Bayesian Hierarchical Methods for Network Meta-Analysis

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

Jing Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Haitao Chu

July, 2014

Acknowledgements

I owe the greatest debt of gratitude to my advisor, Dr. Haitao Chu, for his patient guidance, support, and encouragement. Dr. Haitao Chu has been a great mentor, role model, and friend, all the way from my master study, through to completion of my PhD degree. His guidance has made this a wonderful and rewarding journey. I would like to send a warm and well-deserved thank you to my advisor Dr. Haitao Chu for always being there for me whenever I need advice. Dr. Bradley P. Carlin's intellectual heft is matched only by his genuinely good nature and down-to-earth humility, and I am truly fortunate to have had opportunity to work with him. He is indeed an inspiration for me. Many thanks also go to Dr. James D. Neaton and Dr. Beth A. Virnig, who are my committee members, for their friendly guidance, thought-provoking suggestions, and invaluable input. In a similar vein, I'd like to recognize Drs. Cavan Reilly, William Thomas, and Wei Pan for the contributions that each of them made to my intellectual growth during my years of study at the University of Minnesota. I am also grateful to my undergraduate professor, Dr. Jing Chen, for her generosity, guidance, and continuous emotional support.

I also want to thank my parents and my sister for supporting me, believing in me, and loving me throughout my 5 years here at UMN. Dad and Mom, thank you so much for teaching me respect, confidence, and proper etiquette. Thank you for letting me find my own way and be persistent to achieve my goal. Finally, special thanks to my wonderful husband for showing me the real happiness in life. I can face anything because he is by my side. I cannot find words to utter but I just want to say Thank you darling!

Dedication

This dissertation is dedicated to my family, especially

to my brilliant and outrageously loving and supportive husband, Yiping Yuan;

to my always encouraging and supportive parents, Aibao and Chunhua Zhang, and parents-in law, Zhongwen Yuan and Li Cao.

to my exuberant and sweet younger sister, Lin Zhang.

Abstract

In clinical practice, and at a wider societal level, treatment decisions in medicine need to consider all relevant evidence. Network meta-analysis (NMA) collectively analyzes many randomized controlled trials (RCTs) evaluating multiple interventions relevant to a treatment decision, expanding the scope of a conventional pairwise meta-analysis to simultaneously handle multiple treatment comparisons. NMA synthesizes both direct information, gained from direct comparison for example between treatments A and C, and indirect information obtained from A versus B and C versus B trials, and thus strengthens inference.

Under current contrast-based (CB) methods for NMA of binary outcomes, which do not model the baseline risks and focus on modeling the relative treatment effects, the patient-centered measures including the overall treatment-specific event rates and risk differences are not provided, creating some unnecessary obstacles for patients to comprehensively understand and trade-off efficacy and safety measures. Many NMAs only report odds ratios (ORs) which are commonly misinterpreted as risk ratios (RRs) by many physicians, patients and their care givers. In order to overcome these obstacles of the CB methods, a novel Bayesian hierarchical arm-based (AB) model developed from a missing data perspective is proposed to illustrate how treatment-specific event proportions, risk differences (RD) and relative risks (RR) can be computed in NMAs.

Since most of the trials in NMA only compare two of the treatments of interest, the typical data in a NMA managed as a trial-by-treatment matrix is extremely sparse, like an incomplete block structure with serious missing data problems. The previously proposed AB method assumes a missing at random (MAR) mechanism. However, in RCTs, nonignorable missingness or missingness not at random (MNAR) may occur due to deliberate choices at the design stage. In addition, those undertaking an NMA will often selectively choose treatments to include in the analysis, which will also lead to nonignorable missingness. We then extend the AB method to incorporate nonignorable missingness using *selection models* method.

Meta-analysts undertaking an NMA often selectively choose trials to include in the analysis. Thus inevitably, certain trials are more likely to be included in an NMA. In

addition, it is difficult to include all existing trials that meet the inclusion criteria due to language barriers (i.e., some trials may be published using other languages) and other technical issues. If the omitted trials are quite different from the ones we include, then the estimates will be biased. We obtain empirical evidence on whether these selective inclusions of trials can make a difference in the results, such as treatment effect estimates in an NMA setting, using both the AB and CB methods.

In the opposite direction of the fact that some trials which should have been included but are omitted, some trials may appear to deviate markedly from the others, and thus be inappropriate to be synthesized. we call these trials *outlying trials* or *trial-level outliers*. To the best of our knowledge, while the key NMA assumptions of inconsistency and heterogeneity have been well-studied, few previous authors have considered the issue of trial-level outliers, their detection, and guidance on whether or not to discard them from an NMA. We propose and evaluate Bayesian approaches to detect trial-level outliers in the NMA evidence structures.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Meta-analysis, a statistical technique to assess treatment effects quantitatively by combining the results from several independent studies [2, 3], is now a hallmark of *Comparative Effectiveness Research* (CER) and *Evidence Based Medicine* (EBM) [4][5], two rapidly growing fields whose objective is to assess how various medical interventions result in improved health care outcomes. CER is the study of two or more approaches to a health problem to determine which one results in better health outcomes [4]. EBM is defined as the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients [6][7]. Both CER and EBM require rigorous and systematic analysis of published literature to identify, appraise, select and synthesize *all* high quality research evidence relevant to a particular question.

However, traditional meta-analysis techniques can *only* enable a pairwise comparison at a time (usually, between placebo and an experimental drug). To understand comprehensively the performance of all possible interventions and to facilitate decision making, we have to compare them to one another simultaneously, not just to placebo or some particular standard treatment. *Network meta-analysis* (NMA), a meta-analytic statistical method, expands the scope of conventional pairwise meta-analysis to simultaneously compare multiple treatments in a connected network by synthesizing direct and indirect information [8]. Thus NMA provides the cornerstone for the recent explosion of CER and EBM.

