effects model. If the tuning parameters are set to zero, the resulting effect estimates are identical to those produced by the conventional random-effects model. If the tuning parameters are large enough, the new method leads to the fixed-effects estimates. Therefore, this method can be viewed as trading off between the fixed- and random-effects models.

(iii) *Meta-analysis that combines patient-level data from existing databases with aggregated summary data from a literature search.* This combination can overcome many drawbacks of using only aggregated data. For example, aggregated data are often poorly reported and presented differently across studies, such as reporting odds ratio vs. relative risk. The trend toward data sharing lays a promising foundation for meta-analysis of individual patient data. Taking advantage of the increasing attention from governmental organizations and medical journals, we plan to request individual patient data from data-sharing resources and use them to help evaluate the performance of the new methods.

(iv) *Meta-analysis accounting for post-randomization variables.* Non-compliance with assigned treatments is common in randomized controlled trials and may induce bias in estimated treatment effects. The main existing method, meta-regression, adjusts for study-level baseline covariates, not for arm-level post-randomization variables. We will develop an arm-based approach to jointly modeling both outcome measures and post-randomization variables in both univariate and multivariate meta-analyses, so that the estimated effects of different endpoints can be comparable at the same level of non-compliance.

Additionally, an important step in promoting new statistical methods is to provide open-source user-friendly software. We will continue to develop `R` packages and `SAS` macros so that researchers who do not specialize in (bio)statistics can implement the new methods easily. For example, many frequentist and Bayesian methods for publication bias are based on so-called selection models that not only detect but also adjust for publication bias. These methods are usually complicated and require careful coding, limiting their applications. We plan to release a sophisticated `R` package that includes a comprehensive set of selection-model-based methods for assessing publication bias.

References

[1] Hunter, J. E. and Schmidt, F. L. Cumulative research knowledge and social policy formulation: the critical role of meta-analysis. *Psychology, Public Policy, and Law*, 2(2):324–347, 1996.

[2] Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet*, 360(9349):1903–1913, 2002.

[3] Higgins, J. P. T. and Green, S. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Chichester, UK, 2008.

[4] Sutton, A. J. and Higgins, J. P. T. Recent developments in meta-analysis. *Statistics in Medicine*, 27(5):625–650, 2008.

[5] Berlin, J. A. and Golub, R. M. Meta-analysis as evidence: building a better pyramid. *JAMA*, 312(6):603–606, 2014.

[6] Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010.

[7] Riley, R. D., Higgins, J. P. T., and Deeks, J. J. Interpretation of random effects meta-analyses. *BMJ*, 342:d549, 2011.

[8] Ioannidis, J. P. A., Patsopoulos, N. A., and Evangelou, E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, 335(7626):914–916, 2007.

[9] Cochran, W. G. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129, 1954.

[10] Whitehead, A. and Whitehead, J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, 10(11):1665–1677, 1991.

[11] Hardy, R. J. and Thompson, S. G. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, 17(8):841–856, 1998.

[12] Jackson, D. The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine*, 25(15):2688–2699, 2006.

[13] Higgins, J. P. T. and Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[14] Viechtbauer, W. and Cheung, M. W.-L. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2):112–125, 2010.

[15] Gumedze, F. N. and Jackson, D. A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Medical Research Methodology*, 11(1):19, 2011.

[16] Hedges, L. V. and Olkin, I. *Statistical Method for Meta-Analysis.* Academic Press, Orlando, FL, 1985.

[17] Begg, C. B. and Berlin, J. A. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151(3):419–463, 1988.

[18] Stern, J. M. and Simes, R. J. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*, 315(7109):640–645, 1997.

[19] Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., and Jones, D. R. Empirical assessment of effect of publication bias on meta-analyses. *BMJ*, 320(7249):1574–1577, 2000.

[20] Thornton, A. and Lee, P. Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology*, 53(2):207–216, 2000.

[21] Kicinski, M., Springate, D. A., and Kontopantelis, E. Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine*, 34(20):2781–2793, 2015.

[22] Light, R. J. and Pillemer, D. B. *Summing Up: The Science of Reviewing Research.* Harvard University Press, Cambridge, MA, 1984.

[23] Sterne, J. A. C. and Egger, M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54(10):1046–1055, 2001.

[24] Begg, C. B. and Mazumdar, M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4):1088–1101, 1994.

[25] Egger, M., Davey Smith, G., Schneider, M., and Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634, 1997.

[26] Duval, S. and Tweedie, R. A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449):89–98, 2000.

[27] Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, 26(25):4544–4562, 2007.

[28] Rothstein, H. R., Sutton, A. J., and Borenstein, M. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* John Wiley & Sons, Chichester, UK, 2005.

[29] Tang, J.-L. and Liu, J. L. Y. Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology*, 53(5):477–484, 2000.

[30] Sterne, J. A. C., Gavaghan, D., and Egger, M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53(11):1119–1129, 2000.

[31] Deeks, J. J., Macaskill, P., and Irwig, L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology*, 58(9):882–893, 2005.

[32] Macaskill, P., Walter, S. D., and Irwig, L. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20(4):641–654, 2001.

[33] Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, 295(6):676–680, 2006.

[34] Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., and Higgins, J. P. T. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, 343:d4002, 2011.

[35] Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., and Cooper, N. J. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, 9(1):2, 2009.

[36] Bürkner, P.-C. and Doebler, P. Testing for publication bias in diagnostic meta-analysis: a simulation study. *Statistics in Medicine*, 33(18):3061–3077, 2014.

[37] Dear, K. B. G. and Begg, C. B. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, 7(2):237–245, 1992.

[38] Hedges, L. V. Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2):246–255, 1992.

[39] Silliman, N. P. Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association*, 92(439):926–936, 1997.

[40] Silliman, N. P. Nonparametric classes of weight functions to model publication bias. *Biometrika*, 84(4):909–918, 1997.

[41] Sutton, A. J., Song, F., Gilbody, S. M., and Abrams, K. R. Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research*, 9(5):421–445, 2000.

[42] Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., and Murray, C. J. L. Selected major risk factors and global and regional burden of disease. *The Lancet*, 360(9343):1347–1360, 2002.

[43] Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. *JAMA*, 248(12):1465–1477, 1982.

[44] Stamler, J., Vaccaro, O., Neaton, J. D., and Wentworth, D. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes Care*, 16(2):434–444, 1993.

[45] The Risk and Prevention Study Collaborative Group. n-3 fatty acids in patients with multiple cardiovascular risk factors. *The New England Journal of Medicine*, 368(19):1800–1808, 2013.

[46] Cole, M. G. and Dendukuri, N. Risk factors for depression among elderly community subjects: a systematic review and meta-analysis. *American Journal of Psychiatry*, 160(6):1147–1156, 2003.

[47] Ebrahim, S. and Davey Smith, G. Systematic review of randomised controlled trials of multiple risk factor interventions for preventing coronary heart disease. *BMJ*, 314(7095):1666–1674, 1997.

[48] Flenady, V., Koopmans, L., Middleton, P., Frøen, J. F., Smith, G. C., Gibbons, K., Coory, M., Gordon, A., Ellwood, D., McIntyre, H. D., Fretts, R., and Ezzati, M. Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. *The Lancet*, 377(9774):1331–1340, 2011.

[49] Serghiou, S., Patel, C. J., Tan, Y. Y., Koay, P., and Ioannidis, J. P. A. Field-wide meta-analyses of observational associations can map selective availability of risk factors and the impact of model specifications. *Journal of Clinical Epidemiology*, 71:58–67, 2016.

[50] Lumley, T. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16):2313–2324, 2002.

[51] Lu, G. and Ades, A. E. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20):3105–3124, 2004.

[52] Lu, G. and Ades, A. E. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, 101(474):447–459, 2006.

[53] Salanti, G., Higgins, J. P. T., Ades, A. E., and Ioannidis, J. P. A. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17(3):279–301, 2008.

[54] Lu, G. and Ades, A. E. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009.

[55] Salanti, G., Ades, A. E., and Ioannidis, J. P. A. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology*, 64(2):163–171, 2011.

[56] Salanti, G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, 3(2):80–97, 2012.

[57] Sutton, A. J., Welton, N. J., Cooper, N., Abrams, K. R., and Ades, A. E. *Evidence Synthesis for Decision Making in Healthcare*. John Wiley & Sons, Chichester, UK, 2012.

[58] Dias, S., Welton, N. J., Sutton, A. J., and Ades, A. E. Evidence synthesis for decision making 5: the baseline natural history model. *Medical Decision Making*, 33(5):657–670, 2013.

[59] Gøtzsche, P. C. Is there logic in the placebo? *The Lancet*, 344(8927):925–926, 1994.

[60] Hróbjartsson, A. What are the main methodological problems in the estimation of placebo effects? *Journal of Clinical Epidemiology*, 55(5):430–435, 2002.

[61] Cipriani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J. P. T., Churchill, R., Watanabe, N., Nakagawa, A., Omori, I. M., McGuire, H., Tansella, M., and Barbui, C. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet*, 373(9665):746–758, 2009.

[62] Fisher, L. D., Gent, M., and Büller, H. R. Active-control trials: how would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *American Heart Journal*, 141(1):26–32, 2001.

[63] Hasselblad, V. and Kong, D. F. Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal*, 35(2):435–449, 2001.

[64] Song, F., Altman, D. G., Glenny, A.-M., and Deeks, J. J. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*, 326(7387):472, 2003.

[65] Jansen, J. P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., Lee, K., Boersma, C., Annemans, L., and Cappelleri, J. C. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ispor task force on indirect treatment comparisons good research practices: part 1. *Value in Health*, 14(4):417–428, 2011.

[66] Mills, E. J., Kanters, S., Thorlund, K., Chaimani, A., Veroniki, A.-A., and Ioannidis, J. P. A. The effects of excluding treatments from network meta-analyses: survey. *BMJ*, 347:f5195, 2013.

[67] Zhang, J., Carlin, B. P., Neaton, J. D., Soon, G. G., Nie, L., Kane, R., Virnig, B. A., and Chu, H. Network meta-analysis of randomized clinical trials: reporting the proper summaries. *Clinical Trials*, 11(2):246–262, 2014.

[68] Portnoy, S. and Koenker, R. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators (with discussion). *Statistical Science*, 12(4):279–300, 1997.

[69] DerSimonian, R. and Laird, N. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[70] Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W. S., Tschannen, B., Altman, D. G., Egger, M., and Jüni, P. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ*, 341:c3515, 2010.

[71] Barnett, V. and Lewis, T. *Outliers in Statistical Data*. John Wiley & Sons, New York, NY, 3rd edition, 1994.

[72] Horowitz, J. L. Bootstrap methods for median regression models. *Econometrica*, 66(6):1327–1351, 1998.

[73] Huber, P. J. and Ronchetti, E. M. *Robust Statistics*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2009.

[74] Ismail, I., Keating, S. E., Baker, M. K., and Johnson, N. A. A systematic review and meta-analysis of the effect of aerobic vs. resistance exercise training on visceral fat. *Obesity Reviews*, 13(1):68–91, 2012.

[75] Haentjens, P., Magaziner, J., Colón-Emeric, C. S., Vanderschueren, D., Milisen, K., Velkeniers, B., and Boonen, S. Meta-analysis: excess mortality after hip fracture among older women and men. *Annals of Internal Medicine*, 152(6):380–390, 2010.

[76] Davey, J., Turner, R. M., Clarke, M. J., and Higgins, J. P. T. Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 11(1):160, 2011.

[77] Higgins, J. P. T. and Thompson, S. G. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine*, 23(11):1663–1682, 2004.

[78] Higgins, J. P. T. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, 37(5):1158–1160, 2008.

[79] Hedges, L. V. and Vevea, J. L. Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4):486–504, 1998.

[80] Chalmers, T. C. Problems induced by meta-analyses. *Statistics in Medicine*, 10(6):971–980, 1991.

[81] Poole, C. and Greenland, S. Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology*, 150(5):469–475, 1999.

[82] Henmi, M. and Copas, J. B. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, 29(29):2969–2983, 2010.

[83] Stanley, T. D. and Doucouliagos, H. Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine*, 34(13):2116–2127, 2015.

[84] Walter, S. D. and Cook, R. J. A comparison of several point estimators of the odds ratio in a single 2 × 2 contingency table. *Biometrics*, 47(3):795–811, 1991.

[85] Sweeting, M. J., Sutton, A. J., and Lambert, P. C. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375, 2004.

[86] Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russell Localio, A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1):53–77, 2007.

[87] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[88] Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[89] Ioannidis, J. P. A. and Trikalinos, T. A. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal*, 176(8):1091–1096, 2007.

[90] Sterne, J. A. C., Egger, M., and Davey Smith, G. Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, 323(7304):101–105, 2001.

[91] Harbord, R. M., Egger, M., and Sterne, J. A. C. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, 25(20):3443–3457, 2006.

[92] Thompson, S. G. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*, 309(6965):1351–1355, 1994.

[93] Normand, S.-L. T. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359, 1999.

[94] Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., Rushton, L., and Moreno, S. G. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):575–591, 2010.

[95] Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560, 2003.

[96] MacGillivray, H. L. Skewness and asymmetry: measures and orderings. *The Annals of Statistics*, 14(3):994–1011, 1986.

[97] Wright, S. P. Adjusted P-values for simultaneous inference. *Biometrics*, 48(4):1005–1013, 1992.

[98] Duval, S. and Tweedie, R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, 2000.

[99] Stead, L. F., Perera, R., Bullen, C., Mant, D., Hartmann-Boyce, J., Cahill, K., and Lancaster, T. Nicotine replacement therapy for smoking cessation. *Cochrane Database of Systematic Reviews*, 11:Art. No.: CD000146, 2012.

[100] Hróbjartsson, A. and Gøtzsche, P. C. Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews*, 1:Art. No.: CD003974, 2010.

[101] Liu, C. J. and Latham, N. K. Progressive resistance strength training for improving physical function in older adults. *Cochrane Database of Systematic Reviews*, 3:Art. No.: CD002759, 2009.

[102] Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61(10):991–996, 2008.

[103] Gøtzsche, P. C. Reference bias in reports of drug trials. *BMJ*, 295(6599):654–656, 1987.

[104] Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., and Perneger, T. V. Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology*, 66(3):296–301, 2013.

[105] Chalmers, T. C., Celano, P., Sacks, H. S., and Smith, H. Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine*, 309(22):1358–1361, 1983.

[106] Altman, D. G. Poor-quality medical research: what can journals do? *JAMA*, 287(21):2765–2767, 2002.

[107] Hill, D. L. and Rao, P. V. Tests of symmetry based on Cramér–von Mises statistics. *Biometrika*, 64(3):489–494, 1977.

[108] Antille, A., Kersting, G., and Zucchini, W. Testing symmetry. *Journal of the American Statistical Association*, 77(379):639–646, 1982.

[109] McWilliams, T. P. A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association*, 85(412):1130–1133, 1990.

[110] Chen, H., Manning, A. K., and Dupuis, J. A method of moments estimator for random effect multivariate meta-analysis. *Biometrics*, 68(4):1278–1284, 2012.

[111] Jackson, D., Riley, R., and White, I. R. Multivariate meta-analysis: potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[112] Jackson, D., White, I. R., and Thompson, S. G. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*, 29(12):1282–1297, 2010.

[113] Liu, D., Liu, R. Y., and Xie, M. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340, 2015.

[114] Riley, R. D., Abrams, K. R., Lambert, P. C., Sutton, A. J., and Thompson, J. R. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97, 2007.

[115] Chu, H. and Cole, S. R. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59(12):1331–1332, 2006.

[116] Ma, X., Nie, L., Cole, S. R., and Chu, H. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. *Statistical Methods in Medical Research*, 25(4):1596–1619, 2016.

[117] Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., and Zwinderman, A. H. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58(10):982–990, 2005.

[118] Berkey, C. S., Anderson, J. J., and Hoaglin, D. C. Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine*, 15(5):537–557, 1996.

[119] van Houwelingen, H. C., Zwinderman, K. H., and Stijnen, T. A bivariate approach to meta-analysis. *Statistics in Medicine*, 12(24):2273–2284, 1993.

[120] Riley, R. D. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):789–811, 2009.

[121] Riley, R. D., Thompson, J. R., and Abrams, K. R. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9(1):172–186, 2008.

[122] White, I. R. et al. Multivariate random-effects meta-regression: updates to mvmeta. *Stata Journal*, 11(2):255–270, 2011.

[123] Berlin, J. A., Laird, N. M., Sacks, H. S., and Chalmers, T. C. A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine*, 8(2):141–151, 1989.

[124] Petitti, D. B. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Oxford University Press, New York, NY, 2nd edition, 2000.

[125] Bickel, P. J. and Levina, E. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[126] Veroniki, A. A., Vasiliadis, H. S., Higgins, J. P. T., and Salanti, G. Evaluation of inconsistency in networks of interventions. *International Journal of Epidemiology*, 42(1):332–345, 2013.

[127] Ara, R., Pandor, A., Stevens, J., Rees, A., and Rafia, R. Early high-dose lipid-lowering therapy to avoid cardiac events: a systematic review and economic evaluation. *Health Technology Assessment*, 13(34), 2009.

[128] Ballesteros, J. Orphan comparisons and indirect meta-analysis: a case study on antidepressant efficacy in dysthymia comparing tricyclic antidepressants, selective serotonin reuptake inhibitors, and monoamine oxidase inhibitors by using general linear models. *Journal of Clinical Psychopharmacology*, 25(2):127–131, 2005.

[129] Bucher, H. C., Guyatt, G. H., Griffith, L. E., and Walter, S. D. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, 50(6):683–691, 1997.

[130] Eisenberg, M. J., Filion, K. B., Yavin, D., Bélisle, P., Mottillo, S., Joseph, L., Gervais, A., O'Loughlin, J., Paradis, G., Rinfret, S., and Pilote, L. Pharmacotherapies for smoking cessation: a meta-analysis of randomized controlled trials. *Canadian Medical Association Journal*, 179(2):135–144, 2008.

[131] Elliott, W. J. and Meyer, P. M. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *The Lancet*, 369(9557):201–207, 2007.

[132] Hasselblad, V. Meta-analysis of multitreatment studies. *Medical Decision Making*, 18(1):37–43, 1998.

[133] Goeree, R., O'Brien, B., Hunt, R., Blackhouse, G., Willan, A., and Watson, J. Economic evaluation of long term management strategies for erosive oesophagitis. *Pharmacoeconomics*, 16(6):679–697, 1999.

[134] Lu, G., Ades, A. E., Sutton, A. J., Cooper, N. J., Briggs, A. H., and Caldwell, D. M. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Statistics in Medicine*, 26(20):3681–3699, 2007.

[135] Middleton, L. J., Champaneria, R., Daniels, J. P., Bhattacharya, S., Cooper, K. G., Hilken, N. H., ODonovan, P., Gannon, M., Gray, R., and Khan, K. S. Hysterectomy, endometrial destruction, and levonorgestrel releasing intrauterine system (mirena) for heavy menstrual bleeding: systematic review and meta-analysis of data from individual patients. *BMJ*, 341:c3929, 2010.

[136] Mills, E. J., Wu, P., Spurden, D., Ebbert, J. O., and Wilson, K. Efficacy of pharmacotherapies for short-term smoking abstinance: a systematic review and meta-analysis. *Harm Reduction Journal*, 6(25):1–16, 2009.

[137] Picard, P. and Tramer, M. R. Prevention of pain on injection with propofol: a quantitative systematic review. *Anesthesia & Analgesia*, 90(4):963–969, 2000.

[138] Puhan, M. A., Bachmann, L. M., Kleijnen, J., ter Riet, G., and Kessels, A. G. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. *BMC Medicine*, 7(1):2, 2009.

[139] Thijs, V., Lemmens, R., and Fieuws, S. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. *European Heart Journal*, 29(9):1086–1092, 2008.

[140] Trikalinos, T. A., Alsheikh-Ali, A. A., Tatsioni, A., Nallamothu, B. K., and Kent, D. M. Percutaneous coronary interventions for non-acute coronary artery disease: a quantitative 20-year synopsis and a network meta-analysis. *The Lancet*, 373(9667):911–918, 2009.

[141] Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. CRC press, Boca Raton, FL, 1998.

[142] Begg, C. B. and Pilote, L. A model for incorporating historical controls into a meta-analysis. *Biometrics*, 47(3):899–906, 1991.

[143] Sutton, A. J., Abrams, K. R., Jones, D. R., Jones, D. R., Sheldon, T. A., and Song, F. *Methods for Meta-Analysis in Medical Research*. Chichester, UK, 2000.

[144] Heisel, O., Heisel, R., Balshaw, R., and Keown, P. New onset diabetes mellitus in patients receiving calcineurin inhibitors: a systematic review and meta-analysis. *American Journal of Transplantation*, 4(4):583–595, 2004.

[145] Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014.

[146] Shuster, J. J., Guo, J. D., and Skyler, J. S. Meta-analysis of safety for low event-rate binomial trials. *Research Synthesis Methods*, 3(1):30–50, 2012.

[147] Zhao, H., Hodges, J. S., Ma, H., Jiang, Q., and Carlin, B. P. Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis. *Statistics in Medicine*, 35(20):3524–3536, 2016.

[148] Hutton, B., Salanti, G., Caldwell, D. M., Chaimani, A., Schmid, C. H., Cameron, C., Ioannidis, J. P. A., Straus, S., Thorlund, K., Jansen, J. P., Mulrow, C., Catalá-López, F., Gøtzsche, P. C., Dickersin, K., Boutron, I., Altman, D. G., and Moher, D. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Annals of Internal Medicine*, 162(11):777–784, 2015.

[149] Efthimiou, O., Debray, T. P. A., van Valkenhoef, G., Trelle, S., Panayidou, K., Moons, K. G. M., Reitsma, J. B., Shang, A., and Salanti, G. GetReal in network meta-analysis: a review of the methodology. *Research Synthesis Methods*, 7(3):236–263, 2016.

[150] Hong, H., Chu, H., Zhang, J., and Carlin, B. P. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, 7(1):6–22, 2016.

[151] Lin, L., Zhang, J., Hodges, J. S., and Chu, H. Performing arm-based network meta-analysis in R with the pcnetmeta package. *Journal of Statistical Software*, page in press, 2016.

[152] Nikolakopoulou, A., Chaimani, A., Veroniki, A. A., Vasiliadis, H. S., Schmid, C. H., and Salanti, G. Characteristics of networks of interventions: a description of a database of 186 published networks. *PLoS One*, 9(1):e86754, 2014.

[153] Higgins, J. P. T. and Welton, N. J. Network meta-analysis: a norm for comparative effectiveness? *The Lancet*, 386(9994):628–630, 2015.

[154] Smith, T. C., Spiegelhalter, D. J., and Thomas, A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, 14(24):2685–2699, 1995.

[155] Dias, S., Welton, N. J., Caldwell, D. M., and Ades, A. E. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, 29(7-8):932–944, 2010.

[156] Higgins, J. P. T., Jackson, D., Barrett, J. K., Lu, G., Ades, A. E., and White, I. R. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods*, 3(2):98–110, 2012.

[157] White, I. R., Barrett, J. K., Jackson, D., and Higgins, J. P. T. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods*, 3(2):111–125, 2012.

[158] Mills, E. J., Thorlund, K., and Ioannidis, J. P. A. Demystifying trial networks and network meta-analysis. *BMJ*, 346:f2914, 2013.

[159] Trelle, S., Reichenbach, S., Wandel, S., Hildebrand, P., Tschannen, B., Villiger, P. M., Egger, M., and Jüni, P. Cardiovascular safety of non-steroidal anti-inflammatory drugs: network meta-analysis. *BMJ*, 342:c7086, 2011.

[160] Khera, R., Murad, M. H., Chandar, A. K., Dulai, P. S., Wang, Z., Prokop, L. J., Loomba, R., Camilleri, M., and Singh, S. Association of pharmacological treatments for obesity with weight loss and adverse events: a systematic review and meta-analysis. *JAMA*, 315(22):2424–2434, 2016.

[161] Salanti, G., Kavvoura, F. K., and Ioannidis, J. P. A. Exploring the geometry of treatment networks. *Annals of Internal Medicine*, 148(7):544–553, 2008.

[162] Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[163] Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

[164] Palmer, T. M. and Sterne, J. A. C. *Meta-Analysis in Stata: An Updated Collection From the Stata Journal.* Stata Press, College Station, TX, 2nd edition, 2016.

[165] Parzen, M. I., Wei, L. J., and Ying, Z. A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350, 1994.

[166] Harville, D. A. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[167] Jennrich, R. I. and Schluchter, M. D. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986.

[168] Pinheiro, J. C. and Bates, D. M. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996.

[169] Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[170] Yao, H., Kim, S., Chen, M.-H., Ibrahim, J. G., Shah, A. K., and Lin, J. Bayesian inference for multivariate meta-regression with a partially observed within-study sample covariance matrix. *Journal of the American Statistical Association*, 110(510):528–544, 2015.

[171] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis.* Taylor & Francis, Boca Raton, FL, 3rd edition, 2014.

[172] Wei, Y. and Higgins, J. P. T. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(17):2911–2934, 2013.

[173] Barnard, J., McCulloch, R., and Meng, X.-L. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312, 2000.

[174] Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[175] Robert, C. P. and Casella, G. *Monte Carlo Statistical Methods.* Springer Science+Business Media, New York, NY, 2nd edition, 2004.

[176] Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data.* John Wiley & Sons, Hoboken, NJ, 2nd edition, 2002.

[177] Zeger, S. L., Liang, K.-Y., and Albert, P. S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44(4):1049–1060, 1988.

[178] Carlin, B. P. and Louis, T. A. *Bayesian Methods for Data Analysis.* CRC Press, Boca Raton, FL, 3rd edition, 2008.

[179] Dias, S. and Ades, A. E. Absolute or relative effects? Arm-based synthesis of trial data. *Research Synthesis Methods*, 7(1):23–28, 2016. RSM-07-2015-0036.R1.

[180] Hong, H., Chu, H., Zhang, J., and Carlin, B. P. Rejoinder to the Discussion of "a Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons," by S. Dias and A.E. Ades. *Research synthesis methods*, 7(1):29, 2016.

[181] Serfling, R. J. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York, NY, 1980.

[182] Cramér, H. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ, 1999.

Appendix A

Supplementary Materials

A.1 Sensitivity analysis for the weighted-median-based robust heterogeneity measure

Since the weighted median in Q_m is discontinuous due to the indicator function [165] in Equation (2.2), the approach by Horowitz [72] is applied to approximate the indicator function $\mathbb{I}(t > 0)$ by a smooth function $J(t)$ in the following simulations and case studies. For example, $J(t)$ can be the scaled expit function $J_\epsilon(t) = 1/[1 + \exp(-t/\epsilon)]$, where ϵ is a pre-specified small constant, say 10^{-4}. This section presents sensitivity analysis on the choice of ϵ. We use the data of the case study in Section 2.5.1. Table A.1 presents the results based on $B = 10,000$ resampling iterations.

A.2 Performance of heterogeneity measures in three artificial meta-analyses

This section illustrates that I_r^2 and I_m^2 can be larger than I^2 and provide useful information on assessing heterogeneity. Three artificial meta-analyses were created; each contains ten studies with the same within-study variance 1. The observed effect sizes in half of the studies are $y_i = b$, and those in another half are $y_i = -b$, where b was set to 0.5, 1, and 2. Figure A.1 presents the corresponding forest plots. Note that in these meta-analyses, the condition $w_i(y_i - \bar{\mu})^2 = C$ is satisfied, so the equality in

$I_r^2 \leq I^2 + (1 - 2/\pi)(1 - I^2)$ holds.

Figure A.1(a) shows the meta-analysis with $b = 0.5$. Since the observed effect size of each study is contained in the 95% CIs of all other studies, the collected studies are considered homogeneous; all of I^2, I_r^2, and I_m^2 are calculated as 0. For the meta-analysis with $b = 1.0$ shown in Figure A.1(b), five studies report the effect size -1, lying outside the 95% CIs $(-0.96, 2.96)$ of the other five studies. Despite this, the 95% CIs of the total ten studies overlap in a large region, i.e., $(-0.96, 0.96)$. Therefore, the between-study heterogeneity is moderate, but may not be substantial. The three heterogeneity measures are calculated as $I^2 = 0.10$, $I_r^2 = 0.43$, and $I_m^2 = 0.36$; I^2 may indicate homogeneity but both I_r^2 and I_m^2 imply moderate heterogeneity. Figure A.1(c) shows the meta-analysis with $b = 2$. The 95% CIs of five studies do not overlap with those in the other five studies; therefore, these studies are clearly heterogeneous. The heterogeneity measures are calculated as $I^2 = 0.78$, $I_r^2 = 0.86$, and $I_m^2 = 0.84$; all suggest considerable heterogeneity.

A.3 Complete simulation results for heterogeneity measures

The simulation settings have been detailed in Section 2.4. Tables A.2–A.4 present the complete results.

A.4 Performance of various publication bias tests based on the restricted dataset

Figures A.2–A.4 present the results based on the restricted Cochrane dataset, which consists of the largest meta-analysis from each systematic review. They correspond to Figures 3.1–3.3 in Section 3.2.

A.5 Implementation of multivariate hybrid model

A.5.1 Restricted maximum likelihood method

Since the hybrid model for complete data is a special case of that for missing data by setting $\mathbf{X}_i$ to the $p \times p$ identity matrix $\mathbf{I}_p$, we only discuss the situation of missing data. For the frequentist approach, we consider the restricted maximum likelihood (REML) method, which is commonly used to estimate variance/covariance components [166]. The restricted log-likelihood of the hybrid model is [167]

$$\lambda_{\text{REML}} = \text{Const.} - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{\Phi}_i| - \frac{1}{2}\log\left|\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Phi}_i^{-1}\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{r}_i,$$

where $\boldsymbol{r}_i = \boldsymbol{y}_i - \mathbf{X}_i\widetilde{\boldsymbol{\mu}}$ represents the residuals and

$$\widetilde{\boldsymbol{\mu}} = (\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Phi}_i^{-1}\mathbf{X}_i)^{-1}(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Phi}_i^{-1}\boldsymbol{y}_i).$$

We may treat λ_{REML} as the log-likelihood of the residuals $\boldsymbol{r}_i$, and the REML estimates are obtained by maximizing λ_{REML}. Denote the estimates of $\boldsymbol{\Psi}$ and $\mathbf{R}$ as $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\mathbf{R}}$, respectively. Hence, the overall effect size $\boldsymbol{\mu}$ is estimated as

$$\widehat{\boldsymbol{\mu}} = (\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}_i^{-1}\mathbf{X}_i)^{-1}(\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}_i^{-1}\boldsymbol{y}_i),$$

where $\widehat{\boldsymbol{\Phi}}_i = (\mathbf{D}_i + \mathbf{X}_i\widehat{\boldsymbol{\Psi}}\mathbf{X}_i^{\mathrm{T}})^{1/2}\mathbf{X}_i\widehat{\mathbf{R}}\mathbf{X}_i^{\mathrm{T}}(\mathbf{D}_i + \mathbf{X}_i\widehat{\boldsymbol{\Psi}}\mathbf{X}_i^{\mathrm{T}})^{1/2}$. The covariance matrix of $\widehat{\boldsymbol{\mu}}$ is estimated as $\widehat{\text{Var}}[\widehat{\boldsymbol{\mu}}] = (\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}_i^{-1}\mathbf{X}_i)^{-1}$.

The optimization problem is subject to that the marginal correlation matrix $\mathbf{R}$ is positive definite and its diagonal elements are 1. To ensure these constraints, we consider the spherical decomposition of $\mathbf{R}$ [168]. This technique is basically a reparameterization of the Cholesky decomposition. Specifically, write $\mathbf{R} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L} = (L_{ij})$ is a lower triangular matrix with nonnegative diagonal elements. Let $L_{11} = 1$ and for $i = 2, \dots, p$,

$$L_{ij} = \begin{cases} \cos\theta_{i2} & \text{if } j = 1; \\ \left(\prod_{k=2}^{j} \sin\theta_{ik}\right)\cos\theta_{i,j+1} & \text{if } j = 2, \dots, i-1; \\ \prod_{k=2}^{i} \sin\theta_{ik} & \text{if } j = i. \end{cases}$$

Here, θ_{ij}'s are angle parameters for $2 \leq j \leq i \leq p$. Note that using this parameterization, the diagonal elements of $\mathbf{R}$, $r_{jj} = \sum_{i=1}^{j} L_{ij}^2$, are guaranteed to be 1 by the properties of sine and cosine functions. Moreover, to ensure the uniqueness of L_{ij}'s, the angle parameters are constrained to be $\theta_{ij} \in (0, \pi)$. Since the boundaries of the parameters are linear, and the optimization problem for the REML estimates can be solved by many statistical software, such as the `R` function `optim()`.

A.5.2 Bayesian method

In high-dimensional data analysis, it is well-known that the sample covariance is not consistent if the dimension is close to or greater than the sample size [125, 169]. For the proposed hybrid model, the estimated overall correlation matrix using the REML method may also suffer from the 'curse of dimensionality', especially when the number of factors is large and the data are sparse. The covariance matrix of the overall effect sizes could be poorly estimated, so the 95% confidence intervals (CIs) of the overall effect sizes may have inappropriate coverage probabilities. Alternatively, the Bayesian method may provide better estimates by assigning vague priors to variance/covariance parameters; it has been used in mixed treatment comparisons in which the data are also sparse [54, 67]. The performance of the Bayesian and REML methods will be studied using simulations in Appendix A.6.

In Bayesian analysis, the inverse-Wishart prior is frequently specified for unstructured positive definite matrix [170]; however, the posterior estimates may be sensitive to the selection of hyperparameters for the inverse-Wishart prior [171, 172]. Instead, we consider the separation strategy to specify vague priors for the variance and correlation components separately [54, 173]. Again, the aforementioned spherical decomposition of the marginal correlation matrix $\mathbf{R}$ is applied to guarantee its positive definiteness. Specifically, for each $j = 1, \ldots, p$, we use vague priors $N(0, 10^3)$ for the fixed effects μ_j and uniform priors $U(0, 10)$ for the between-study standard deviations ψ_j [174]. For the correlation matrix that is parameterized using the angle parameters θ_{ij} $(2 \leq j \leq i \leq p)$, we specify uniform priors $\theta_{ij} \sim U(0, \pi)$. We implement the Bayesian method using the MCMC algorithm through the software `JAGS` version 4.2.0 (`http://mcmc-jags.sourceforge.net/`). The medians of posterior samples are used as the point estimate, and the 2.5% and 97.5% quantiles are used as the lower and upper bounds of the 95% credible interval, respectively.

Practitioners need to be cautious for the convergence of the hybrid model. The REML method may fail to converge when the dimension of factors is high and the sample size is small. An estimated overall correlation close to ±1 may lead to poor convergence and unstable estimated covariance of the overall effect sizes [121]. We may check the sensitivity of the REML estimates by specifying several different initial values for maximizing restricted log-likelihood; large changes of the results may indicate that the estimates are unstable, possibly due to high dimension. The Bayesian method may be preferred in such situations; this method has been popular in the literature of mixed treatment comparisons which also deal with sparse data [54,67]. When using the Bayesian method, researchers still need to pay attention on checking the stabilization and convergence of MCMC algorithm by various criteria [162,175].