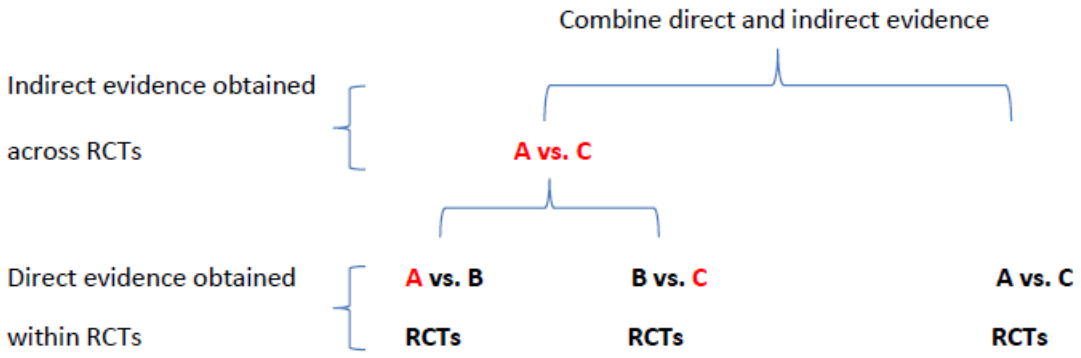

Figure 1.1: Illustration of direct and indirect information

1.1 Background and current development

We now first introduce the evidence synthesis process of NMA with a simple three treatment network as is shown in Figure 1.1 adapted from Li et al. [9]. Suppose the primary interest is the comparison of A versus C, NMA techniques allow us to combine evidence from trials directly comparing treatments A and C (AC trials), trials directly comparing A and B (AB trials), and trials directly comparing B and C (BC trials). AC trials are direct information while AB and BC trials are indirect information.

There is a broad consensus that the "best" evidence on the effect of treatment C relative to treatment A is provided by head-to-head trials, which provide a "direct" estimate. However, even if this is accepted, several reasons can be advanced for taking a wider view of what the legitimate evidence base should be. First, it may be that there are no A vs C trials, but that instead an "indirect" estimate can be formed from the results of A vs B and B vs C trials. For example, two new active treatments may have been compared with placebo, or to an established standard treatment, but manufacturers have proved reluctant to carry out the head-to-head comparisons that would be of most clinical interest. A second reason might be that, even if direct AC evidence exists, it may be sparse; the volume of indirect evidence can be much greater. This is, in fact, a very common situation. In a nutshell, by incorporating indirect evidence, NMA enables comparison of interest even if there is no direct information and strengthens inference when direct information is available.

There is by now a considerable literature for NMA, see, for example, by several working groups, including the Pharmaceutical Benefits Advisory Committee in Australia [10], the Canadian Agency for Drugs and Technologies in Health [11], and the National Institute for Health and Clinical Excellence (NICE) in the United Kingdom [12]. The most popular and to current method is the *contrast-based* (CB) Bayesian hierarchical modeling [13, 14, 15, 16, 17, 18]. This approach chooses one of the treatments as the baseline and focuses on estimating the relative treatment effects, e.g. the odds ratio (OR). Lu and Ades [16, 17] proposed Bayesian NMA models under the CB framework for a binary outcome and gained popularity in this field, and subsequently the NICE group (e.g., Dias et al. [19] and some others [20, 21, 22]) extended these models to other types of outcomes (e.g., continuous and count data). In addition, Salanti et al. [23] introduced an *arm-based* (AB) parameterization for NMAs and compared it with the CB parameterization.

There are two major issues for NMA: *heterogeneity* and *inconsistency* [15, 16, 17, 24, 25, 26, 27]. Heterogeneity aims to assess the dispersion of effect sizes from study to study, then take them into account when interpreting the data [28]. Since each study is conducted under different conditions and populations, study-specific effect sizes may vary even when they are drawn from an underlying population of study effects that has a common mean. If the effect size is consistent, then we will usually focus on the summary effect; if the effect size varies modestly, then we might still report the summary effect but note that the true effect in any given study could be somewhat lower or higher than this value; if the effect varies substantially from one study to the next, our attention will shift from the summary effect to the random effects, or even the dispersion itself. Random-effects models are usually used to take charge of heterogeneity.

Evidence inconsistency is usually defined, informally, as the disagreement between the direct and indirect point estimates of a particular comparison of treatments within a broader network of evidence. For simplicity, let us again use the three treatment network in Figure 1.1. This simple network has a single triangular closed loop of evidence shown in Figure 1.2. In the CB framework, consistency $d_{AB} = d_{CB} - d_{CA}$ is usually assumed, where d_{hk} represents the relative treatment effect between treatment h and treatment k. Inconsistency arises when this equality does not hold [15, 17, 29, 24, 30, 31]. In contrast, in the AB framework, instead of loop-based definition, Zhao et al. [32] used

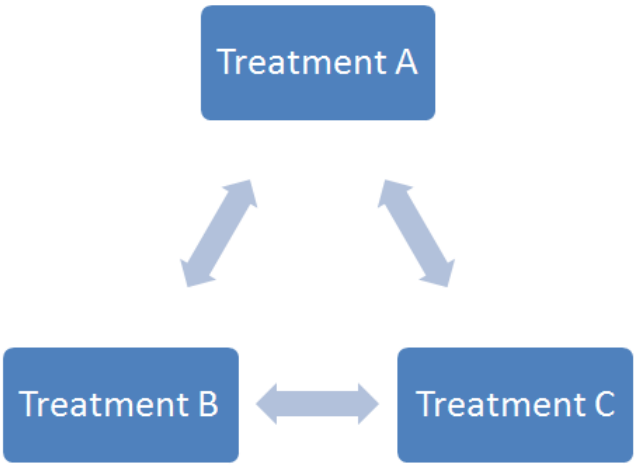

Figure 1.2: Loop for inconsistency

fixed and random effects to define and detect inconsistency.