A.6 Simulations for multivariate meta-analysis of multiple factors

We conducted simulations in various settings to compare the performance of the proposed hybrid model (Model H) with the ideal model that uses within-study correlations (Model M), the model that ignores within-study correlations but accounts for between-study correlations (Model M_0), and the univariate model that ignores both types of correlations (Model U). Bias and root mean squared error (RMSE) of point estimate and 95% CI/CrI coverage probability are used to evaluate the models' performance. We set the number of studies in each simulated MVMA-MF dataset to 30 and considered 5 factors in total. Without loss of generality, the true overall effect sizes of the 5 factors were set to 0, i.e., $\boldsymbol{\mu} = (0,0,0,0,0)^{\mathrm{T}}$. Also, the between-study standard deviation τ was fixed as 1 for each factor; the within-study standard deviation σ of each factor was set to 0.5, 1, or 2. These choices for σ represent different extents of heterogeneity between studies; since the between-study variance τ^2 was fixed, the studies tend to be more homogeneous as the within-study variance σ^2 increases. Moreover, we considered the exchangeable correlation structure for both the between- and within-study correlation matrices, $\mathbf{R}_{\mathrm{B}} = (r_{\mathrm{B}ij})$ and $\mathbf{R}_{\mathrm{W}} = (r_{\mathrm{W}ij})$, which are determined by the correlation parameters ρ_{B} and ρ_{W} respectively; that is, $r_{\mathrm{B}ij} = \rho_{\mathrm{B}}$ and $r_{\mathrm{W}ij} = \rho_{\mathrm{W}}$ for

$1 \leq i \neq j \leq 5$. The between-study correlation was fixed as $\rho_{\mathrm{B}} = 0.5$, and ρ_{W} was drawn from $U(0, 0.3)$, $U(0.3, 0.6)$, or $U(0.6, 0.9)$ to represent different extents of within-study correlations. Hence, the simulated studies have different marginal correlation matrices, and the settings do not favor the assumption in the proposed hybrid model. For each setting, 1000 MVMA-MF datasets were simulated using the ideal Model M, i.e., Equation (5.1), with $\mathbf{S}_i = \sigma^2 \mathbf{R}_{\mathrm{W}}$ and $\mathbf{T} = \tau^2 \mathbf{R}_{\mathrm{B}}$. Finally, three scenarios of missingness were considered: (I) all 5 factors were observed in all studies, i.e., the data were complete; (II) the data of factors 1, 3, and 5 in 10 studies were missing completely at random; and (III) the smallest 10 effect sizes of factors 1, 3, and 5 were missing. The missingness that is not at random in scenario (III) can be considered as the effect of publication bias. Moreover, we also considered a missingness scenario that is similar to (III) but contains more missing values: (III′) the smallest 25 effect sizes of factors 1, 3, and 5 were missing; in this case, the three factors were only available from 5 studies, so the simulated MVMA-MF dataset was much sparser than the previous settings. Both the REML and Bayesian methods were applied to implement the four models. For the Bayesian method, the results of each simulated MVMA-MF dataset was based on one chain with a run of 10,000 updates after a 10,000-run burn-in period.

To save space, here we present the results of factors 1, 2, and 3 in some settings in Table A.5; the results of factors 4 and 5 are fairly similar to those of factors 2 and 1, respectively. First, recall that the data of factors 2 and 4 were complete under each scenario; their corresponding results produced by the four models are almost the same. All models lead to nearly unbiased estimated effect sizes and proper 95% CI/CrI coverage probabilities for these two factors under each scenario. Second, the results of factors 1, 3, and 5 produced by the four models differ little when the data are complete under scenario (I) or missing completely at random under scenario (II); this is expected from the perspective of missing data analysis [176]. Third, if the missingness is not at random under scenario (III), the results produced by the four models are noticeably different. The univariate model leads to the largest bias and RMSE and the lowest 95% CI/CrI coverage probability. Since Model M_0 still accounts for the between-study correlations, its performance is similar to the proposed Model H when ρ_{W} is small compared to ρ_{B}. However, the proposed Model H outperforms Model M_0 when ρ_{W} is larger than ρ_{B}. This is due to that the within-study level dominates the estimation of the overall effect sizes in such a situation, but Model M_0 ignores correlations at this level. Finally, note that Model M is ideal as it uses the within-study correlations that are usually unavailable in real data. Although Model H does not use the within-study correlations, Table A.5 shows that the biases and RMSEs produced by Model H are fairly close to the ideal Model M across various settings. Also, the 95% CI/CrI coverage probability produced by Model H is generally higher than those produced by Models M_0 and U.

In most situations, the biases and RMSEs of point estimates obtained using the REML method are close to those obtained using the Bayesian method. However, under the missingness scenarios (III) and (III′), the 95% CrI coverage probabilities for factors 1, 3, and 5 obtained using the Bayesian method are generally higher than those obtained using the REML method. As noted in Section 5.3.3, this may be due to that the estimated covariance matrix is inconsistent when many observations are missing and the dimension is close to the sample size. In addition, when using the hybrid model to analyze the data under scenario (III′), the optimization algorithm for the REML estimates did not converge for many simulated replicates, likely due to the sparsity of the data (only five samples observed for each of factors 1, 3, and 5). We ran around 2500 iterations to obtain 1000 datasets that produced converged REML estimates, and these 1000 datasets were used to produce the results in Table A.5 for scenario (III′). Also, under this scenario, the biases and RMSEs produced by the hybrid model using the REML method are noticeably different from those using the Bayesian method for factors 1, 3, and 5; again, the differences may be caused by the poor estimates of the REML method for sparse data. Hence, the Bayesian method is possibly preferred to implement the multivariate hybrid model when the dimension of factors is high but the number of observations is limited.

A.7 Estimating population-averaged absolute risks for the smoking cessation data

Consider using both the arm-based and contrast-based models to estimate absolute risks for the smoking cessation data presented by Hasselblad [132] and Lu and Ades [52]. This network meta-analysis dataset consists of 24 studies on a total of 16,737 participants, comparing the effects of self-help (B), individual counseling (C), and group counseling (D) vs. no contact (A). It is straightforward to estimate the population-averaged absolute risks using the arm-based model.2 To illustrate that the arm-based model does not simply estimate the population-averaged absolute risks for each treatment arm independently, we also consider separate logit and probit random effects models on each treatment to estimate the corresponding population-averaged absolute risks. Specifically, the random effects model for a treatment is $y_i \sim bin(n_i, p_i)$, $g(p_i) = u + v_i$, $v_i \sim N(0, \sigma^2)$, where i indexes different studies, and y_i and n_i represent the number of events and participants on a given treatment arm. We used a vague prior for the fixed effect u and an inverse gamma prior for the variance σ^2. The link function $g(\cdot)$ is either the logit or probit link. The treatment's population-averaged absolute risk can be estimated for the logit link as $\frac{1}{1+\exp(-u/\sqrt{1+C^2\sigma^2})}$ where $C = \frac{16\sqrt{3}}{15\pi}$, and for the probit link as $\Phi(u/\sqrt{1+\sigma^2})$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function [177].

To estimate absolute risks using contrast-based NMA, we first selected a reference treatment group and used the above logit random effects model to estimate absolute risk distribution for the reference group, which was further used to estimate the population-averaged absolute risks of other treatment groups. The related `WinBUGS` code for the 'random effect models for multiple arm trials' is available at `http://www.bristol.ac.uk/social-community-medicine/projects/mpes/mtc/`. Specifically, based on separate logit random effects models on each treatment, we used $N(-2.62, 2.68^{-1})$ as the prior of logit absolute risks for treatment A, $N(-1.94, 1.23^{-1})$ for treatment B, $N(-1.69, 1.69^{-1})$ for treatment C, and $N(-1.44, 1.51^{-1})$ for treatment D, when each is chosen as the reference group, respectively. The results were based on 500,000 MCMC iterations with 500,000 additional burn-in iterations, and are listed in Table A.6.

Table A.6 illustrates differences between the population-averaged absolute risks estimated by the arm-based model and by separate logit or probit models. In particular, because the arm-based approach models the absolute risks of treatment arms jointly to account for correlations among them within a study, the posterior of population-averaged absolute risks from the arm-based model generally have narrower 95% credible intervals than those from separate models. In addition, the contrast-based model leads to much wider 95% credible intervals. This may arise because the contrast-based model only uses the point estimates of u and σ^2 from the separate logit/probit random effects models as a 'fixed' prior distribution for the reference group, and the absolute risk estimates of other treatments greatly depend on this prior information.

A.8 The arm-based and contrast-based models

Assume that a network meta-analysis reviews I studies on K treatments, where each study investigates a subset of the K treatments. Label the studies $i = 1$ to I and the treatments $k = 1$ to K. Let T_i be the subset of the K treatments that is compared in ith study. Further, in the ith study, let n_{ik} be the number of participants allocated to treatment k ($k \in T_i$), and let y_{ik} be the number of events. For binary outcome, both types of NMA models are based on the binomial likelihood $y_{ik} \sim bin(n_{ik}, p_{ik})$ for $k \in T_i$; they differ in the way they model the underlying absolute risks p_{ik} in each study's treatment group.

The arm-based model [67] is specified as follows:

$$g(p_{ik}) = \mu_k + \nu_{ik};$$
$$(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{\Sigma}_K),$$

where $g(\cdot)$ is a link function and $\mathbf{\Sigma}_K$ is the variance-covariance matrix of the vector of random effects $(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})^{\mathrm{T}}$. Let σ_k^2, $k = 1, 2, \ldots, K$ denote the diagonal elements of Σ_K. The μ_k's are fixed effects for the treatments. Notice that the ν_{ik}'s are correlated within each study via the multivariate normal distribution; thus the arm-based model respects within-study randomization. When the link function $g(\cdot)$ is the probit link (i.e., $\Phi^{-1}(\cdot)$), the population-averaged absolute risk of treatment k is $\pi_k = \mathrm{E}[p_{ik}\mu_k, \sigma_k] = \Phi\left(\mu_k/\sqrt{1+\sigma_k^2}\right)$ [177]. With this estimate for absolute risks, we can calculate odds ratios, relative risks, and risk differences. This study uses an inverse-Wishart prior for $\mathbf{\Sigma}_K$, $\mathbf{\Sigma}_K^{-1} \sim W(\mathbf{V}_K, K)$, where $\mathbf{V}_K$ is a $K \times K$ matrix with diagonal elements 1 and off-diagonal elements 0.005. The inverse-Wishart prior is commonly used for variance-covariance matrices and is considered vague [171, 173]. Also, since it is conjugate, this prior allows some mathematical simplicity [178]. Zhang et al. [67] give practical computer code implementing MCMC for this model, and the `R` package 'pcnetmeta' (`http://cran.r-project.org/package=pcnetmeta`) provides user-friendly functions.

A popular contrast-based model proposed by Lu and Ades [51,54] specifies a baseline treatment $b(i)$ in the ith study. For convenience, we simply denote $b(i)$ as b. The Bayesian hierarchical model for this approach is

$$g(p_{ik}) = \mu_i + X_{ik}\delta_{ibk};$$
$$\delta_{ibk} \sim N(d_{bk}, \sigma^2_{bk}).$$

In this model, X_{ik} is a dummy variable taking the value 1 if $k \neq b$ and 0 if $k = b$. Also, μ_i is the baseline effect for treatment b in the ith study, and δ_{ibk} is the relative effect of treatment k compared to the baseline b on the logit scale. This model treats the μ_i's as nuisances and uses non-informative priors for them. This model focuses on the treatment contrasts δ_{ibk} and the parameter of interest is the overall relative effect $d_{hk} = d_{bk} - d_{bh}$; therefore, this model is described as contrast-based. Practical computer code is available at `http://www.bristol.ac.uk/social-community-medicine/projects/mpes/mtc/`.

This study uses the two models as described above. Since the contrast-based model cannot estimate absolute effects, some authors have proposed the so-called 'contrast-based + baseline' model [58], which is specified as

$$g(p_{ik}) = \mu_i + X_{ik}\delta_{i1k};$$
$$\delta_{i1k} \sim N(d_{1k}, \sigma^2_{1k});$$
$$\mu_i \sim N(m, \sigma^2_m).$$

Here, instead of being treated as a nuisance, μ_i is modeled as the absolute effect of the 'reference' treatment 1. The absolute effect of treatment k is estimated by $a_k = m + d_{1k}$. This model not only assumes that the relative effects δ_{i1k} are exchangeable between studies, but also requires exchangeability between studies of the absolute effect μ_i for the 'reference' treatment. Also, this model can be reduced to the arm-based model: we may rewrite $g(p_{ik}) = (m + d_{1k}) + (\widetilde{\mu}_{i1} + \widetilde{\delta}_{i,1k})$ for $k \neq 1$, where $\widetilde{\mu}_{i1} = \mu_{i1} - m \sim N(0, \tau^2_m)$ and $\widetilde{\delta}_{i,1k} = \delta_{i,1k} - d_{1k} \sim N(0, \sigma^2_{1k})$. Therefore, $m + d_{1k}$ and $\widetilde{\mu}_{i1} + \widetilde{\delta}_{i,1k}$ correspond to the treatment-specific fixed effect μ_k and the random effect ν_{ik} in the arm-based model, respectively. More details of NMA models can be found in the work by Hong et al. [150] with discussion [179, 180]. Due to its similarity to the arm-based model and its strong assumptions, this study does not consider the 'contrast-based + baseline' model.

A.9 An example of excluding a treatment to form a reduced network

Table A.7 gives an example to illustrate the data available for arm-based and contrast-based models after excluding each treatment being considered. This network consists of six studies labeled A to F, which evaluate the efficacy of three treatments, labeled 1 to 3. Studies A to E are two-armed, while Study F is three-armed. The six rightmost columns show data that can be used by the arm-based and contrast-based methods if each of the three treatments is excluded. For the arm-based model, in the complete network and reduced networks each retained treatment is investigated in at least 3 studies. However, for the contrast-based model only exclusion of treatment 3 is eligible for consideration under our criteria (i.e., that each retained treatment is investigated in at least 3 studies).

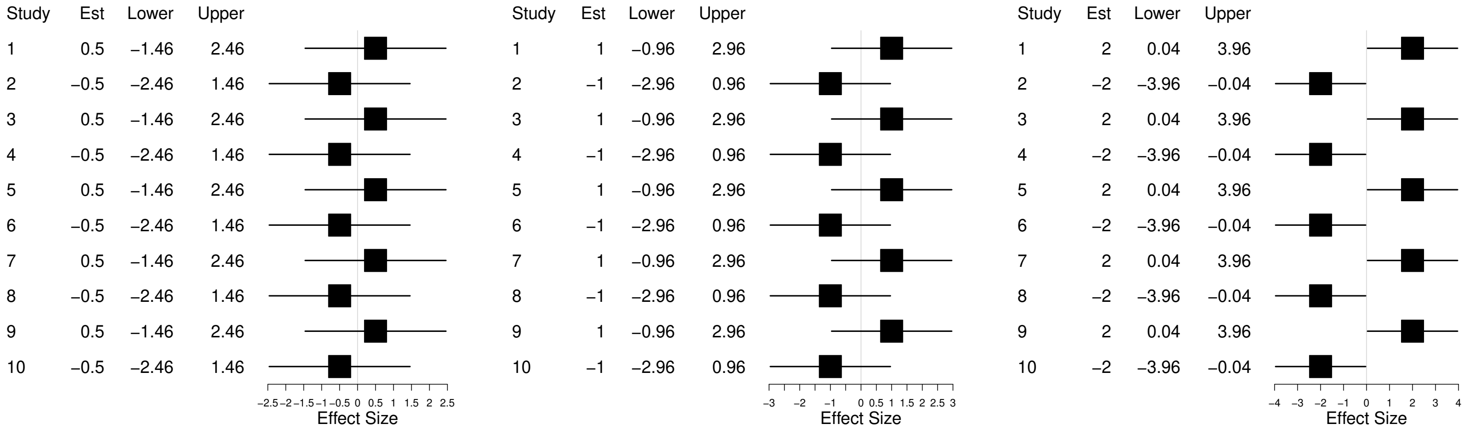
The foregoing criterion would imply that no treatment exclusions need to be ruled out for the arm-based method. However, certain treatment exclusions, such as the removal of treatment placebo in *Eisenberg 2008*, would create a disconnected network. Therefore, we ruled out such treatment exclusions in the present study.

A.10 Additional simulations comparing pairwise and network meta-analyses in the absence of evidence cycles

Tables A.8 and A.9 present the simulation results for Cases (i) and (ii) of the five treatments' heterogeneity standard deviations, respectively.

A.11 Complete data for the network meta-analysis of non-acute coronary artery disease

Table A.10 presents the dataset of the network meta-analysis performed by Trikalinos et al. [140]. It investigates the effects of four treatments for non-acute coronary artery disease.



(a) $I^2 = I_r^2 = I_m^2 = 0$.

(b) $I^2 = 0.10$, $I_r^2 = 0.43$, $I_m^2 = 0.36$.

(c) $I^2 = 0.78$, $I_r^2 = 0.86$, $I_m^2 = 0.84$.

Figure A.1: Forest plots of the three artificial meta-analyses. The column 'Est' contains the observed effect size in each study; the columns 'Lower' and 'Upper' contain the lower and upper bounds of the corresponding 95% CI.

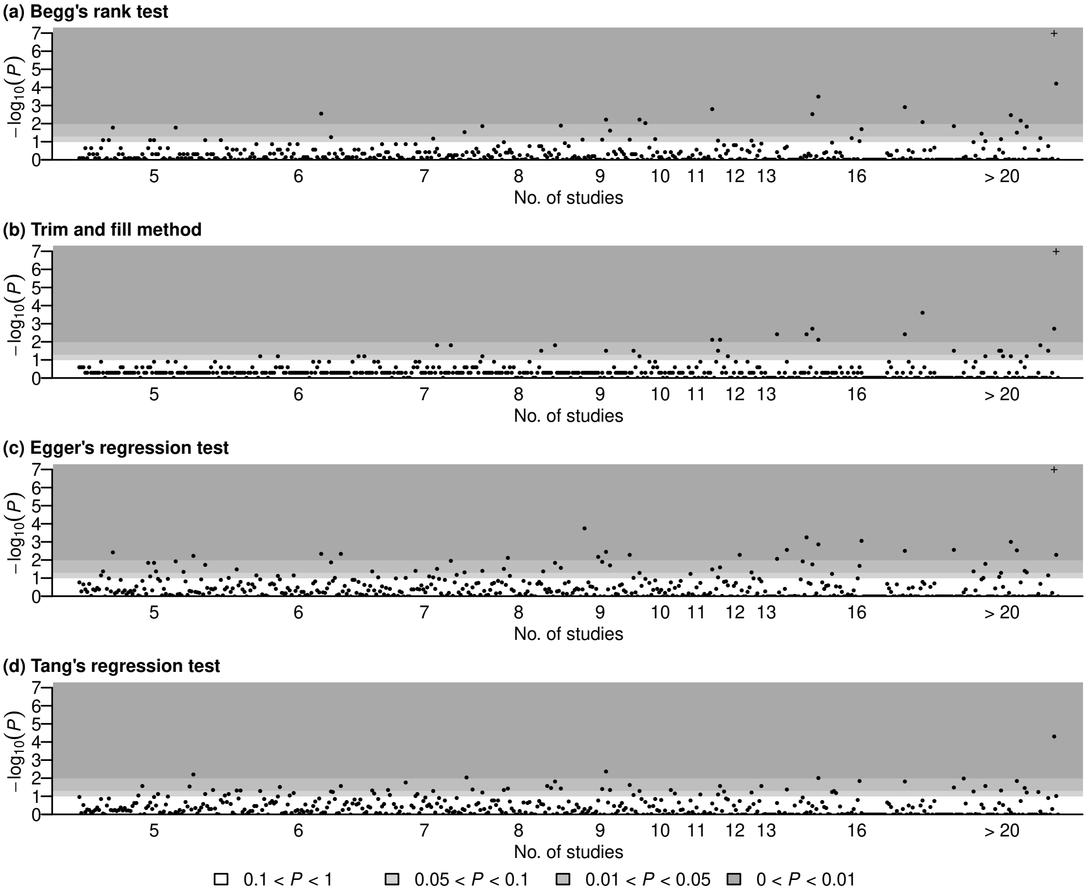


Figure A.2: The P-values produced by the various publication bias tests for the 499 Cochrane meta-analyses with continuous outcomes in the restricted dataset. Plus signs indicate P-values $< 10^{-7}$.

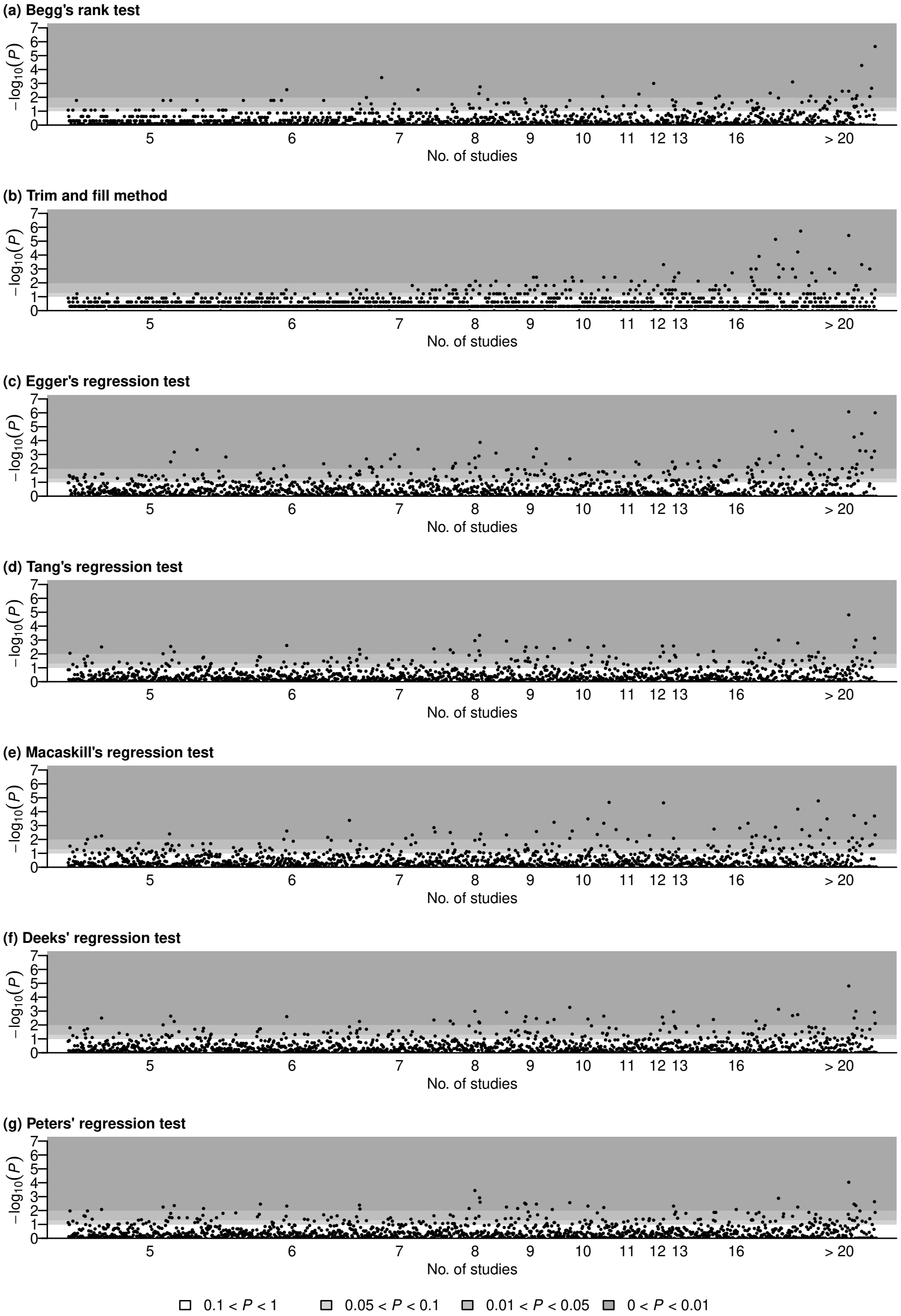


Figure A.3: The P-values produced by the various publication bias tests for the 1380 Cochrane meta-analyses with binary outcomes in the restricted dataset. Plus signs indicate P-values $< 10^{-7}$.

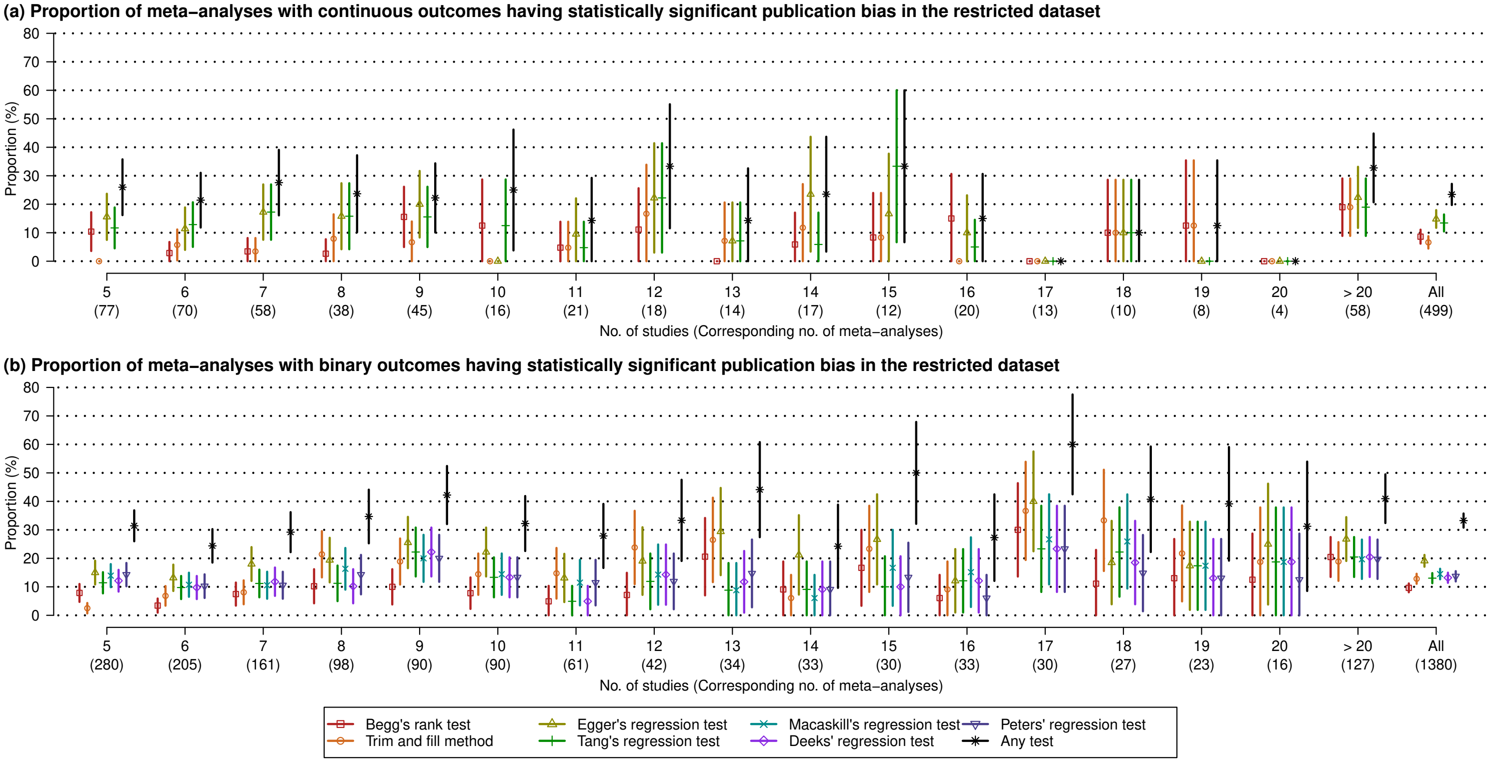


Figure A.4: Proportions of the Cochrane meta-analyses having statistically significant publication bias (P-value < 0.1) based on the various tests in the restricted dataset and their 95% confidence intervals. 'Any test' implies the proportion of having statistically significant publication bias detected by at least one test. The label 'All' on the horizontal axis represents all extracted meta-analyses with continuous/binary outcomes.

Table A.1: Sensitivity analysis on the choice of ϵ for the weighted-median-based heterogeneity measures.

ϵ	Q_m	P-value	$\widehat{\tau}_m$ (95% CI)	H_m (95% CI)	I_m^2 (95% CI)
10^{-2}	31.340	0.006	0.298 (0, 0.561)	1.354 (1, 1.884)	0.455 (0, 0.718)
10^{-3}	31.273	0.006	0.296 (0, 0.563)	1.352 (1, 1.886)	0.453 (0, 0.719)
10^{-4}	31.259	0.006	0.296 (0, 0.563)	1.351 (1, 1.886)	0.452 (0, 0.719)
10^{-5}	31.259	0.006	0.296 (0, 0.563)	1.351 (1, 1.886)	0.452 (0, 0.719)

Table A.2: Type I error rates and powers of three heterogeneity tests for the simulated meta-analyses containing 10 studies with outliers in Scenario I.

Outlier pattern	Size/power[†]			RMSE			CP (%)		
	$Q^{‡}$	Q_r	Q_m	$\widehat{\tau}^2_{\mathrm{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$	$\widehat{\tau}^2_{\mathrm{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$
Scenario I (contamination) with $\tau^2 = 0$ (homogeneity) and $s_i \sim U(0.5, 1)$:									
No outliers	0.04 (0.05)	0.04	0.04	0.18	0.22	0.16	100	100	100
C	0.74 (0.74)	0.53	0.47	1.04	0.80	0.56	99	98	100
(C, C)	0.97 (0.97)	0.93	0.91	1.70	1.68	1.16	92	92	98
$(C, -C)$	0.98 (0.98)	0.93	0.92	2.14	1.55	1.24	91	91	97
(C, C, C)	0.99 (0.99)	0.99	0.99	2.21	2.66	1.91	48	55	78
$(C, C, -C)$	1.00 (1.00)	0.99	1.00	3.01	2.57	2.07	48	57	79
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(0.5, 1)$:									
No outliers	0.75 (0.74)	0.72	0.70	0.72	0.82	0.71	76	88	82
C	0.99 (0.98)	0.97	0.97	2.24	1.89	1.37	98	98	99
(C, C)	1.00 (1.00)	1.00	1.00	3.57	3.58	2.59	92	93	98
$(C, -C)$	1.00 (1.00)	1.00	1.00	4.44	3.49	2.77	92	92	98
(C, C, C)	1.00 (1.00)	1.00	1.00	4.47	5.29	3.94	67	71	88
$(C, C, -C)$	1.00 (1.00)	1.00	1.00	6.35	5.60	4.54	63	70	86
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(1, 2)$:									
No outliers	0.26 (0.26)	0.22	0.22	1.27	1.52	1.21	76	88	81
C	0.88 (0.89)	0.78	0.75	5.34	4.28	3.06	98	98	99
(C, C)	0.99 (0.99)	0.98	0.98	8.73	8.68	6.15	92	92	98
$(C, -C)$	1.00 (1.00)	0.99	0.99	10.81	8.09	6.48	91	92	97
(C, C, C)	1.00 (1.00)	1.00	1.00	11.02	13.15	9.66	56	61	83
$(C, C, -C)$	1.00 (1.00)	1.00	1.00	15.66	13.46	10.97	56	62	82
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(2, 5)$:									
No outliers	0.08 (0.08)	0.08	0.07	4.23	5.28	3.77	77	87	82
C	0.81 (0.81)	0.64	0.60	27.07	20.71	14.06	98	98	100
(C, C)	0.98 (0.98)	0.96	0.95	44.75	44.94	30.35	90	90	97
$(C, -C)$	1.00 (1.00)	0.96	0.96	55.85	39.94	31.44	90	91	97
(C, C, C)	1.00 (1.00)	1.00	1.00	56.81	69.04	49.87	44	53	75
$(C, C, -C)$	1.00 (1.00)	1.00	1.00	81.37	68.52	55.13	45	54	74

RMSE: root mean squared error; CP: coverage probability of 95% confidence interval.

[†] Size (type I error rate) for homogeneous studies ($\tau^2 = 0$) and power for heterogeneous studies ($\tau^2 > 0$) at the significance level $\alpha = 0.05$.

[‡] The sizes/powers outside the parentheses are produced by the resampling method; those inside the parentheses are obtained using Q's theoretical distribution under the null hypothesis.

Table A.3: Type I error rates and powers of three heterogeneity tests for the simulated meta-analyses containing 30 studies with outliers in Scenario I.

Outlier pattern	Size/power†			RMSE			CP (%)		
	$Q^{\ddagger}$	Q_r	Q_m	$\widehat{\tau}^2_{\mathrm{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$	$\widehat{\tau}^2_{\mathrm{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$
Scenario I (contamination) with $\tau^2 = 0$ (homogeneity) and $s_i \sim U(0.5, 1)$:									
No outliers	0.05 (0.06)	0.05	0.05	0.10	0.12	0.10	98	99	99
C	0.55 (0.55)	0.27	0.25	0.37	0.24	0.20	97	97	98
(C, C)	0.89 (0.89)	0.66	0.60	0.63	0.42	0.35	88	90	94
$(C, -C)$	0.92 (0.92)	0.61	0.61	0.68	0.40	0.36	89	90	94
(C, C, C)	0.98 (0.98)	0.90	0.87	0.88	0.64	0.53	65	74	83
$(C, C, -C)$	0.99 (0.98)	0.89	0.88	0.99	0.61	0.55	64	73	83
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(0.5, 1)$:									
No outliers	0.98 (0.99)	0.98	0.98	0.40	0.43	0.41	88	93	91
C	1.00 (1.00)	1.00	1.00	0.84	0.63	0.55	97	97	98
(C, C)	1.00 (1.00)	1.00	1.00	1.37	1.00	0.85	93	94	96
$(C, -C)$	1.00 (1.00)	1.00	1.00	1.45	0.97	0.85	93	94	96
(C, C, C)	1.00 (1.00)	1.00	1.00	1.86	1.44	1.22	76	83	90
$(C, C, -C)$	1.00 (1.00)	1.00	1.00	2.05	1.40	1.25	77	84	91
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(1, 2)$:									
No outliers	0.48 (0.49)	0.42	0.43	0.74	0.81	0.75	89	93	91
C	0.89 (0.89)	0.78	0.77	1.97	1.36	1.17	98	97	98
(C, C)	0.99 (0.99)	0.94	0.94	3.33	2.29	1.93	91	92	96
$(C, -C)$	0.99 (0.99)	0.94	0.94	3.50	2.17	1.93	91	92	96
(C, C, C)	1.00 (1.00)	0.99	0.99	4.60	3.41	2.85	70	80	88
$(C, C, -C)$	1.00 (1.00)	0.99	0.99	5.03	3.24	2.90	71	81	88
Scenario I (contamination) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(2, 5)$:									
No outliers	0.11 (0.11)	0.09	0.09	2.32	2.64	2.25	88	92	91
C	0.70 (0.70)	0.43	0.41	9.89	5.96	5.02	97	97	99
(C, C)	0.96 (0.96)	0.81	0.78	17.19	10.92	8.97	90	91	94
$(C, -C)$	0.96 (0.96)	0.76	0.77	18.10	10.02	8.94	90	91	95
(C, C, C)	1.00 (1.00)	0.95	0.94	23.87	16.90	13.59	65	74	82
$(C, C, -C)$	1.00 (1.00)	0.95	0.95	26.10	15.49	13.78	64	74	83

RMSE: root mean squared error; CP: coverage probability of 95% confidence interval.

† Size (type I error rate) for homogeneous studies ($\tau^2 = 0$) and power for heterogeneous studies ($\tau^2 > 0$) at the significance level $\alpha = 0.05$.

‡ The sizes/powers outside the parentheses are produced by the resampling method; those inside the parentheses are obtained using Q's theoretical distribution under the null hypothesis.

Table A.4: Powers of three heterogeneity tests for the simulated meta-analyses containing 30 studies with outliers in Scenario II.

Outlier pattern	Power			RMSE			CP (%)		
	$Q^{\ddagger}$	Q_r	Q_m	$\widehat{\tau}^2_{\text{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$	$\widehat{\tau}^2_{\text{DL}}$	$\widehat{\tau}^2_r$	$\widehat{\tau}^2_m$
Scenario II (heavy tail) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(0.5, 1)$:									
df = 3	0.92 (0.92)	0.89	0.88	1.45	0.59	0.56	72	79	73
df = 5	0.98 (0.98)	0.95	0.95	0.55	0.45	0.45	84	90	86
df = 10	0.98 (0.98)	0.97	0.97	0.43	0.43	0.42	88	93	90
Scenario II (heavy tail) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(1, 2)$:									
df = 3	0.41 (0.40)	0.35	0.35	1.53	0.88	0.82	83	90	87
df = 5	0.46 (0.46)	0.40	0.40	0.82	0.82	0.77	88	93	90
df = 10	0.48 (0.49)	0.42	0.42	0.76	0.82	0.77	88	94	90
Scenario II (heavy tail) with $\tau^2 = 1$ (heterogeneity) and $s_i \sim U(2, 5)$:									
df = 3	0.10 (0.10)	0.10	0.10	2.66	2.71	2.33	88	92	91
df = 5	0.09 (0.09)	0.09	0.08	2.18	2.54	2.17	88	93	92
df = 10	0.10 (0.10)	0.08	0.09	2.18	2.48	2.12	88	93	91

RMSE: root mean squared error; CP: coverage probability of 95% confidence interval.

[‡] The sizes/powers outside the parentheses are produced by the resampling method; those inside the parentheses are obtained using Q's theoretical distribution under the null hypothesis.

Table A.5: The simulation results produced by Models M, H, M_0, and U in various settings of within-study variances and correlations under different missingness scenarios.