1.2 Overview of the contribution in this thesis

A limitation of reporting for many current contrast-based NMA methods for binary outcomes is that the only summary statistic usually reported is OR[33][34][35][36][37][38][39][40][41]. ORs are often mistakenly thought as RRs by physicians, patients and their care givers, although it is well-known that RRs and ORs diverge when events are common (i.e., event rates are higher than 10%) [42][43][6][44]. Absolute measures including treatment-specific event rates and RDs contain important information that cannot be expressed by ORs [28]. Thus both relative measures and absolute measures should be reported and reporting only OR is not proper. However, to the best of our knowledge, only a few published NMAs [45][46] have reported RR, and none have reported the treatment-specific event rates and RDs. This limitation in reporting arises because many current statistical approaches and software [47][48][49][50][51][52][53][54][55] are not capable of estimating treatment-specific response proportions and summary statistics such as the risk difference (RD) and RR. We instead propose a novel arm-based Bayesian hierarchical method from the missing data perspective [8]. More specifically, we treat the sparsity as missing at random (MAR), which is totally different from Salanti et al. [23].

Our proposed method provides both direct estimates including event rates and RDs and indirect estimates including RRs and ORs.

A thorny problem in NMA involves nonignorable data missingness. It can happen due to deliberate study design. For example, clinicians often select treatments that have appeared to be more effective based on previous RCTs or their own personal medical experience, which may lead to a higher probability of missingness for relatively ineffective treatments. Another situation that can lead to nonignorable missingness is when meta-analysts undertaking an NMA selectively choose treatments to include in the analysis. For example, some NMAs exclude placebo or "no treatment" from consideration because it is sometimes believed that placebo trials vary over time; or are set in favorable conditions to appease regulatory authorities [56]. Other NMAs may include only the treatments available in particular location or time period, only those of perceived dose relevance, or (often in the case of industry submissions to health technology assessment bodies) only specific competing treatments [57][58]. In these cases, simply ignoring missingness (as the CB method does) or considering all missing data to be MAR as [8] can lead to bias [59]. To handle this problem, we jointly model the data and the missing indicator using a *selection models* approach.

A very interesting paper by Mills et al. [58] investigated empirically the effects of excluding treatments from NMA. It concluded that excluding treatments sometimes could have important effects on the results and could diminish the usefulness of the research to clinicians if important comparisons were missing. In the same vein, those undertaking an NMA will often selectively choose trials to include in the analysis. Inevitably, certain trials are more likely to be included in an NMA. In addition, it is difficult to include all existing trials that meet the inclusion criteria due to language barriers (i.e., some trials may be published using other languages) and other technical issues. Intuitively, if the omitted studies are a random subset of all relevant studies, the failure to include these studies will only result in less information, wider confidence intervals, and less powerful tests, but will not have any systematic impact on the NMA points estimates. However, if the omitted studies are systematically different from the ones we include, then these estimates will be biased. We thus obtained empirical evidence on whether these selective inclusions of trials could make a difference in the results such as treatment effect estimates in an NMA, where both the AB and CB methods were applied.

After investigating the influence of omission of trials that should have been included, we next study trials that are included but appear to deviate markedly from the others, and thus might actually be inappropriate to be synthesized. We call these trials "outlying" trials. To the best of our knowledge, while the key NMA assumptions of inconsistency and heterogeneity have been well-studied, few previous authors have considered the issue of trial-level outliers, their detection, and guidance on whether or not to discard them from an NMA. We thus propose four outlier detection measures for NMA in order to identify outlying trials and suggest to leave them out of the evidence synthesis.

The remainder of this thesis is structured as follows. First, Chapter 2 introduces the novel arm-based Bayesian hierarchical method under the missing at random assumption. We compare our approach to other alternative methods using two hypothetical NMA data sets, and then re-analyze two published network meta-analyses and show how more comprehensive and proper summary statistics can be reported using the proposed method. Chapter 3 presents our Bayesian *selection models* method aiming to handle the nonignorable missingness problem. We compare the proposed method, which models the observed data and missing indicator simultaneously, with the methods that consider the missingness as missing completely at random (MCAR) or MAR. In Chapter 4, we investigate empirically the impact of excluding trials that should have been included under both AB and CB framework. Chapter 5 proposes and evaluates Bayesian approaches to detect trial-level outliers in NMA evidence structures. The four detection measures are: *relative distance* (RD), Bayesian *standardized trial residual* (STR), *Bayesian p-value*, and *scale mixtures of normals* (SMN). Finally, Chapter 6 summarizes our findings and discusses potential areas for future work.

Chapter 2

Network meta-analysis of randomized clinical trials: reporting the proper summaries

In this chapter, we introduce the novel arm-based Bayesian hierarchical method which enables proper summaries, including direct summaries such as event rates and RDs, and indirect summaries such as RRs and ORs. This chapter begins in Section 2.1 with the background of reporting in NMAs. Section 2.2 provides details of this arm-based Bayesian NMA method and also introduces the existing CB method. Section 2.3 presents our data analysis results for two hypothetical data sets, aiming to compare the AB method with the CB method and traditional pairwise meta-analysis method. It also presents the re-analysis of two published NMAs. Section 2.4 concludes with a summary and discussion of limitations.

2.1 Background of reporting in NMA

To the best of our knowledge, only a few published NMAs [45][46] have reported RR, but none have reported the treatment-specific event rates and RDs. They focus on treatment contrasts where one of the arms of each study is chosen as "baseline". Since many NMAs do not have a common control arm such as a placebo or standard intervention

and different trials may have different "baselines", specifying a common distribution for baseline groups is generally not interpretable. Thus, many current NMA methods treat the underlying baseline risks as nuisance parameters and therefore fail to estimate the treatment-specific response proportions.

Although a few [51][22][60][61][62] discussed the transformation from the ORs to RRs and RDs, they depend on a strong assumption that either the event rate in a reference treatment group can be accurately estimated from some external data, or by summarizing only trials with the reference arm with a separate (random effects) model. In many cases, such external data are not available limiting the applicability of the former approach. Furthermore, even if some external data are available, it may come from a different population than what the NMA may represent. From the theory of missing data analysis [63], these current NMA methods are unbiased only under a strong assumption of missing completely at random (i.e., all trials randomly choose to include or not include the reference arm).

2.2 Statistical methods

Consider a collection of RCTs $i = 1, 2, \ldots, I$, each of which only includes a subset of the complete collection of K treatments. Let k_i be the number of treatments, and S_i be the set of treatments that are compared in the i^{th} trial. Trials with $k_i \geq 3$ are called "multi-arm" trials, in contrast to $k_i = 2$ for "two-arm" trials. For our binary data, let $D_i = (y_{ik}, n_{ik}), k \in S_i, i = 1, 2, \ldots, I$ denote the available data from the i^{th} trial, where n_{ik} is the total number of subjects and y_{ik} is the total number of responses for the k^{th} arm in the i^{th} trial. The corresponding probability of response is denoted by p_{ik}. In this section, we first briefly review the most commonly used contrast-based approach, then present our novel arm-based approach illustrating how to accurately estimate the overall treatment-specific event rates from the perspective of missing data analysis. At last, we evaluate the performance of a few alternative methods using two hypothetical examples.

2.2.1 The contrast-based (CB) approach

Let b_i be the specified "baseline" treatment for the i^{th} trial, commonly denoted as b for simplicity. Let $X_{ik} = 1$ if $k \neq b$ and $X_{ik} = 0$ if $k = b$. The The most commonly used CB models use the following Bayesian hierarchical model [53],

$$
\begin{aligned}
& y_{ik} \overset{\text{ind}}{\sim} \text{Bin}(n_{ik}, p_{ik}),\ i = 1, \ldots, I,\ k \in S_i, \\
& \text{logit}(p_{ik}) = \mu_i + X_{ik}\delta_{ibk}, \delta_{ibk} \overset{\text{ind}}{\sim} N(d_{bk}, \sigma^2_{bk}), \\
& d_{hk} = d_{bk} - d_{bh}, \\
& \text{Corr}(\delta_{ibh}, \delta_{ibk}) = \gamma^b_{hk},\ \ b \neq h \neq k \in S_i.
\end{aligned}
$$

where μ_i is the specified baseline effect that is commonly regarded as a nuisance parameter; X_{ik} is the indicator for baseline, taking value 0 when $k = b$ and 1 when $k \neq b$; $b(i)$ is the specified baseline treatment in trial i, commonly denoted as b for simplicity as above; and δ_{ibk} represents the contrast between treatment k and b for the i^{th} trial and is assumed to be a random effect with a normal distribution with mean d_{bk} and variance σ^2_{bk}. $d_{hk} = d_{bk} - d_{bh}$ represents consistency and $\gamma^{(b)}_{hk}$ represents the correlation between δ_{ibh} and δ_{ibk}.

2.2.2 The arm-based (AB) approach

We view the analytic challenges associated with NMA from the perspective of missing data analysis [63][59][64][65][66]. The basic idea of this arm-based approaches to NMA (which focus on modeling the event proportions for each treatment arm), in contrast to the contrast-based approaches (which focus on modeling the relative treatment effects, e.g., ORs, comparing treatments), has been briefly discussed by Salanti et al. [67], but thoroughly not from the missing data perspective. When viewed from this perspective, the proportion of patients responding to each treatment and associated summary statistics such as the RD, RR and OR can be estimated. Specifically, we assume that each study hypothetically compares all treatments, many of which are missing by design and thus can be considered as missing at random [59].

Specifically, we consider the multivariate Bayesian hierarchical mixed model (MBHMM), which extend the bivariate generalized linear mixed model for the meta-analysis of comparative studies of two arms [68]. First, we assume conditional on $P_i = p_{ik}$, the elements

y_{ik} of $Y = y_{ik}$ are independently binomially distributed with probability mass function

$$P(Y_i = y_i) \sim \prod_{k \in S_i} (p_{ik})^{y_{ik}} (1 - p_{ik})^{n_{ik} - y_{ik}}. \tag{2.1}$$

Second, we assume a multivariate normal distribution (MVN) for p_{ik} on a probit transformed scale. In the absence of any individual level covariates, the model is specified as

$$\Phi^{-1}(p_{ik}) = \mu_k + \sigma_k \nu_{ik}, \;\; (\nu_{i1}, \ldots, \nu_{iK})^T \sim MVN(0, R_K), \tag{2.2}$$

where $\Phi()$ is the standard normal cumulative distribution function, $(\mu_1, \ldots, \mu_K)$ are treatment-specific fixed effects, R_K is a positive definite correlation matrix with off-diagonal elements $\rho_{k_1 k_2}$, σ_k is the standard deviation for the random effects ν_{ik}. Let $diag(\sigma_1, \ldots, \sigma_K)$ be a diagonal matrix with elements σ_i, the covariance matrix is thus $\Sigma_K = \text{diag}(\sigma_1, \ldots, \sigma_K) \times R_K \times \text{diag}(\sigma_1, \ldots, \sigma_K)$. Here, σ_k captures trial-level heterogeneity in response to treatment k, and R_k captures the within-study dependence among treatments. The population-averaged event rate can then be calculated, for example with the probit link, as

$$\pi_k = E(p_{ik} | \mu_k, \sigma_k) = \int_{-\infty}^{\infty} \Phi(\mu_k + \sigma_k z) \phi(z) dz = \Phi(\mu_k / \sqrt{1 + \sigma_k^2}), k = 1, \ldots, K, \tag{2.3}$$

where $\Phi()$ is the standard normal cumulative density function and $\phi()$ is the standard normal density function. We can also use some other link functions, for example *logit* link, under which condition, the population-averaged event rate is $\pi_k \approx expit(\mu_k / \sqrt{1 + C^2 \sigma_k^2})$, where $C = 16\sqrt{3}/(15\pi)$ and $expit(x) = e^x/(1 + e^x)$. The marginal OR, RR, and RD are defined as $OR_{kl} = [\pi_k/(1 - \pi_k)]/[\pi_l/(1 - \pi_l)]$, $RR_{kl} = \pi_k/\pi_l$ and $RD_{kl} = \pi_k - \pi_l$ for a pairwise comparison between treatments k and l ($k \neq l$).