Model	Factor 1			Factor 2			Factor 3		
	Bias	RMSE	CP (%)	Bias	RMSE	CP (%)	Bias	RMSE	CP (%)
$\tau = 1$, $\sigma = 1$, $\rho_B = 0.5$, $\rho_W \sim U(0.6, 0.9)$:									
No missing data (I)									
M	0.00 (0.00)	0.26 (0.26)	94 (94)	0.00 (0.00)	0.26 (0.26)	93 (92)	0.00 (0.00)	0.25 (0.25)	95 (94)
H	0.00 (0.00)	0.26 (0.26)	94 (95)	0.00 (0.00)	0.26 (0.26)	93 (95)	0.00 (0.00)	0.25 (0.25)	95 (96)
M_0	0.00 (0.00)	0.26 (0.26)	95 (97)	0.00 (0.00)	0.26 (0.26)	94 (96)	0.00 (0.00)	0.25 (0.25)	96 (97)
U	0.00 (0.00)	0.26 (0.26)	94 (95)	0.00 (0.00)	0.26 (0.26)	94 (94)	0.00 (0.00)	0.25 (0.25)	95 (96)
Factors 1, 3, and 5 in 10 studies are missing completely at random (II)									
M	0.00 (0.00)	0.30 (0.29)	93 (95)	0.00 (0.00)	0.26 (0.26)	94 (93)	0.00 (0.01)	0.30 (0.28)	93 (95)
H	0.00 (0.00)	0.31 (0.29)	92 (96)	0.00 (0.00)	0.26 (0.26)	93 (95)	0.00 (0.01)	0.30 (0.29)	92 (96)
M_0	0.00 (0.00)	0.30 (0.29)	96 (97)	0.00 (0.00)	0.26 (0.26)	94 (95)	0.00 (0.01)	0.29 (0.28)	95 (98)
U	−0.01 (0.00)	0.32 (0.32)	94 (95)	0.00 (0.00)	0.26 (0.26)	94 (94)	−0.01 (0.02)	0.32 (0.31)	94 (96)
Factors 1, 3, and 5 in 10 studies are missing not at random (III)									
M	0.46 (0.48)	0.54 (0.56)	51 (54)	0.00 (0.00)	0.26 (0.26)	94 (94)	0.46 (0.48)	0.54 (0.55)	48 (52)
H	0.49 (0.51)	0.57 (0.58)	43 (54)	0.00 (0.00)	0.26 (0.26)	94 (96)	0.48 (0.52)	0.56 (0.59)	41 (49)
M_0	0.56 (0.58)	0.62 (0.64)	44 (53)	0.00 (0.00)	0.26 (0.26)	94 (95)	0.55 (0.61)	0.62 (0.67)	41 (46)
U	0.75 (0.75)	0.80 (0.80)	16 (19)	0.00 (0.00)	0.26 (0.26)	94 (94)	0.75 (0.75)	0.80 (0.80)	14 (18)
$\tau = 1$, $\sigma = 1$, $\rho_B = 0.5$, $\rho_W \sim U(0, 0.3)$:									
Factors 1, 3, and 5 in 10 studies are missing not at random (III)									
M	0.68 (0.69)	0.74 (0.74)	23 (26)	0.01 (0.01)	0.26 (0.26)	94 (95)	0.68 (0.70)	0.73 (0.75)	22 (23)
H	0.67 (0.68)	0.73 (0.74)	21 (30)	0.01 (0.01)	0.26 (0.26)	94 (96)	0.67 (0.69)	0.73 (0.74)	21 (27)
M_0	0.69 (0.71)	0.75 (0.76)	23 (26)	0.01 (0.01)	0.26 (0.26)	94 (95)	0.70 (0.72)	0.75 (0.77)	21 (24)
U	0.75 (0.75)	0.80 (0.80)	16 (19)	0.01 (0.01)	0.26 (0.26)	94 (95)	0.75 (0.75)	0.80 (0.80)	13 (16)
$\tau = 1$, $\sigma = 2$, $\rho_B = 0.5$, $\rho_W \sim U(0.6, 0.9)$:									
Factors 1, 3, and 5 in 10 studies are missing not at random (III)									
M	0.58 (0.57)	0.74 (0.73)	70 (74)	0.01 (0.01)	0.42 (0.41)	96 (96)	0.59 (0.59)	0.74 (0.74)	71 (73)
H	0.60 (0.63)	0.76 (0.78)	66 (74)	0.01 (0.01)	0.42 (0.42)	96 (97)	0.61 (0.65)	0.76 (0.80)	65 (73)
M_0	0.90 (0.96)	1.00 (1.05)	56 (62)	0.01 (0.01)	0.42 (0.42)	97 (97)	0.90 (1.01)	1.00 (1.10)	55 (57)
U	1.18 (1.18)	1.26 (1.26)	25 (29)	0.01 (0.01)	0.42 (0.42)	94 (96)	1.19 (1.19)	1.27 (1.27)	24 (28)
Factors 1, 3, and 5 in 25 studies are missing not at random (III′)									
M	1.60 (1.57)	1.76 (1.72)	32 (76)	−0.01 (0.01)	0.41 (0.42)	95 (95)	1.60 (1.58)	1.77 (1.73)	31 (75)
H	1.46 (2.36)	1.64 (2.75)	26 (65)	−0.02 (0.01)	0.43 (0.42)	93 (96)	1.47 (2.45)	1.67 (2.73)	27 (65)
M_0	2.92 (3.05)	2.99 (3.11)	4 (73)	−0.01 (0.01)	0.41 (0.42)	96 (96)	2.95 (3.14)	3.02 (3.20)	3 (62)
U	3.19 (3.21)	3.26 (3.28)	1 (11)	−0.01 (0.01)	0.41 (0.42)	94 (95)	3.21 (3.23)	3.27 (3.30)	1 (9)

The results outside parentheses are obtained using the REML method; those inside parentheses are obtained using the Bayesian method. RMSE: root mean square error; CP: 95% CI/CrI coverage probability.

Table A.6: Population-averaged absolute risks of the four treatments in the smoking cessation network meta-analysis. They are obtained by the arm-based model, contrast-based model using different reference treatments, and separate logit/probit random effects models on each treatment.

	Population-averaged absolute risks (posterior mean with 95% credible intervals)						
Treatment	Contrast-based model				Separate logit random effects models	Separate probit random effects models	Arm-based model (using probit link)
	Reference treatment (# of studies including this treatment)						
	A (19)	B (6)	C (19)	D (6)			
A	0.078 (0.021, 0.194)	0.110 (0.012, 0.378)	0.093 (0.016, 0.280)	0.098 (0.013, 0.325)	0.075 (0.055, 0.104)	0.072 (0.045, 0.108)	0.083 (0.058, 0.117)
B	0.126 (0.027, 0.334)	0.156 (0.024, 0.456)	0.144 (0.023, 0.422)	0.147 (0.020, 0.454)	0.174 (0.084, 0.352)	0.182 (0.069, 0.384)	0.170 (0.086, 0.304)
C	0.162 (0.044, 0.379)	0.207 (0.028, 0.587)	0.180 (0.039, 0.455)	0.188 (0.029, 0.525)	0.175 (0.128, 0.241)	0.173 (0.118, 0.245)	0.185 (0.135, 0.248)
D	0.203 (0.048, 0.491)	0.248 (0.034, 0.665)	0.225 (0.042, 0.574)	0.220 (0.046, 0.540)	0.231 (0.125, 0.403)	0.244 (0.106, 0.450)	0.233 (0.127, 0.382)

Table A.7: An example for treatment exclusion in network meta-analysis.

		Full network	Usable data					
Study	Treatment	(no. of events /	Remove Treatment 1		Remove Treatment 2		Remove Treatment 3	
ID	ID	no. of	Arm-	Contrast-	Arm-	Contrast-	Arm-	Contrast-
		participants)	based	based	based	based	based	based
A	1	$y_{\mathrm{A1}}/n_{\mathrm{A1}}$	–	–	$y_{\mathrm{A1}}/n_{\mathrm{A1}}$	–	$y_{\mathrm{A1}}/n_{\mathrm{A1}}$	$y_{\mathrm{A1}}/n_{\mathrm{A1}}$
A	2	$y_{\mathrm{A2}}/n_{\mathrm{A2}}$	$y_{\mathrm{A2}}/n_{\mathrm{A2}}$	–	–	–	$y_{\mathrm{A2}}/n_{\mathrm{A2}}$	$y_{\mathrm{A2}}/n_{\mathrm{A2}}$
B	1	$y_{\mathrm{B1}}/n_{\mathrm{B1}}$	–	–	$y_{\mathrm{B1}}/n_{\mathrm{B1}}$	–	$y_{\mathrm{B1}}/n_{\mathrm{B1}}$	$y_{\mathrm{B1}}/n_{\mathrm{B1}}$
B	2	$y_{\mathrm{B2}}/n_{\mathrm{B2}}$	$y_{\mathrm{B2}}/n_{\mathrm{B2}}$	–	–	–	$y_{\mathrm{B2}}/n_{\mathrm{B2}}$	$y_{\mathrm{B2}}/n_{\mathrm{B2}}$
C	1	$y_{\mathrm{C1}}/n_{\mathrm{C1}}$	–	–	$y_{\mathrm{C1}}/n_{\mathrm{C1}}$	–	$y_{\mathrm{C1}}/n_{\mathrm{C1}}$	$y_{\mathrm{C1}}/n_{\mathrm{C1}}$
C	2	$y_{\mathrm{C2}}/n_{\mathrm{C2}}$	$y_{\mathrm{C2}}/n_{\mathrm{C2}}$	–	–	–	$y_{\mathrm{C2}}/n_{\mathrm{C2}}$	$y_{\mathrm{C2}}/n_{\mathrm{C2}}$
D	2	$y_{\mathrm{D2}}/n_{\mathrm{D2}}$	$y_{\mathrm{D2}}/n_{\mathrm{D2}}$	$y_{\mathrm{D2}}/n_{\mathrm{D2}}$	–	–	$y_{\mathrm{D2}}/n_{\mathrm{D2}}$	–
D	3	$y_{\mathrm{D3}}/n_{\mathrm{D3}}$	$y_{\mathrm{D3}}/n_{\mathrm{D3}}$	$y_{\mathrm{D3}}/n_{\mathrm{D3}}$	$y_{\mathrm{D3}}/n_{\mathrm{D3}}$	–	–	–
E	1	$y_{\mathrm{E1}}/n_{\mathrm{E1}}$	–	–	$y_{\mathrm{E1}}/n_{\mathrm{E1}}$	$y_{\mathrm{E1}}/n_{\mathrm{E1}}$	$y_{\mathrm{E1}}/n_{\mathrm{E1}}$	–
E	3	$y_{\mathrm{E3}}/n_{\mathrm{E3}}$	$y_{\mathrm{E3}}/n_{\mathrm{E3}}$	–	$y_{\mathrm{E3}}/n_{\mathrm{E3}}$	$y_{\mathrm{E3}}/n_{\mathrm{E3}}$	–	–
F	1	$y_{\mathrm{F1}}/n_{\mathrm{F1}}$	–	–	$y_{\mathrm{F1}}/n_{\mathrm{F1}}$	$y_{\mathrm{F1}}/n_{\mathrm{F1}}$	$y_{\mathrm{F1}}/n_{\mathrm{F1}}$	$y_{\mathrm{F1}}/n_{\mathrm{F1}}$
F	2	$y_{\mathrm{F2}}/n_{\mathrm{F2}}$	$y_{\mathrm{F2}}/n_{\mathrm{F2}}$	$y_{\mathrm{F2}}/n_{\mathrm{F2}}$	–	–	$y_{\mathrm{F2}}/n_{\mathrm{F2}}$	$y_{\mathrm{F2}}/n_{\mathrm{F2}}$
F	3	$y_{\mathrm{F3}}/n_{\mathrm{F3}}$	$y_{\mathrm{F3}}/n_{\mathrm{F3}}$	$y_{\mathrm{F3}}/n_{\mathrm{F3}}$	$y_{\mathrm{F3}}/n_{\mathrm{F3}}$	$y_{\mathrm{F3}}/n_{\mathrm{F3}}$	–	–

Table A.8: Biases (outside brackets), mean squared errors (inside parentheses), and 95% credible interval coverage probabilities (%, inside square brackets) of the estimated relative effects produced by the Smith model (pairwise meta-analysis) and the Lu–Ades model (network meta-analysis) in simulations. The data were simulated assuming that treatment effects were homogeneous across studies.

Network	Treatment	Network meta-analysis			Pairwise meta-analysis		
shape	contrast	FE	RE1	RE2	FE	RE1	RE2
Shape 1	d_{12}	−0.01	−0.01	0.00	0.00	−0.01	0.00
		(0.10)	(0.11)	(0.10)	(0.10)	(0.11)	(0.10)
		[96]	[100]	[98]	[96]	[100]	[98]
	d_{15}	0.00	0.00	0.00	0.00	0.00	0.00
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
		[95]	[98]	[97]	[95]	[98]	[97]
	d_{23}	0.01	0.01	0.00	0.00	0.01	0.00
		(0.15)	(0.16)	(0.15)	(0.15)	(0.16)	(0.15)
		[96]	[100]	[97]	[96]	[100]	[97]
	d_{45}	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
		[94]	[98]	[97]	[94]	[98]	[97]
Shape 2	d_{12}	0.01	0.00	0.00	0.00	−0.01	0.00
		(0.10)	(0.11)	(0.10)	(0.10)	(0.11)	(0.10)
		[97]	[100]	[98]	[96]	[100]	[98]
	d_{13}	0.02	0.00	0.00	0.00	−0.01	0.00
		(0.14)	(0.15)	(0.14)	(0.14)	(0.15)	(0.14)
		[96]	[100]	[97]	[96]	[100]	[97]
	d_{15}	0.02	0.00	0.01	0.00	0.00	0.01
		(0.19)	(0.20)	(0.19)	(0.19)	(0.20)	(0.19)
		[97]	[100]	[97]	[96]	[100]	[98]
	d_{45}	0.00	0.00	0.00	0.00	0.00	0.00
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
		[94]	[98]	[96]	[94]	[98]	[97]
Shape 3	d_{12}	0.00	0.00	0.01	0.00	−0.01	0.00
		(0.10)	(0.11)	(0.10)	(0.10)	(0.11)	(0.10)
		[97]	[100]	[98]	[96]	[100]	[98]
	d_{13}	0.01	0.00	0.01	0.00	−0.01	0.00
		(0.14)	(0.15)	(0.14)	(0.14)	(0.15)	(0.14)
		[96]	[100]	[97]	[96]	[100]	[98]
	d_{15}	0.01	0.00	0.01	0.00	0.00	0.01
		(0.13)	(0.15)	(0.14)	(0.13)	(0.14)	(0.14)
		[96]	[99]	[96]	[96]	[100]	[97]
	d_{45}	0.00	0.00	0.00	0.00	0.00	0.00
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
		[94]	[98]	[97]	[94]	[98]	[96]

FE: fixed-effects model; RE1: random-effects model with different heterogeneity variances for different treatment contrasts; RE2: random-effects model with a common heterogeneity variance.

d_{hk}: treatment k compared to h.

Monte Carlo standard errors of all biases, mean squared errors, and coverage probabilities are less than 0.02, 0.01, and 2%, respectively.

Table A.9: Biases (outside brackets), mean squared errors (inside parentheses), and 95% credible interval coverage probabilities (%, inside square brackets) of the estimated relative effects produced by the Smith model (pairwise meta-analysis) and the Lu–Ades model (network meta-analysis) in simulations. The data were simulated using a common heterogeneity standard deviation for all treatments.

Network	Treatment	Network meta-analysis			Pairwise meta-analysis		
shape	contrast	FE	RE1	RE2	FE	RE1	RE2
Shape 1	d_{12}	-0.01^{a}	0.01^{a}	0.01^{a}	0.00^{a}	0.01^{a}	0.01^{a}
		(0.76^{e})	(0.55^{d})	(0.54^{d})	(0.76^{e})	(0.55^{d})	(0.54^{d})
		[54]	[98]	[95]	[54]	[98]	[95]
	d_{15}	0.02	0.01	0.01	0.02	0.01	0.01
		(0.24)	(0.12)	(0.12)	(0.24)	(0.12)	(0.12)
		[48]	[96]	[96]	[48]	[96]	[96]
	d_{23}	-0.02^{b}	-0.02^{a}	-0.02^{a}	-0.02^{b}	-0.02^{a}	-0.02^{a}
		(1.18^{f})	(0.80^{e})	(0.79^{e})	(1.18^{f})	(0.80^{e})	(0.79^{e})
		[53]	[99]	[95]	[53]	[98]	[95]
	d_{45}	0.01^{a}	0.00	0.00	0.01^{a}	0.00	0.00
		(0.52^{d})	(0.27)	(0.27)	(0.52^{d})	(0.27)	(0.27)
		[50]	[97]	[96]	[50]	[97]	[96]
Shape 2	d_{12}	0.01^{a}	0.00^{a}	0.01^{a}	0.00^{a}	0.01^{a}	0.01^{a}
		(0.76^{e})	(0.55^{d})	(0.54^{d})	(0.76^{e})	(0.55^{d})	(0.54^{d})
		[55]	[98]	[95]	[54]	[98]	[95]
	d_{13}	0.02^{b}	0.00^{a}	0.02^{a}	-0.01^{b}	0.01^{a}	0.01^{a}
		(1.15^{f})	(0.83^{e})	(0.82^{e})	(1.15^{f})	(0.83^{e})	(0.82^{e})
		[52]	[98]	[95]	[50]	[98]	[95]
	d_{15}	0.02^{c}	-0.01^{b}	0.00^{b}	-0.01^{c}	0.00^{b}	0.00^{b}
		(1.66^{f})	(1.17^{f})	(1.17^{f})	(1.66^{f})	(1.17^{f})	(1.16^{f})
		[50]	[98]	[95]	[49]	[98]	[95]
	d_{45}	-0.01	-0.03	-0.03	-0.01	-0.03	-0.03
		(0.23)	(0.12)	(0.12)	(0.23)	(0.12)	(0.12)
		[47]	[95]	[95]	[46]	[95]	[95]
Shape 3	d_{12}	0.00^{a}	0.00^{a}	0.01^{a}	0.00^{a}	0.01^{a}	0.01^{a}
		(0.76^{e})	(0.55^{d})	(0.54^{d})	(0.76^{e})	(0.55^{d})	(0.54^{d})
		[55]	[98]	[95]	[54]	[98]	[95]
	d_{13}	0.00^{b}	0.01^{a}	0.02^{a}	-0.01^{b}	0.01^{a}	0.01^{a}
		(1.15^{f})	(0.83^{e})	(0.82^{e})	(1.15^{f})	(0.83^{e})	(0.82^{e})
		[51]	[98]	[94]	[50]	[98]	[94]
	d_{15}	0.03^{b}	0.02^{a}	0.03^{a}	0.02^{b}	0.02^{a}	0.03^{a}
		(1.02^{e})	(0.69^{e})	(0.69^{e})	(1.02^{e})	(0.69^{e})	(0.68^{e})
		[54]	[98]	[95]	[54]	[98]	[94]
	d_{45}	-0.01	-0.03	-0.03	-0.01	-0.03	-0.03
		(0.23)	(0.12)	(0.12)	(0.23)	(0.12)	(0.12)
		[47]	[95]	[95]	[46]	[95]	[95]

FE: fixed-effects model; RE1: random-effects model with different heterogeneity variances for different treatment contrasts; RE2: random-effects model with a common heterogeneity variance.

d_{hk}: treatment k compared to h.