Since improper prior distributions may lead to an improper posterior in some complex models [69][70], we selected minimally informative but proper priors. Specifically, we chose a weakly informative prior $N(0, \tau_\mu^2)$ for μ_k with $\tau_\mu^2 = 1000$, and a Wishart prior for the precision matrix, i.e., $\Sigma_K^{-1} \sim W(V, n)$, where the degrees of freedom $n = K$, V is a known $K \times K$ matrix with diagonal elements equal 1.0, and off-diagonal elements equal 0.005. It turned out that the above prior corresponded to a 95% CI of 0.45 to 32.10 for the standard deviation parameters and a 95% CI of 1.00 to 1.00 for the correlation parameters, which is computed via simulations using the R function rWishart(). The

Washart distribution is the conjugate prior of the precision matrix of a multivariate-normal random vector in Bayesian statistics, which facilitates the computation of the unstructured posterior covariance matrix.

We implemented our method within a fully Bayesian framework using Markov chain Monte Carlo (MCMC) methods with the WinBUGS software [71][72]. Weakly informative priors were used and posterior samples were drawn using Gibbs and Metropolis-Hastings algorithms [73][73] with convergence assessed using trace plots, sample auto-correlations, and other standard convergence diagnostics [74][75]. A generous burn-in period of 1,000,000 iterations was used, with 1,000,000 subsequent iterations retained for accurate posterior treatment effect estimates. By borrowing information across multiple treatments, the multivariate Bayesian hierarchical mixed model that we utilize reduces potential bias when missing is not completely at random, compared to a naive approach of estimating population-averaged treatment-specific event proportions or rates based solely on studies that used a particular treatment. With this Bayesian approach, we used the 95% posterior credible intervals to assess statistical significance (according to whether the CI included the null value) instead of p-values [76]. The corresponding WinBUGS code is presented in the appendix.

2.2.3 Evaluation of different approaches

To investigate the performance of the proposed arm-based multivariate Bayesian hierarchical mixed model, we create two hypothetical network meta-analysis data sets under either a homogenous relative risk (RR) or a homogenous rate difference (RD) assumption. Each network meta-analysis includes 11 trials and 3 treatment arms. Because in a typical network meta-analysis, most trials only compare a subset of all treatments of interest, we let two trials compare all three treatments, and three trials each comparing A and B, B and C, A and C, respectively. The total numbers of patients are equal to 1000 for arm A, 2000 for arm B, and 500 for arm C in all trials. The response rates for arm A are assigned from a uniform distribution ranging from 0.10 to 0.40 in ascending order for the 11 trials. The corresponding numbers of responses for arm B and C in each trial are assigned based on a fixed RR or a fixed RD assumption. Specifically, the RR of B vs. A is 1.50 and C vs. A is 2.00 under the fixed RR assumption, and the RD of B vs. A is 15% and C vs. A is 25% under the fixed RD assumption. To simplify

illustration, we ignore the random sampling error and assume the number of events is equal to the response rates multiplied by the total number of patients.

We analyzed the above two hypothetical data using four methods. The first is based on Cochran-Mantel-Haenszel procedure with estimates of the log OR and variance as discussed in Yusuf et al. (we refer to this as Petos method) [77]. With this fixed effect method, inferences are based on the direct head-to-head pairwise comparisons. The second and third methods are the Lu & Ades contrast-based network meta-analysis method under either a homogeneous variance (i.e., the HOM model) or an unstructured heterogeneous variance assumption (i.e., the ID model)35. It combines the direct and indirect evidence, but it is not able to estimate the population-averaged treatment-specific event rates. The fourth is the arm-based network meta-analysis method that we have proposed. By borrowing information across treatment arms, it is able to estimate the treatment-specific event rates. The hypothetical data and the assumptions underlying these four methods are given in the web appendix wTable 1 and wTable 2, respectively.

2.3 Results

2.3.1 Comparison of four methods with hypothetical data

Table 2.1 presents the ORs based on the pairwise head-to-head comparisons for each hypothetical trial. The difference between the mean ORs from the observed data versus the mean ORs from the full data illustrates the potential bias of summarizing treatment effects based only on trials with particular treatment arms, i.e., the direct head-to-head comparisons. As evidenced by these two examples, the direction of bias can be either toward the null or away from the null, depending on the underlying data generating and missing data generating mechanisms, which limits the application and generalizability of methods based on direct head-to-head comparisons. For example, the true mean OR of B vs. A under a fixed RR assumption is 1.85, as compared to the mean OR of 1.66 based on the available direct head-to-head comparisons. The true mean OR of B vs. A under a fixed RD assumption is 2.15, as compared to the mean OR of 2.45 based on the available direct head-to-head comparisons.