Monte Carlo standard error of bias: a, 0.02–0.03; b, 0.03–0.04; c, 0.04–0.05; otherwise, less than 0.02. Monte Carlo standard error of mean squared error: d, 0.02–0.03; e, 0.03–0.05; f, 0.05–0.08; otherwise, less than 0.02. Monte Carlo standard errors of all coverage probabilities are less than 2%.

Table A.10: The effects of four treatments for non-acute coronary artery disease.

Study	Treatment ID				Study	Treatment ID			
ID	1	2	3	4	ID	1	2	3	4
1	7/107	5/105			33		0/60	0/60	
2	1/44	0/44			34		0/66	0/31	
3	6/72	6/72			35		0/48	0/48	
4	10/50	9/51			36		4/189	1/192	
5	15/112	16/115			37		1/182	1/169	
6	43/514	43/504			38		1/71	1/74	
7	22/105	6/96			39		0/143	0/145	
8	1/34		2/32		40		0/195	4/393	
9	84/1084		87/1082		41		0/106	1/96	
10	0/51		0/50		42		10/111	0/110	
11	95/1138		85/1148		43		1/100	1/100	
12		8/257	15/259		44		1/22	0/23	
13		3/202	3/205		45		1/122	0/124	
14		0/59	0/58		46			1/26	0/24
15		0/42	0/42		47			0/58	1/117
16		10/60	4/60		48			0/38	1/152
17		4/410	4/413		49			5/519	5/522
18		2/176	2/178		50			0/30	0/31
19		1/59	0/57		51			0/134	0/135
20		3/54	6/56		52			2/136	0/131
21		1/30	0/30		53			8/652	9/662
22		14/796	4/794		54			8/576	7/569
23		5/223	6/229		55			2/227	0/219
24		1/208	3/202		56			1/10	0/20
25		0/55	0/55		57			4/525	7/533
26		0/43	1/42		58			5/118	9/120
27		3/365	3/370		59			1/177	2/175
28		4/322	5/286		60			0/50	0/50
29		3/200	2/204		61			5/250	7/250
30		0/196	0/192		62			1/159	1/163
31		3/146	0/154		63			1/100	0/100
32		0/126	3/125						

The outcome is death due to non-acute coronary artery disease.

Each entry shows (number of deaths)/(total number of patients).

Blank entries represent treatments that were not investigated in the corresponding studies.

Treatment IDs: (1) medical therapy; (2) percutaneous transluminal balloon coronary angioplasty; (3) bare-metal stents; and (4) drug-eluting stents.

Appendix B

Proofs of Propositions and Theorems

B.1 Asymptotic values of heterogeneity measures

The proofs will frequently use the property about the mean of folded normal distribution: if $X \sim N(\mu, \sigma^2)$, then $\mathrm{E}|X| = \sigma\sqrt{\frac{2}{\pi}}e^{-\mu^2/(2\sigma^2)} + \mu(1 - 2\Phi(-\mu/\sigma))$, where $\Phi(\cdot)$ is the cumulative density function of standard normal distribution. Let $\lfloor x \rfloor$ be the largest integer less than or equal to x.

Proof of Proposition 1. Note that $\bar{\mu} = \frac{\sum_{i=1}^n w_i y_i/n}{\sum_{i=1}^n w_i/n} \xrightarrow{P} \frac{\mathrm{E}[w_1 y_1]}{\mathrm{E}[w_1]} = \mu$, we have

$$
\begin{aligned}
Q/n &= \frac{1}{n}\sum_{i=1}^n w_i[(y_i - \mu) - (\bar{\mu} - \mu)]^2 \\
&= \frac{1}{n}\sum_{i=1}^n w_i(y_i - \mu)^2 - 2(\bar{\mu} - \mu)\cdot\frac{1}{n}\sum_{i=1}^n w_i(y_i - \mu) + (\bar{\mu} - \mu)^2\cdot\frac{1}{n}\sum_{i=1}^n w_i \\
&\xrightarrow{P} \mathrm{E}[w_1(y_1 - \mu)^2] = 1.
\end{aligned}
$$

Therefore, $I^2 = 1 - \frac{1}{Q/(n-1)} \xrightarrow{P} 0$.

For Q_r, applying the triangle inequality $|x| - |y| \leq |x - y|$, we have

$$
\begin{aligned}
\sqrt{w_i}|y_i - \bar{\mu}| - \sqrt{w_i}|y_i - \mu| \leq \sqrt{w_i}|\bar{\mu} - \mu|; \\
\sqrt{w_i}|y_i - \mu| - \sqrt{w_i}|y_i - \bar{\mu}| \leq \sqrt{w_i}|\bar{\mu} - \mu|.
\end{aligned}
$$

Averaging each of the above two inequalities for $i = 1, \dots, n$, we have

$$\left| Q_r/n - \frac{1}{n}\sum_{i=1}^{n} \sqrt{w_i}|y_i - \mu| \right| \le |\bar{\mu} - \mu| \cdot \frac{1}{n}\sum_{i=1}^{n}\sqrt{w_i} \xrightarrow{P} 0.$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{w_i}|y_i - \mu| \xrightarrow{P} \mathrm{E}[|\sqrt{w_1}(y_1 - \mu)|] = \sqrt{2/\pi}.$$

Therefore, $Q_r/n \xrightarrow{P} \sqrt{2/\pi}$, and $I_r^2 = 1 - \frac{n-1}{n} \cdot \frac{2/\pi}{(Q_r/n)^2} \xrightarrow{P} 0$.

For Q_m, by the theory of M-estimation [73], the weighted median $\widehat{\mu}_m \xrightarrow{P} \mu$. Similarly applying the triangle inequality, we have

$$\left| Q_m/n - \frac{1}{n}\sum_{i=1}^{n} \sqrt{w_i}|y_i - \mu| \right| \le |\widehat{\mu}_m - \mu| \cdot \frac{1}{n}\sum_{i=1}^{n}\sqrt{w_i} \xrightarrow{P} 0.$$

Hence, $Q_m/n \xrightarrow{P} \mathrm{E}[|\sqrt{w_1}(y_1 - \mu)|] = \sqrt{2/\pi}$ and $I_m^2 = 1 - \frac{2/\pi}{(Q_m/n)^2} \xrightarrow{P} 0$. □

Proof of Proposition 2. Now, the weights w_i have a common value $w = 1/\sigma^2$. Under the random-effects setting, the weighted average and weighted median still converge to the true overall effect size μ in probability. Similarly to the derivations in Proposition 1, $Q/n \xrightarrow{P} \mathrm{E}[w(y_1 - \mu)^2] = (\sigma^2 + \tau^2)/\sigma^2$; both Q_r/n and Q_m/n converge to $\mathrm{E}[|\sqrt{w}(y_1 - \mu)|] = \sqrt{\frac{2}{\pi}}\sqrt{(\sigma^2 + \tau^2)/\sigma^2}$. Hence, $I^2 = 1 - \frac{1}{Q/(n-1)} \xrightarrow{P} I_0^2$, $I_r^2 = 1 - \frac{n-1}{n} \cdot \frac{2/\pi}{(Q_r/n)^2} \xrightarrow{P} I_0^2$, and $I_m^2 = 1 - \frac{2/\pi}{(Q_m^2/n)^2} \overset{P}{\xrightarrow{P}} I_0^2$, where $I_0^2 = \tau^2/(\sigma^2 + \tau^2)$. □

Proof of Proposition 3. Without loss of generality, let $y_i = z_i + C$ for $i = 1, \dots, \lfloor n\eta \rfloor$ and $y_i = z_i$ for $i = \lfloor n\eta \rfloor + 1, \dots, n$, where $z_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2 + \tau^2)$. Denote the weights $w_i = w = 1/\sigma^2$.

Note that $\bar{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\lfloor n\eta \rfloor}{n} \cdot \frac{\sum_{i=1}^{\lfloor n\eta \rfloor}(z_i + C)}{\lfloor n\eta \rfloor} + \frac{n - \lfloor n\eta \rfloor}{n} \cdot \frac{\sum_{i=n-\lfloor n\eta \rfloor+1}^{n} z_i}{n - \lfloor n\eta \rfloor} \xrightarrow{P} \eta(\mu + C) + (1 - \eta)\mu = \mu + \eta C$. Therefore,

$$\begin{aligned} Q/n &= \frac{w}{n}\sum_{i=1}^{n}[(y_i - \mu - \eta C) - (\bar{\mu} - \mu - \eta C)]^2 \\ &= \frac{w}{n}\sum_{i=1}^{n}(y_i - \mu - \eta C)^2 - 2w(\bar{\mu} - \mu - \eta C) \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu - \eta C) + w(\bar{\mu} - \mu - \eta C)^2. \end{aligned}$$

The last two terms on the right hand side converge to 0 in probability. For the first term, note that

$$\begin{aligned}
&\frac{w}{n}\sum_{i=1}^{n}(y_i-\mu-\eta C)^2\\
&= w\frac{\lfloor n\eta\rfloor}{n}\cdot\frac{\sum_{i=1}^{\lfloor n\eta\rfloor}(z_i-\mu+(1-\eta)C)^2}{\lfloor n\eta\rfloor}+w\frac{n-\lfloor n\eta\rfloor}{n}\cdot\frac{\sum_{i=n-\lfloor n\eta\rfloor+1}^{n}(z_i-\mu-\eta C)^2}{n-\lfloor n\eta\rfloor}.
\end{aligned}$$

It converges in probability to $w\eta\mathrm{E}[(z_1-\mu+(1-\eta)C)^2]+w(1-\eta)\mathrm{E}[(z_1-\mu-\eta C)^2]=\eta[\sigma^2+\tau^2+(1-\eta)^2C^2]/\sigma^2+(1-\eta)(\sigma^2+\tau^2+\eta^2C^2)/\sigma^2=(\sigma^2+\tau^2)/\sigma^2+\eta(1-\eta)C^2/\sigma^2=(1-I_0^2)^{-1}+r_1r_2$, where $I_0^2=\tau^2/(\sigma^2+\tau^2)$, $r_1=(1-\eta)C/\sigma$, and $r_2=\eta C/\sigma$. Therefore,

$$Q/n\xrightarrow{P}(1-I_0^2)^{-1}+r_1r_2,$$

and

$$I^2=1-\frac{1}{Q/(n-1)}\xrightarrow{P}1-[(1-I_0^2)^{-1}+r_1r_2]^{-1}.$$

To derive the asymptotic value of I_r^2, we apply the triangle inequality again as in the proof of Proposition 1, and obtain

$$\left|Q_r/n-\frac{\sqrt{w}}{n}\sum_{i=1}^{n}|y_i-\mu-\eta C|\right|\le\sqrt{w}|\bar{\mu}-\mu-\eta C|\xrightarrow{P}0.$$

Note that

$$\begin{aligned}
&\frac{\sqrt{w}}{n}\sum_{i=1}^{n}|y_i-\mu-\eta C|\\
&=\sqrt{w}\frac{\lfloor n\eta\rfloor}{n}\cdot\frac{\sum_{i=1}^{\lfloor n\eta\rfloor}|z_i-\mu+(1-\eta)C|}{\lfloor n\eta\rfloor}+\sqrt{w}\frac{n-\lfloor n\eta\rfloor}{n}\cdot\frac{\sum_{i=n-\lfloor n\eta\rfloor+1}^{n}|z_i-\mu-\eta C|}{n-\lfloor n\eta\rfloor}\\
&\xrightarrow{P}\sqrt{w}\eta\mathrm{E}[|z_1-\mu+(1-\eta)C|]+\sqrt{w}(1-\eta)\mathrm{E}[|z_1-\mu-\eta C|]\\
&=\frac{\eta}{\sigma}\left[\sqrt{\sigma^2+\tau^2}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{(1-\eta)^2C^2}{2(\sigma^2+\tau^2)}\right)+(1-\eta)C\left(1-2\Phi\left(-\frac{(1-\eta)C}{\sqrt{\sigma^2+\tau^2}}\right)\right)\right]\\
&\quad+\frac{1-\eta}{\sigma}\left[\sqrt{\sigma^2+\tau^2}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\eta^2C^2}{2(\sigma^2+\tau^2)}\right)-\eta C\left(1-2\Phi\left(\frac{\eta C}{\sqrt{\sigma^2+\tau^2}}\right)\right)\right]\\
&=\eta\left[\sqrt{\frac{2}{\pi}}(1-I_0^2)^{-1/2}\exp\left(-\frac{1}{2}r_1^2(1-I_0^2)\right)+r_1\left(1-2\Phi\left(-r_1(1-I_0^2)^{1/2}\right)\right)\right]\\
&\quad+(1-\eta)\left[\sqrt{\frac{2}{\pi}}(1-I_0^2)^{-1/2}\exp\left(-\frac{1}{2}r_2^2(1-I_0^2)\right)-r_2\left(1-2\Phi\left(r_2(1-I_0^2)^{1/2}\right)\right)\right].
\end{aligned}$$

Therefore, Q_r/n also converges to the value above in probability, and

$$\begin{aligned} I_r^2 &= 1 - \frac{n-1}{n} \cdot \frac{2/\pi}{(Q_r/n)^2} \\ &\xrightarrow{P} 1 - \left\{ \eta \left[(1-I_0^2)^{-1/2} \exp\left(-\frac{1}{2} r_1^2 (1-I_0^2)\right) + \sqrt{\frac{\pi}{2}} r_1 \left(1 - 2\Phi\left(-r_1(1-I_0^2)^{1/2}\right)\right)\right] \right. \\ &\quad \left. + (1-\eta) \left[(1-I_0^2)^{-1/2} \exp\left(-\frac{1}{2} r_2^2 (1-I_0^2)\right) - \sqrt{\frac{\pi}{2}} r_2 \left(1 - 2\Phi\left(r_2(1-I_0^2)^{1/2}\right)\right)\right] \right\}^{-2}. \end{aligned}$$

Finally, we derive the asymptotic value of I_m^2. The weighted median $\widehat{\mu}_m$ is defined as the solution to $\sum_{i=1}^n \psi(\theta) = 0$, where $\psi(\theta) = w[\mathbb{I}(\theta \geq y_i) - 0.5]$. Equivalently, $\widehat{\mu}_m$ is the solution to $\sum_{i=1}^n \widetilde{\psi}(\theta) = 0$, where $\widetilde{\psi}(\theta) = \mathbb{I}(\theta \geq y_i) - 0.5$ as we assume that the weights are equal. By the theory of M-estimation [73], $\widehat{\mu}_m \xrightarrow{P} \mu_0$, where μ_0 is the solution to $\mathrm{E}[\widetilde{\psi}(\theta)] = 0$. Specifically,

$$\begin{aligned} \mathrm{E}[\widetilde{\psi}(\theta)] &= \Pr(\theta \geq y_i) - 0.5 \\ &= \Pr(\theta \geq y_i, 1 \leq i \leq \lfloor n\eta \rfloor) + \Pr(\theta \geq y_i, \lfloor n\eta \rfloor + 1 \leq i \leq n) - 0.5 \\ &= \eta \Pr(z_i \leq \theta - C) + (1-\eta) \Pr(z_i \leq \theta) - 0.5 \\ &= \eta \Phi\left(\frac{\theta - \mu - C}{\sqrt{\sigma^2 + \tau^2}}\right) + (1-\eta) \Phi\left(\frac{\theta - \mu}{\sqrt{\sigma^2 + \tau^2}}\right) - 0.5. \end{aligned}$$

Therefore, μ_0 satisfied the following equation:

$$\eta \Phi\left(-\frac{\mu + C - \mu_0}{\sqrt{\sigma^2 + \tau^2}}\right) + (1-\eta) \Phi\left(\frac{\mu_0 - \mu}{\sqrt{\sigma^2 + \tau^2}}\right) = 0.5. \tag{B.1}$$

Applying the triangle inequality as in the proof of Proposition 1, we have

$$\left| Q_m/n - \frac{\sqrt{w}}{n} \sum_{i=1}^n |y_i - \mu_0| \right| \leq \sqrt{w} |\widehat{\mu}_m - \mu_0| \xrightarrow{P} 0.$$

Note that

$$
\begin{aligned}
&\frac{\sqrt{w}}{n}\sum_{i=1}^{n}|y_i-\mu_0| \\
&= \sqrt{w}\frac{\lfloor n\eta\rfloor}{n}\cdot\frac{\sum_{i=1}^{\lfloor n\eta\rfloor}|z_i-\mu_0+C|}{\lfloor n\eta\rfloor}+\sqrt{w}\frac{n-\lfloor n\eta\rfloor}{n}\cdot\frac{\sum_{i=n-\lfloor n\eta\rfloor+1}^{n}|z_i-\mu_0|}{n-\lfloor n\eta\rfloor} \\
&\xrightarrow{P}\sqrt{w}\eta\mathrm{E}[|z_1-\mu_0+C|]+\sqrt{w}(1-\eta)\mathrm{E}[|z_1-\mu_0|] \\
&= \frac{\eta}{\sigma}\left[\sqrt{\sigma^2+\tau^2}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{(\mu-\mu_0+C)^2}{2(\sigma^2+\tau^2)}\right)+(\mu-\mu_0+C)\left(1-2\Phi\left(-\frac{\mu-\mu_0+C}{\sqrt{\sigma^2+\tau^2}}\right)\right)\right] \\
&\quad+\frac{1-\eta}{\sigma}\left[\sqrt{\sigma^2+\tau^2}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{(\mu-\mu_0)^2}{2(\sigma^2+\tau^2)}\right)+(\mu-\mu_0)\left(1-2\Phi\left(-\frac{\mu-\mu_0}{\sqrt{\sigma^2+\tau^2}}\right)\right)\right] \\
&= \eta\left[\sqrt{\frac{2}{\pi}}(1-I_0^2)^{-1/2}\exp\left(-\frac{1}{2}s_1^2(1-I_0^2)\right)+s_1\left(1-2\Phi\left(-s_1(1-I_0^2)^{1/2}\right)\right)\right] \\
&\quad+(1-\eta)\left[\sqrt{\frac{2}{\pi}}(1-I_0^2)^{-1/2}\exp\left(-\frac{1}{2}s_2^2(1-I_0^2)\right)-s_2\left(1-2\Phi\left(s_2(1-I_0^2)^{1/2}\right)\right)\right],
\end{aligned}
$$

where $s_1=(\mu+C-\mu_0)/\sigma$ and $s_2=(\mu_0-\mu)/\sigma$. Therefore,

$$
\begin{aligned}
I_m^2 &= 1-\frac{2/\pi}{(Q_m/n)^2} \\
&\xrightarrow{P}1-\left\{\eta\left[(1-I_0^2)^{-1/2}\exp\left(-\frac{1}{2}s_1^2(1-I_0^2)\right)+\sqrt{\frac{\pi}{2}}s_1\left(1-2\Phi\left(-s_1(1-I_0^2)^{1/2}\right)\right)\right]\right. \\
&\quad\left.+(1-\eta)\left[(1-I_0^2)^{-1/2}\exp\left(-\frac{1}{2}s_2^2(1-I_0^2)\right)-\sqrt{\frac{\pi}{2}}s_2\left(1-2\Phi\left(s_2(1-I_0^2)^{1/2}\right)\right)\right]\right\}^{-2}.
\end{aligned}
$$

Notice that $s_2=C/\sigma-s_1$ and Equation (B.1) can be rewritten as

$$
\eta\Phi\left(-s_1(1-I_0^2)^{1/2}\right)+(1-\eta)\Phi\left((C/\sigma-s_1)(1-I_0^2)^{1/2}\right)=0.5; \qquad \text{(B.2)}
$$

that is, s_1 is the solution to Equation (B.2). This completes the proof. □

B.2 Asymptotic distribution of the publication bias measure T_S

Let $X_1,\ldots,X_n$ be iid random variables and denote the kth central moment $\beta_k=\mathrm{E}(X_1-\beta)^k$, where $\beta=\mathrm{E}(X_1)$. Also, denote the sample kth central moment $m_k=n^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^k$, where $\bar{X}=n^{-1}\sum_{i=1}^{n}X_i$. We have the following lemma regarding the asymptotic distribution of the sample kth central moment.

Lemma 1. *If $X_1, \dots, X_n$ are iid with mean β and $\beta_{2k} < \infty$ for $k \geq 1$, then*

$$m_k - \beta_k = \frac{1}{n} \sum_{i=1}^{n} \Big[(X_i - \beta)^k - \beta_k - k\beta_{k-1}(X_i - \beta) \Big] + o_p(n^{-1/2}).$$

as $n \to \infty$.

Proof of Lemma 1. See page 72 in [181]. □

Now, let us back to the notation in the main text. Specifically, let $x_i = (s_i^2 + \tau^2)^{-1/2}$ and $z_i = y_i(s_i^2 + \tau^2)^{-1/2}$. The regression test is $z_i = \alpha + \mu x_i + \epsilon_i$, where ϵ_i's are iid following a distribution with mean zero; $\widehat{\alpha}$ and $\widehat{\mu}$ are the least squares estimates of α and μ respectively, and the residuals $\widehat{\epsilon}_i = y_i - \widehat{\mu} x_i - \widehat{\alpha}$. Also, $\beta_k = \mathrm{E}(\epsilon_1 - \beta)^k$ is the kth central moment of ϵ_i's, where $\beta = \mathrm{E}(\epsilon_1) = 0$, and $m_k = n^{-1} \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^k$. The true skewness of ϵ_i's is $\gamma = \beta_3 / \beta_2^{3/2}$. Let $\widehat{m}_k = n^{-1} \sum_{i=1}^{n} (\widehat{\epsilon}_i - \bar{\widehat{\epsilon}})^k$ be the sample kth central moment by plugging in the residuals $\widehat{\epsilon}_i$, where $\bar{\widehat{\epsilon}} = n^{-1} \sum_{i=1}^{n} \widehat{\epsilon}_i = 0$. The sample skewness of $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^{\mathrm{T}}$ is $\mathrm{Skew}(\boldsymbol{\epsilon}) = m_3 / s^3$, where $s = \sqrt{n m_2 / (n-1)}$, and $T_S = \mathrm{Skew}(\widehat{\boldsymbol{\epsilon}})$ is obtained by plugging $\widehat{\boldsymbol{\epsilon}} = (\widehat{\epsilon}_1, \dots, \widehat{\epsilon}_n)^{\mathrm{T}}$ in $\mathrm{Skew}(\boldsymbol{\epsilon})$.

Proof of Proposition 4. First, we show that $\sqrt{n}(\mathrm{Skew}(\boldsymbol{\epsilon}) - \gamma) \xrightarrow{D} N(0, v)$ as $n \to \infty$, where

$$v = 9 + \frac{35}{4} \beta_2^{-3} \beta_3^2 - 6\beta_2^{-2} \beta_4 + \beta_2^{-3} \beta_6 + \frac{9}{4} \beta_2^{-5} \beta_3^2 \beta_4 - 3\beta_2^{-4} \beta_3 \beta_5.$$

Because $\mathrm{Skew}(\boldsymbol{\epsilon}) = [(n-1)/n]^{3/2} m_3 / m_2^{3/2}$, $\mathrm{Skew}(\boldsymbol{\epsilon})$ have the same asymptotic distribution as $m_3 / m_2^{3/2}$. By Lemma 1, we have

$$\begin{bmatrix} m_2 \\ m_3 \end{bmatrix} - \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \epsilon_i^2 - \beta_2 \\ \epsilon_i^3 - \beta_3 - 3\beta_2 \epsilon_i \end{bmatrix} + o_p(n^{-1/2}).$$

Therefore,

$$\sqrt{n} \left(\begin{bmatrix} m_2 \\ m_3 \end{bmatrix} - \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} \right) \xrightarrow{D} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \beta_4 - \beta_2^2 & \beta_5 - 4\beta_2\beta_3 \\ \beta_5 - 4\beta_2\beta_3 & \beta_6 - \beta_3^2 - 6\beta_2\beta_4 + 9\beta_2^3 \end{bmatrix} \right).$$

Denote the asymptotic covariance matrix above as $\boldsymbol{\Sigma}$. Let $g(r, s) = s / r^{3/2}$, then $g'(r, s) = \left(-\frac{3}{2} s r^{-5/2}, r^{-3/2}\right)^{\mathrm{T}}$. By the delta method,

$$\sqrt{n}(g(m_2, m_3) - g(\beta_2, \beta_3)) \xrightarrow{D} N(0, [g'(\beta_2, \beta_3)]^{\mathrm{T}} \boldsymbol{\Sigma} [g'(\beta_2, \beta_3)]);$$

that is,

$$\sqrt{n}(\mathrm{Skew}(\boldsymbol{\epsilon}) - \gamma) \xrightarrow{D} N\left(0, v\right).$$

Second, we show that $\sqrt{n}(T_S - \mathrm{Skew}(\boldsymbol{\epsilon})) \xrightarrow{D} 0$ as $n \to \infty$. We write $\mathrm{Skew}(\boldsymbol{\epsilon}) = [(n-1)/n]^{3/2} f(\boldsymbol{\delta})$, where $f(\boldsymbol{\delta}) = m_3/m_2^{3/2}$ is a continuous and differentiable function of $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)^{\mathrm{T}} = (\bar{\epsilon}^2, \bar{\epsilon}^3, \bar{\epsilon} \cdot \overline{\epsilon^2}, \overline{\epsilon^2}, \overline{\epsilon^3})^{\mathrm{T}}$; here, $\overline{\epsilon^k} = n^{-1}\sum_{i=1}^{n} \epsilon_i^k$. Specifically, $f(\boldsymbol{\delta}) = (\delta_5 - 3\delta_3 + 2\delta_2)(\delta_4 - \delta_1)^{-3/2}$; it is free of n. Also, $T_S = \mathrm{Skew}(\widehat{\boldsymbol{\epsilon}}) = [(n-1)/n]^{3/2} f(\widehat{\boldsymbol{\delta}})$, where $\widehat{\boldsymbol{\delta}} = \left(\left(\bar{\widehat{\epsilon}}\right)^2, \left(\bar{\widehat{\epsilon}}\right)^3, \bar{\widehat{\epsilon}} \cdot \overline{\widehat{\epsilon}^2}, \overline{\widehat{\epsilon}^2}, \overline{\widehat{\epsilon}^3}\right)^{\mathrm{T}}$, and $\overline{\widehat{\epsilon}^k} = n^{-1}\sum_{i=1}^{n} \widehat{\epsilon}_i^k$. Because the average of the residuals is $\bar{\widehat{\epsilon}} = 0$, we have $\widehat{\boldsymbol{\delta}} = \left(0, 0, 0, \overline{\widehat{\epsilon}^2}, \overline{\widehat{\epsilon}^3}\right)^{\mathrm{T}}$. By multivariate Taylor expansion,

$$f(\widehat{\boldsymbol{\delta}}) = f(\boldsymbol{\delta}) + [\boldsymbol{h}(\boldsymbol{\delta})]^{\mathrm{T}}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O_p(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2),$$

where $\boldsymbol{h}(\boldsymbol{\delta}) = \bigtriangledown f(\boldsymbol{\delta})$ is the gradient of $f(\boldsymbol{\delta})$ and $\|\cdot\|$ is the Euclidean norm. Specifically,

$$\boldsymbol{h}(\boldsymbol{\delta}) = \begin{bmatrix} h_1(\boldsymbol{\delta}) \\ h_2(\boldsymbol{\delta}) \\ h_3(\boldsymbol{\delta}) \\ h_4(\boldsymbol{\delta}) \\ h_5(\boldsymbol{\delta}) \end{bmatrix} = \begin{bmatrix} \frac{3}{2}(\delta_5 - 3\delta_3 + 2\delta_2)(\delta_4 - \delta_1)^{-5/2} \\ 2(\delta_4 - \delta_1)^{3/2} \\ -3(\delta_4 - \delta_1)^{3/2} \\ -\frac{3}{2}(\delta_5 - 3\delta_3 + 2\delta_2)(\delta_4 - \delta_1)^{-5/2} \\ (\delta_4 - \delta_1)^{3/2} \end{bmatrix}.$$

Since $\delta_1, \delta_2, \delta_3 \xrightarrow{P} 0$, $\delta_4 \xrightarrow{P} \beta_2 > 0$, and $\delta_5 \xrightarrow{P} \beta_3$, we have $h_j(\boldsymbol{\delta}) = O_p(1)$ for $j = 1, \ldots, 5$. Now, we focus on

$$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta} = \left(-\bar{\epsilon}^2, -\bar{\epsilon}^3, -\bar{\epsilon} \cdot \overline{\epsilon^2}, \overline{\widehat{\epsilon}^2} - \overline{\epsilon^2}, \overline{\widehat{\epsilon}^3} - \overline{\epsilon^3}\right)^{\mathrm{T}}.$$

Due to $\bar{\epsilon} = O_p(n^{-1/2})$, we have $\widehat{\delta}_1 - \delta_1 = -\bar{\epsilon}^2 = O_p(n^{-1})$, $\widehat{\delta}_2 - \delta_2 = -\bar{\epsilon}^3 = O_p(n^{-3/2})$, and $\widehat{\delta}_3 - \delta_3 = -\bar{\epsilon} \cdot \overline{\epsilon^2} = O_p(n^{-1/2})$. Note that

$$\widehat{\epsilon}_i = (\alpha - \widehat{\alpha}) + (\mu - \widehat{\mu})x_i + \epsilon_i,$$

and $\widehat{\alpha} - \alpha = O_p(n^{-1/2})$, $\widehat{\mu} - \mu = O_p(n^{-1/2})$. Also, by the assumption that x_i's have finite third moment and the weak law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} x_i^k = O_p(1)$ for $k = 1, 2, 3$.

Consequently, we have

$$\begin{aligned}
\widehat{\delta}_4 - \delta_4 &= \overline{\widehat{\epsilon}^2} - \overline{\epsilon^2} \\
&= \frac{1}{n}\sum_{i=1}^{n}[(\alpha-\widehat{\alpha}) + (\mu-\widehat{\mu})x_i + \epsilon_i]^2 - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \\
&= (\alpha-\widehat{\alpha})^2 + (\mu-\widehat{\mu})^2\frac{\sum_{i=1}^{n}x_i^2}{n} + 2(\alpha-\widehat{\alpha})(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i}{n} \\
&\quad + 2(\alpha-\widehat{\alpha})\bar{\epsilon} + 2(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i\epsilon_i}{n} \\
&= O_p(n^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\delta}_5 - \delta_5 &= \overline{\widehat{\epsilon}^3} - \overline{\epsilon^3} \\
&= \frac{1}{n}\sum_{i=1}^{n}[(\alpha-\widehat{\alpha}) + (\mu-\widehat{\mu})x_i + \epsilon_i]^3 - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^3 \\
&= (\alpha-\widehat{\alpha})^3 + (\mu-\widehat{\mu})^3\frac{\sum_{i=1}^{n}x_i^3}{n} + 3(\alpha-\widehat{\alpha})^2(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i}{n} \\
&\quad + 3(\alpha-\widehat{\alpha})(\mu-\widehat{\mu})^2\frac{\sum_{i=1}^{n}x_i^2}{n} + 3(\alpha-\widehat{\alpha})^2\bar{\epsilon} + 3(\mu-\widehat{\mu})^2\frac{\sum_{i=1}^{n}x_i^2\epsilon_i}{n} \\
&\quad + 6(\alpha-\widehat{\alpha})(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i\epsilon_i}{n} + 3(\alpha-\widehat{\alpha})\overline{\epsilon^2} + 3(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i\epsilon_i^2}{n} \\
&= 3(\alpha-\widehat{\alpha})\overline{\epsilon^2} + 3(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i\epsilon_i^2}{n} + O_p(n^{-1}) \\
&= O_p(n^{-1/2}).
\end{aligned}$$

Therefore, $O_p(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2) = O_p(n^{-1})$, implying

$$\begin{aligned}
f(\widehat{\boldsymbol{\delta}}) - f(\boldsymbol{\delta}) &= [\boldsymbol{h}(\boldsymbol{\delta})]^{\mathrm{T}}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O_p(n^{-1}) \\
&= \sum_{j=1}^{5}h_j(\boldsymbol{\delta})(\widehat{\delta}_j - \delta_j) + O_p(n^{-1}) \\
&= h_3(\boldsymbol{\delta})(\widehat{\delta}_3 - \delta_3) + h_5(\boldsymbol{\delta})(\widehat{\delta}_5 - \delta_5) + O_p(n^{-1}) \\
&= 3(\delta_4-\delta_1)^{3/2}\cdot\bar{\epsilon}\cdot\overline{\epsilon^2} + (\delta_4-\delta_1)^{3/2}\left[3(\alpha-\widehat{\alpha})\overline{\epsilon^2} + 3(\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i\epsilon_i^2}{n}\right] + O_p(n^{-1}) \\
&= 3(\delta_4-\delta_1)^{3/2}\left\{[(\alpha-\widehat{\alpha}) + \bar{\epsilon}]\,\overline{\epsilon^2} + (\mu-\widehat{\mu})\frac{\sum_{i=1}^{n}x_i\epsilon_i^2}{n}\right\} + O_p(n^{-1})
\end{aligned}$$

Note that $\sum_{i=1}^{n} \widehat{\epsilon}_i = 0$, so $(\alpha - \widehat{\alpha}) + \bar{\epsilon} = (\widehat{\mu} - \mu)\frac{\sum_{i=1}^{n} x_i}{n}$. Consequently,

$$\begin{aligned}
f(\widehat{\boldsymbol{\delta}}) - f(\boldsymbol{\delta}) &= 3(\delta_4 - \delta_1)^{3/2} \left\{ (\widehat{\mu} - \mu)\frac{\sum_{i=1}^{n} x_i}{n}\overline{\epsilon^2} - (\widehat{\mu} - \mu)\frac{\sum_{i=1}^{n} x_i \epsilon_i^2}{n} \right\} + O_p(n^{-1}) \\
&= 3(\delta_4 - \delta_1)^{3/2}(\widehat{\mu} - \mu) \left\{ \frac{\sum_{i=1}^{n} x_i}{n}\overline{\epsilon^2} - \frac{\sum_{i=1}^{n} x_i \epsilon_i^2}{n} \right\} + O_p(n^{-1}) \\
&= O_p(n^{-1/2})\big\{[\mathrm{E}(x_1) + O_p(n^{-1/2})][\beta_2 + O_p(n^{-1/2})] \\
&\qquad\qquad - [\mathrm{E}(x_1\epsilon_1^2) + O_p(n^{-1/2})]\big\} + O_p(n^{-1}) \\
&= O_p(n^{-1/2}) \Big\{ [\mathrm{E}(x_1)\beta_2 + O_p(n^{-1/2})] - [\mathrm{E}(x_1)\beta_2 + O_p(n^{-1/2})] \Big\} + O_p(n^{-1}) \\
&= O_p(n^{-1}).
\end{aligned}$$

This leads to $\sqrt{n}(f(\widehat{\boldsymbol{\delta}}) - f(\boldsymbol{\delta})) \xrightarrow{D} 0$; hence, $\sqrt{n}(T_S - \mathrm{Skew}(\boldsymbol{\epsilon})) \xrightarrow{D} 0$, and $\sqrt{n}(T_S - \gamma) \xrightarrow{D} N(0, v)$.