Table 2.2 compares the population-averaged treatment-specific event rate estimates from the observed data vs. that from the full data based on the new method. It shows

Table 2.1: The odds ratios based on pairwise head-to-head comparisons

	I. Fixed RR			II. Fixed RD		
	B vs. A	**C vs. A**	**C vs. B**	**B vs. A**	**C vs. A**	**C vs. B**
Trial 1	1.59	2.25	1.42	3.00	4.85	1.62
Trial 2	1.62	2.35	1.45	2.60	4.10	1.58
Trial 3	1.66	***2.47***	***1.49***	2.36	***3.65***	***1.55***
Trial 4	1.70	***2.61***	***1.54***	2.20	***3.35***	***1.53***
Trial 5	1.75	***2.79***	***1.60***	2.08	***3.14***	***1.51***
Trial 6	***1.80***	3.00	***1.67***	***2.00***	3.00	***1.50***
Trial 7	***1.86***	3.27	***1.76***	***1.94***	2.90	***1.49***
Trial 8	***1.93***	3.63	***1.88***	***1.90***	2.83	***1.49***
Trial 9	***2.02***	***4.12***	2.04	***1.87***	***2.79***	1.50
Trial 10	***2.12***	***4.85***	2.28	***1.84***	***2.78***	1.51
Trial 11	***2.25***	***6.00***	2.67	***1.83***	***2.79***	1.52
Mean OR_1	1.66	2.90	1.97	2.45	3.54	1.54
Mean OR_2	1.85	3.40	1.80	2.15	3.29	1.53

OR=Odds Ratio; RR=Relative Risk; RD=Rate Difference; Mean OR_1 is the mean of ORs from the observed data assuming the italic cells are not available as in many NMAs; mean OR_2 is the mean of ORs from the full data assuming all the italic cells are observed and available.

that with this approach, estimates of the population-averaged treatment-specific event rates are nearly unbiased. In addition, the information loss due to missing data is mostly recovered as evidenced by the similarity of the length of the posterior credible intervals.

Table 2.3 compares the relative treatment effect estimates for the four methods using the observed data (which assume that the greyed cells in web appendix wTable 1 are not available as in many NMAs) and the full data (which assume that each trial has three arms and there is no missing arms), respectively. Under the hypothetical data generating mechanisms, all 4 model assumptions are incorrect, and the true ORs are not well defined. Thus, we choose the estimates from the full data as the true ORs under each model assumption. The closer the estimates from the observed data are to that from the full data, the less bias of the method. Under both fixed RR and fixed RD assumptions, Petos method is potentially biased since it incorporates only

Table 2.2: Population averaged event rate estimates under fixed RR and RD assumptions

	Event Rates	Treatment A	Treatment B	Treatment C
Fixed RR	True	**0.25**	**0.375**	**0.50**
	Observed data	0.25(0.19,0.34)	0.37(0.28,0.46)	0.50(0.38,0.61)
	Full data	0.25(0.19,0.31)	0.37(0.29,0.45)	0.50(0.38,0.59)
Fixed RD	True	**0.25**	**0.40**	**0.50**
	Observed data	0.24(0.18,0.33)	0.40(0.33,0.48)	0.50(0.43,0.57)
	Full data	0.25(0.19,0.32)	0.40(0.34,0.46)	0.50(0.43,0.56)

Results based on the proposed method; OR = Odds Ratio; RR = Relative Risk; RD = Rate Difference.

the direct information (the available head-to-head comparisons of two treatments). For example, under the fixed RR assumption, the estimated OR from Petos method is 1.63 comparing treatment B vs. A using the observed data set, while the corresponding OR from the full data set is 1.83 illustrating some biases. Lu & Ades contrast-based method shows potential biases, which is consistent with the results from simulation studies55. For example, under the fixed RR assumption, the estimated ORs of B vs. A from the observed data are 1.60 (95% CI 1.39, 1.81) and 1.66 (1.44, 1.85) under the Lu and Ades HOM and ID model assumptions, while the corresponding estimated ORs from the full data is 1.87 (1.66, 2.09) and 1.88 (1.75, 2.00), respectively. In contrast, using our proposed arm-based method, estimates for the ORs, RRs and RDs under both fixed RR and RD assumptions are nearly unbiased.

2.3.2 Re-analyses of two network meta-analyses recently published in The Lancet

Comparative efficacy and acceptability of 12 antidepressants

Cipriani et al. [34] Comprehensively summarized results of 117 randomized controlled trials (25,928 participants) from 1991 to 2007, and compared 12 new-generation antidepressants in terms of efficacy and acceptability in acute-phase treatment of major depression. The main outcomes were the proportions of patients who responded to a treatment or discontinued the allocated treatment (dropped out). Response was defined

as the total number of patients who had a reduction of at least 50% from baseline score at 8 weeks on the Hamilton depression rating scale (HDRS).

Table 2.4 presents a summary of the efficacy results using the proposed method. A similar table that only cited ORs and 95% CIs was reported by Cipriani et al6. The population-averaged treatment-specific response proportions are given in the diagonal entries in the table. These proportions range from 0.48 (95% CI 0.41 to 0.55) for reboxetine (REB) to 0.62 (95% CI 0.57 to 0.67) for mirtazapine (MIR). The upper and lower triangular panels report the RRs and RDs of all pairwise comparisons. Table 2.6 summarizes the treatment discontinuation proportions using the proposed method in the same format as the efficacy results. The population-averaged treatment-specific dropout rates (diagonal entries in the table) range from 0.21 for citalopram (CIT) (95% CI 0.17 to 0.26) and escitalopram (ESC) (95% CI 0.17 to 0.26) to 0.29 for REB (95% CI 0.23 to 0.37), fluoxetine (FVX) (95% CI 0.23 to 0.37), and milnacipran (MIL) (95% CI 0.21 to 0.37).

Table 2.3: Relative treatment effect estimates under fixed RR and RD

		Observed Data			Full Data		
		B vs. A	**C vs. A**	**C vs. B**	**B vs. A**	**C vs. A**	**C vs. B**
I. Fixed RR							
OR	Peto	1.63 (1.50,1.77)	3.06 (2.74,3.41)	1.93 (1.75,2.13)	1.83 (1.74,1.93)	3.36 (3.13,3.61)	1.78 (1.67,1.90)
	HOM	1.60 (1.39,1.81)	3.18 (2.75,3.64)	1.99 (1.73,2.31)	1.87 (1.66,2.09)	3.29 (2.89,3.71)	1.76 (1.50,2.06)
	ID	1.66 (1.44,1.85)	3.23 (2.66,4.07)	1.98 (1.56,2.44)	1.88 (1.75,2.00)	3.30 (2.78,3.90)	1.76 (1.48,2.09)
	New	1.72 (1.29,2.30)	2.97 (2.20,4.12)	1.74 (1.34,2.24)	1.78 (1.52,2.09)	2.96 (2.38,3.63)	1.66 (1.40,1.96)
RR	***True***	***1.50***	***2.00***	***1.33***	***1.50***	***2.00***	***1.33***
	New	1.45 (1.18,1.78)	2.00 (1.63,2.41)	1.33 (1.19,1.57)	1.50 (1.34,1.66)	2.00 (1.77,2.21)	1.33 (1.22,1.45)
II. Fixed RD							
OR	Peto	2.20 (2.03,2.37)	3.45 (3.10,3.83)	1.54 (1.41,1.69)	1.99 (1.89,2.09)	3.23 (3.01,3.46)	1.53 (1.44,1.63)
	HOM	2.28 (2.08,2.54)	3.36 (3.03,3.76)	1.47 (1.32,1.63)	2.06 (1.94,2.20)	3.15 (2.92,3.41)	1.53 (1.42,1.65)
	ID	2.31 (2.07,2.59)	3.40 (3.00,3.93)	1.47 (1.30,1.66)	2.07 (1.94,2.21)	3.16 (2.91,3.42)	1.53 (1.41,1.65)
	New	2.09 (1.48,2.85)	3.17 (2.28,4.31)	1.52 (1.19,1.95)	1.99 (1.66,2.36)	2.98 (2.44,3.59)	1.50 (1.28,1.75)
RD	***True***	***0.15***	***0.25***	***0.10***	***0.15***	***0.25***	***0.10***
	New	0.16 (0.09,0.22)	0.25 (0.19,0.32)	0.10 (0.04,0.16)	0.15 (0.11,0.18)	0.25 (0.21,0.28)	0.10 (0.06,0.14)

Note: HOM represent the contrast-based NMA with a homogeneous variance assumption; ID represents the contrast-based NMA with an unstructured heterogeneous variance assumption; New represnts our proposed arm-based NMA with an unstructured heterogeneous variance assumption.

Table 2.4: Population averaged responses rates (proportions), relative risks, and risk differences of the 12 antidepressants

	BUP	CIT	DUL	ESC	FLU	FVX	MIL	MIR	PAR	REB	SER	VEN
BUP	0.570 (0.522,0.615)	1.020 (0.901,1.164)	1.086 (0.937,1.268)	0.946 (0.853,1.049)	1.070 (0.972,1.175)	1.087 (0.944,1.263)	1.099 (0.897,1.358)	0.921 (0.824,1.030)	1.040 (0.940,1.149)	**1.189** **(1.015,1.417)**	0.970 (0.877,1.072)	0.953 (0.865,1.052)
CIT	0.011 (-0.059,0.084)	0.558 (0.499,0.615)	1.065) (0.908,1.250)	0.927 (0.830,1.030)	1.049 (0.937,1.161)	1.065 (0.918,1.240)	1.077 (0.869,1.337)	0.903 (0.795,1.018)	1.019 (0.905,1.138)	1.166* (0.987,1.384)	0.951 (0.843,1.062)	0.935 (0.829,1.043)
DUL	0.045 (-0.037,0.125)	0.034 (-0.053,0.118)	0.524 (0.457,0.595)	**0.871** **(0.757,0.992)**	0.984 (0.855,1.124)	1.001 (0.839,1.196)	1.012 (0.804,1.275)	**0.848** **(0.727,0.982)**	0.957 (0.834,1.088)	1.094* (0.905,1.336)	0.893* (0.770,1.029)	0.878* (0.757,1.009)
ESC	-0.033 (-0.092,0.028)	-0.044 (-0.105,0.017)	**-0.078** **(-0.150,-0.005)**	0.602 (0.557,0.646)	**1.131** **(1.040,1.229)**	**1.150** **(1.005,1.328)**	1.162 (0.950,1.430)	0.974 (0.878,1.081)	**1.099** **(1.007,1.199)**	**1.258** **(1.080,1.490)**	1.025 (0.935,1.125)	1.008 (0.921,1.103)
FLU	0.037 (-0.015,0.090)	0.026 (-0.034,0.084)	-0.008 (-0.078,0.065)	**0.070** **(0.022,0.118)**	0.533 (0.499,0.564)	1.016 (0.899,1.163)	1.027 (0.854,1.248)	**0.861** **(0.789,0.943)**	0.972 (0.904,1.045)	1.112* (0.964,1.302)	**0.907** **(0.840,0.982)**	**0.891** **(0.832,0.957)**
FVX	0.046 (-0.032,0.123)	0.034 (-0.047,0.114)	0.001 (-0.092,0.094)	**0.078** **(0.003,0.153)**	0.008 (-0.059,0.076)	0.524 (0.459,0.590)	1.010 (0.817,1.257)	**0.848** **(0.735,0.968)**	0.957 (0.834,1.084)	1.094* (0.907,1.327)	0.892* (0.776,1.017)	**0.877** (0.762,1.000)
MIL	0.051 (-0.064,0.155)	0.040 (-0.080,0.147)	0.006 (-0.119,0.121)	0.084 (-0.031,0.185)	0.014 (-0.091,0.106)	0.005 (-0.111,0.113)	0.518 (0.425,0.626)	0.838 (0.680,1.024)	0.946 (0.775,1.144)	1.083* (0.851,1.379)	0.882 (0.719,1.074)	0.868 (0.709,1.054)
MIR	-0.049 (-0.115,0.018)	-0.060 (-0.132,0.010)	**-0.094** **(-0.176,-0.011)**	-0.016 (-0.079,0.047)	**-0.086** **(-0.138,-0.033)**	**-0.094** **(-0.170,-0.019)**	-0.100 (-0.203,0.015)	0.619 (0.568,0.668)	**1.129** **(1.028,1.237)**	**1.291** **(1.102,1.535)**	1.053 (0.951,1.164)	1.036 (0.939,1.138)
PAR	0.022 (-0.034,0.078)	0.010 (-0.053,0.073)	-0.023 (-0.092,0.047)	**0.055** **(0.004,0.105)**	-0.015 (-0.054,0.024)	-0.024 (-0.093,0.045)	-0.029 (-0.125,0.078)	**0.071** **(0.016,0.125)**	0.548 (0.511,0.583)	1.144* (0.986,1.352)	0.933* (0.856,1.019)	**0.917** **(0.843,0.998)**
REB	**0.091** **(0.008,0.174)**	0.080* (-0.007,0.162)	0.045* (-0.050,0.143)	**0.123** **(0.043,0.204)**	0.054* (-0.019,0.125)	0.045* (-0.049,0.139)	0.040* (-0.077,0.166)	**0.140** **(0.055,0.223)**	0.069* (-0.008,0.146)	0.479 (0.408,0.549)	**0.816** **(0.689,0.948)**	**0.802** **(0.680,0.930)**
SER	-0.018 (-0.075,0.040)	-0.029 (-0.095,0.035)	−0.063* (-0.139,0.017)	0.015 (-0.039,0.070)	**-0.055** **(-0.099,-0.010)**	−0.063* (-0.135,0.009)	-0.069 (-0.169,0.043)	0.031 (-0.030,0.092)	−0.040* (-0.089,0.010)	**-0.108** **(-0.187,-0.029)**	0.588 (0.543,0.629)	0.983 (0.902,1.071)
VEN	-0.028 (-0.083,0.030)	-0.039 (-0.105,0.025)	−0.073* (-0.149,0.005)	0.005 (-0.049,0.059)	**-0.065** **(-0.105,-0.024)**	**-0.073** **(-0.146,-0.000)**	-0.079 (-0.177,0.032)	0.021 (-0.038,0.079)	**-0.050** **(-0.098,-0.001)**	**-0.118** **(-0.196,-0.040)**	-0.010 (-0.061,0.041)	0.598 (0.555,0.637)