Finally, we show that $\widehat{v} \xrightarrow{P} v$. By continuous mapping theorem, it is sufficient to show that $\widehat{m}_k \xrightarrow{P} \beta_k$ for $k = 2, \ldots, 6$. Recall that $\beta_k = \mathrm{E}(\epsilon_1^k)$ and $\widehat{m}_k = n^{-1}\sum_{i=1}^{n} \widehat{\epsilon}_i^k$. Since $\widehat{\epsilon}_i = (\alpha - \widehat{\alpha}) + (\mu - \widehat{\mu})x_i + \epsilon_i = \epsilon_i + O_p(n^{-1/2})$, we have $\widehat{m}_k = n^{-1}\sum_{i=1}^{n}(\epsilon_i + O_p(n^{-1/2}))^k = n^{-1}\sum_{i=1}^{n} \epsilon_i^k + o_p(1) = \beta_k + o_p(1)$; that is, $\widehat{m}_k \xrightarrow{P} \beta_k$. By Slutsky's theorem, $\sqrt{n}(T_S - \gamma)/\sqrt{\widehat{v}} \xrightarrow{D} N(0, 1)$; this completes the proof. □

Proof of Corollary 1. Under H_0'', we have $\epsilon_i \sim N(0, \sigma^2)$, so $\beta_{2k} = (2k-1)!!\sigma^{2k}$ and $\beta_{2k-1} = 0$ for $k \geq 1$. Here, $c!! = c \cdot (c-2) \cdot (c-4) \cdots$ is the double factorial. Specifically, $\beta_2 = \sigma^2$, $\beta_4 = 3\sigma^4$, and $\beta_6 = 15\sigma^6$. In the proof of Proposition 1, we showed that $\sqrt{n}(T_S - \gamma) \xrightarrow{D} N(0, v)$. Under H_0'', v is simplified as $v = 9 - 6(\sigma^2)^{-2} \cdot 3\sigma^4 + (\sigma^2)^{-3} \cdot 15\sigma^6 = 6$. This completes the proof. □

B.3 Properties of pairwise and network meta-analyses

Proof of Proposition 6. For simplicity, let $P = (K-2)(K-1)/2$, the dimension of $\boldsymbol{e}_\mathrm{f}$, so $\mathbf{A}$ is $P \times (K-1)$. First, if there are two distinct transformation matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ such that $\boldsymbol{e}_\mathrm{f} = \mathbf{A}_1\boldsymbol{e}_\mathrm{b}$ and $\boldsymbol{e}_\mathrm{f} = \mathbf{A}_2\boldsymbol{e}_\mathrm{b}$, then $(\mathbf{A}_1 - \mathbf{A}_2)\boldsymbol{e}_\mathrm{b} = \mathbf{0}$. Let $\mathbf{A}_1 = (\boldsymbol{a}_{11}, \ldots, \boldsymbol{a}_{1P})^\mathrm{T}$ and $\mathbf{A}_2 = (\boldsymbol{a}_{21}, \ldots, \boldsymbol{a}_{2P})^\mathrm{T}$. Since $\mathbf{A}_1 \neq \mathbf{A}_2$, there is at least one $k = 1, \ldots, P$ such that $\boldsymbol{a}_{1k} \neq \boldsymbol{a}_{2k}$. Consequently, $(\boldsymbol{a}_{1k} - \boldsymbol{a}_{2k})^\mathrm{T}\boldsymbol{e}_\mathrm{b} = 0$; this implies evidence cycles in $\boldsymbol{e}_\mathrm{b}$ and contradicts the definition of basic parameters.

Second, to investigate the entries of $\mathbf{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_P)^{\mathrm{T}}$, consider the pth indirect treatment contrast in $\boldsymbol{e}_{\mathrm{f}}$ ($p = 1, \ldots, P$), denoted as d_{hk} ($h < k$); it corresponds to the vector $\boldsymbol{a}_p = (a_{p1}, \ldots, a_{p,K-1})^{\mathrm{T}}$ in $\mathbf{A}$. Due to the network's connectivity, treatments h and k must be linked through a certain path; the argument in the above paragraph implies that the path is unique. Suppose that this unique path contains $M + 1$ vertices ($M \geq 2$), say $\ell_0 = h, \ell_1, \ldots, \ell_{M-1}, \ell_M = k$. Consequently, iteratively using the evidence consistency equation, the indirect treatment contrast d_{hk} can be obtained from M direct treatment contrasts; that is, $d_{hk} = d_{\ell_0\ell_1} + d_{\ell_1\ell_2} + \cdots + d_{\ell_{M-1}\ell_M}$. Recall that $\boldsymbol{e}_{\mathrm{b}} = (e_1, \ldots, e_{K-1})^{\mathrm{T}}$ contains all direct treatment contrasts. For each $i = 1, \ldots, M$, there is some e_{j_i} ($j_i = 1, \ldots, K-1$) such that $d_{\ell_{i-1}\ell_i} = x_i e_{j_i}$, where $x_i = 1$ if $\ell_{i-1} < \ell_i$ and $x_i = -1$ if $\ell_{i-1} > \ell_i$. Consequently, we can write $d_{hk} = \sum_{i=1}^{M} x_i e_{j_i}$. On the other hand, $d_{hk} = \boldsymbol{a}_p^{\mathrm{T}} \boldsymbol{e}_{\mathrm{b}} = \sum_{j=1}^{K-1} a_{pj} e_j$. Therefore, if $j \notin \{j_1, \ldots, j_M\}$, $a_{pj} = 0$; if $j = j_i$ for some i, $a_{pj} = x_i$. This completes the proof. □

Proof of Theorem 2. Both the Lu–Ades and Smith models use the evidence consistency equation (i.e., $\boldsymbol{e}_{\mathrm{f}} = \mathbf{A}\boldsymbol{e}_{\mathrm{b}}$) to impute the indirect treatment contrasts, so the posterior distributions of $\boldsymbol{e}_{\mathrm{f}}$ produced by the two models are entirely determined by $\boldsymbol{e}_{\mathrm{b}}$. Since Theorem 1 showed that the two models produce identical posterior distributions of $\boldsymbol{e}_{\mathrm{b}}$, the posterior distributions of $\boldsymbol{e}_{\mathrm{f}}$ must also be identical. Furthermore, we make regularity assumptions that $\varphi_{\mathrm{f}}(\boldsymbol{t})$ and $p(\boldsymbol{e}_{\mathrm{f}} \mid \mathcal{D})$ are in L^P space. Given $\boldsymbol{e}_{\mathrm{f}}$'s characteristic function $\varphi_{\mathrm{f}}(\boldsymbol{t})$, its posterior distribution is

$$p(\boldsymbol{e}_{\mathrm{f}} \mid \mathcal{D}) = \frac{1}{(2\pi)^P} \int_{\mathbb{R}^P} e^{-i\boldsymbol{t}^{\mathrm{T}}\boldsymbol{e}_{\mathrm{f}}} \varphi_{\mathrm{f}}(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t};$$

see Equation (10.6.3) in [182]. Note that

$$\varphi_{\mathrm{f}}(\boldsymbol{t}) = E\left(e^{i\boldsymbol{t}^{\mathrm{T}}\boldsymbol{e}_{\mathrm{f}}} \mid \mathcal{D}\right) = E\left(e^{i\boldsymbol{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{e}_{\mathrm{b}}} \mid \mathcal{D}\right) = E\left(e^{i(\mathbf{A}^{\mathrm{T}}\boldsymbol{t})^{\mathrm{T}}\boldsymbol{e}_{\mathrm{b}}} \mid \mathcal{D}\right) = \varphi_{\mathrm{b}}(\mathbf{A}^{\mathrm{T}}\boldsymbol{t}).$$

Since $p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) = \prod_{j=1}^{K-1} p(e_j \mid \mathcal{D}_j)$, for $\boldsymbol{s} = (s_1, \ldots, s_{K-1})^{\mathrm{T}}$, we have

$$\begin{aligned}\varphi_{\mathrm{b}}(\boldsymbol{s}) = E\left(e^{i\boldsymbol{s}^{\mathrm{T}}\boldsymbol{e}_{\mathrm{b}}} \mid \mathcal{D}\right) &= E\left(e^{i\sum_{j=1}^{K-1} s_j e_j} \mid \mathcal{D}\right) \\ &= \prod_{j=1}^{K-1} E\left(e^{i s_j e_j} \mid \mathcal{D}\right) = \prod_{j=1}^{K-1} \int_{\mathbb{R}} e^{i s_j e_j} p(e_j \mid \mathcal{D}_j) \, \mathrm{d}e_j.\end{aligned}$$

This completes the proof. □

Proof of Theorem 3. We retain the notation $\widetilde{\boldsymbol{\mu}}_j$, $\boldsymbol{y}_j$, $\boldsymbol{\xi}_j$ ($j = 1, \ldots, K-1$), $\boldsymbol{\mu}$, $\boldsymbol{y}$, and $\boldsymbol{\xi}$ defined in the proof of Theorem 1 for treatment networks without evidence cycles; however, now σ is a scalar, not a vector, representing the common heterogeneity standard deviation of all treatment contrasts. Consequently, under the assumption of equal heterogeneity standard deviations, the Lu–Ades random-effects model gives the posterior distribution of $\boldsymbol{e}_{\mathrm{b}}$ as:

$$\begin{aligned}
p_{\mathrm{LA}}(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) &= \iint p(\boldsymbol{e}_{\mathrm{b}}, \sigma, \boldsymbol{\mu} \mid \mathcal{D}) \, \mathrm{d}\sigma \, \mathrm{d}\boldsymbol{\mu} \\
&\propto \iint f(\boldsymbol{y} \mid \boldsymbol{e}_{\mathrm{b}}, \sigma, \boldsymbol{\mu}, \boldsymbol{\xi}) p(\boldsymbol{e}_{\mathrm{b}}) p(\sigma) p(\boldsymbol{\mu}) \, \mathrm{d}\sigma \, \mathrm{d}\boldsymbol{\mu} \\
&= \iint \left\{ \prod_{j=1}^{K-1} f(\boldsymbol{y}_j \mid e_j, \sigma, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\widetilde{\boldsymbol{\mu}}_j) \right\} p(\sigma) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\sigma \\
&= \int \left\{ \prod_{j=1}^{K-1} \int f(\boldsymbol{y}_j \mid e_j, \sigma, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\widetilde{\boldsymbol{\mu}}_j) \, \mathrm{d}\widetilde{\boldsymbol{\mu}}_j \right\} p(\sigma) \, \mathrm{d}\sigma.
\end{aligned}$$

Like the proof of Theorem 1, the first two steps above are consequences of the properties of conditional probability and the likelihood of the outcome measure $\boldsymbol{y}$; these are also valid for the Smith random-effects models. The third step is due to the partition of studies in the network without cycles, i.e., $\mathcal{S} = \bigcup_{j=1}^{K-1} \mathcal{S}_j$, and the outcome measures $\boldsymbol{y}_j$ in studies $\mathcal{S}_j$ depending on e_j but not the other basic parameters in $\boldsymbol{e}_{\mathrm{b}}$. This study partition also naturally holds when the Smith random-effects models are used for different sets of studies $\mathcal{S}_j$. Therefore, by simultaneously using the Smith random-effects models conditional on the common heterogeneity standard deviation σ, the posterior distribution of $\boldsymbol{e}_{\mathrm{b}}$ is also

$$\begin{aligned}
p_{\mathrm{S}}(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) &\propto \iint \left\{ \prod_{j=1}^{K-1} f(\boldsymbol{y}_j \mid e_j, \sigma, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\widetilde{\boldsymbol{\mu}}_j) \right\} p(\sigma) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\sigma \\
&= \int \left\{ \prod_{j=1}^{K-1} \int f(\boldsymbol{y}_j \mid e_j, \sigma, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\widetilde{\boldsymbol{\mu}}_j) \, \mathrm{d}\widetilde{\boldsymbol{\mu}}_j \right\} p(\sigma) \, \mathrm{d}\sigma.
\end{aligned}$$

That is, $p_{\mathrm{LA}}(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) = p_{\mathrm{S}}(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D})$. This completes the proof. □

Proof of Proposition 7. Among all connected networks with K treatments, we consider the network which has the largest number of acyclic comparisons. This network may not be unique, but it must contain no evidence cycles. Otherwise, the removal of certain comparisons in evidence cycles would add acyclic comparisons; this contradicts the fact that this network has the most acyclic comparisons among all networks with K treatments. As this network does not have cycles, it is a spanning tree and contains $K-1$ treatment comparisons. Hence, the number of acyclic comparisons does not exceed $K-1$; i.e., $J \leq K-1$. □

Proof of Theorem 4. If there is evidence inconsistency in the sub-network consisting of $\mathcal{S}^\star$, let $\boldsymbol{w}^\star$ be the inconsistency factors. Consequently, the functional parameters $\boldsymbol{e}_\mathrm{f}^\star$ in the sub-network are determined by $\boldsymbol{e}_\mathrm{b}^\star$ and $\boldsymbol{w}^\star$. The remaining proof is similar to that of Theorem 1. In the Smith and Lu–Ades fixed-effects models, let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^\mathrm{T}$ be the study-specific baseline effects, $\widetilde{\boldsymbol{\mu}}_j = (\mu_i; i \in \mathcal{S}_j)^\mathrm{T}$ be the baseline effects in studies $\mathcal{S}_j$, and $\boldsymbol{\mu}^\star = (\mu_i; i \in \mathcal{S}^\star)^\mathrm{T}$ be those in studies $\mathcal{S}^\star$. Denote $\boldsymbol{y}_j = (y_{ik}; i \in \mathcal{S}_j, k \in \mathcal{T}_i)^\mathrm{T}$, $\boldsymbol{\xi}_j = (\xi_{ik}; i \in \mathcal{S}_j, k \in \mathcal{T}_i)^\mathrm{T}$, and $\boldsymbol{y}^\star = (y_{ik}; i \in \mathcal{S}^\star, k \in \mathcal{T}_i)^\mathrm{T}$, $\boldsymbol{\xi}^\star = (\xi_{ik}; i \in \mathcal{S}^\star, k \in \mathcal{T}_i)^\mathrm{T}$. We have

$$\begin{aligned}
p(\boldsymbol{e}_\mathrm{b} \mid \mathcal{D}) &= \iint p(\boldsymbol{e}_\mathrm{b}, \boldsymbol{\mu}, \boldsymbol{w}^\star \mid \mathcal{D})\,\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}\boldsymbol{w}^\star \\
&\propto \iint f(\boldsymbol{y} \mid \boldsymbol{e}_\mathrm{b}, \boldsymbol{\mu}, \boldsymbol{w}^\star, \boldsymbol{\xi}) p(\boldsymbol{e}_\mathrm{b}) p(\boldsymbol{\mu}) p(\boldsymbol{w}^\star)\,\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}\boldsymbol{w}^\star \\
&= \iint f(\boldsymbol{y}^\star \mid \boldsymbol{e}_\mathrm{b}^\star, \boldsymbol{\mu}^\star, \boldsymbol{w}^\star, \boldsymbol{\xi}^\star) p(\boldsymbol{e}_\mathrm{b}^\star) p(\boldsymbol{\mu}^\star) p(\boldsymbol{w}^\star)\,\mathrm{d}\boldsymbol{\mu}^\star\,\mathrm{d}\boldsymbol{w}^\star \\
&\qquad \times \prod_{j=1}^{J} \int f(\boldsymbol{y}_j \mid e_j, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\widetilde{\boldsymbol{\mu}}_j)\,\mathrm{d}\widetilde{\boldsymbol{\mu}}_j \\
&\propto p(\boldsymbol{e}_\mathrm{b}^\star \mid \mathcal{D}^\star) \prod_{j=1}^{J} p(e_j \mid \mathcal{D}_j).
\end{aligned}$$

The notation in the equation above is similar to the notation in the proof of Theorem 1, and $p(\boldsymbol{w}^\star)$ is the prior of $\boldsymbol{w}^\star$.

In the random-effects models, further denote σ_j as the heterogeneity standard deviation of the acyclic treatment comparison e_j ($j = 1, \ldots, J$), and let $\boldsymbol{\sigma}^\star$ be the vector of heterogeneity standard deviations in the sub-network consisting of $\mathcal{S}^\star$ and

$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_J, (\boldsymbol{\sigma}^\star)^{\mathrm{T}})^{\mathrm{T}}$ be the vector of heterogeneity standard deviations in the entire network. As in the foregoing, we have

$$\begin{aligned}
p(\boldsymbol{e}_{\mathrm{b}} \mid \mathcal{D}) &= \iiint p(\boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{w}^\star \mid \mathcal{D}) \,\mathrm{d}\boldsymbol{\sigma} \,\mathrm{d}\boldsymbol{\mu} \,\mathrm{d}\boldsymbol{w}^\star \\
&\propto \iiint f(\boldsymbol{y} \mid \boldsymbol{e}_{\mathrm{b}}, \boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{w}^\star, \boldsymbol{\xi}) p(\boldsymbol{e}_{\mathrm{b}}) p(\boldsymbol{\sigma}) p(\boldsymbol{\mu}) p(\boldsymbol{w}^\star) \,\mathrm{d}\boldsymbol{\sigma} \,\mathrm{d}\boldsymbol{\mu} \,\mathrm{d}\boldsymbol{w}^\star \\
&= \iiint f(\boldsymbol{y}^\star \mid \boldsymbol{e}_{\mathrm{b}}^\star, \boldsymbol{\sigma}^\star, \boldsymbol{\mu}^\star, \boldsymbol{w}^\star, \boldsymbol{\xi}^\star) p(\boldsymbol{e}_{\mathrm{b}}^\star) p(\boldsymbol{\sigma}^\star) p(\boldsymbol{\mu}^\star) p(\boldsymbol{w}^\star) \,\mathrm{d}\boldsymbol{\sigma}^\star \,\mathrm{d}\boldsymbol{\mu}^\star \,\mathrm{d}\boldsymbol{w}^\star \\
&\qquad \times \prod_{j=1}^{J} \iint f(\boldsymbol{y}_j \mid e_j, \sigma_j, \widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\xi}_j) p(e_j) p(\sigma_j) p(\widetilde{\boldsymbol{\mu}}_j) \,\mathrm{d}\sigma_j \,\mathrm{d}\widetilde{\boldsymbol{\mu}}_j \\
&\propto p(\boldsymbol{e}_{\mathrm{b}}^\star \mid \mathcal{D}^\star) \prod_{j=1}^{J} p(e_j \mid \mathcal{D}_j).
\end{aligned}$$

If all evidence cycles are consistent, the inconsistency factors $\boldsymbol{w}^\star$ can be simply ignored in the equations above. This completes the proof. □