Drugs are reported in alphabetical order. Diagonal panels are the population averaged response rates (i.e., proportion of patients who had at least 50% reduction from the baseline score on HDRS); upper triangular and lower triangular panels are the relative risks (RRs) and risk differences (RDs) of the first drug in alphabetical order compared with the second drug in alphabetical order, respectively. Drugs with higher response rate are more effective; RRs larger than 1.0 or positive RDs favor the first drug in alphabetical order. To obtain comparisons in the opposite direction, reciprocals should be taken for RR and opposite sign should be used for RD. Statistically significant results are in bold and underlined. Comparisons statistically significant here but not in Cipriani et al.[34] or vice versa are noted with *. For all summaries, we report both the Bayesian posterior medians and the 95% credible intervals. BUR=bupropion, CIT=citalopram, DUL=duloxetine, ESC=escitalopram, FLU=fluoxetine, FVX=fluvoxamine (FVX), MIL=milnacipran, MIR=mirtazapine, PAR=paroxetine, REB=reboxetine, SER=sertraline, and VEN=venlafaxine.

Table 2.5: Population averaged dropout rates (proportions), relative risks, and risk differences of the 12 antidepressantss

	BUP	CIT	DUL	ESC	FLU	FVX	MIL	MIR	PAR	REB	SER	VEN
BUP	0.252 (0.210,0.295)	1.202 (0.925,1.539)	0.921 (0.695,1.221)	1.200 (0.940,1.521)	0.982 (0.813,1.170)	0.879 (0.654,1.170)	0.874 (0.646,1.220)	1.002 (0.783,1.278)	0.951 (0.777,1.150)	0.869* (0.653,1.168)	1.170 (0.938,1.453)	0.957 (0.785,1.157)
CIT	0.042 (-0.018,0.098)	0.209 (0.173,0.257)	0.768 (0.575,1.032)	1.000 (0.783,1.267)	0.817 (0.663,1.008)	**0.732** **(0.544,0.985)**	0.728 (0.530,1.038)	0.834 (0.648,1.086)	**0.790*** **(0.638,0.986)**	**0.725** **(0.542,0.978)**	0.974 (0.777,1.232)	0.795 (0.635,1.005)
DUL	-0.021 (-0.100,0.050)	-0.063 (-0.140,0.007)	0.273 (0.216,0.343)	1.303* (0.990,1.693)	1.065 (0.831,1.350)	0.954 (0.684,1.318)	0.950 (0.671,1.372)	1.088 (0.813,1.445)	1.031 (0.812,1.299)	0.943 (0.679,1.315)	1.270* (0.968,1.656)	1.039 (0.796,1.343)
ESC	0.042 (-0.015,0.096)	0.000 (-0.052,0.051)	0.063* (-0.002,0.134)	0.209 (0.174,0.256)	0.817 (0.669,1.004)	**0.732** **(0.544,0.990)**	0.730 (0.533,1.031)	0.834 (0.649,1.088)	**0.790** **(0.645,0.982)**	**0.725** **(0.541,0.988)**	0.974 (0.779,1.231)	**0.797** **(0.642,0.996)**
FLU	-0.005 (-0.050,0.042)	-0.047 (-0.091,0.002)	0.017 (-0.045,0.087)	-0.047 (-0.089,0.001)	0.257 (0.232,0.283)	0.896 (0.699,1.147)	0.890 (0.687,1.206)	1.021 (0.844,1.248)	0.968 (0.855,1.098)	0.885* (0.698,1.149)	**1.192*** **(1.017,1.411)**	0.975 (0.851,1.120)
FVX	-0.035 (-0.121,0.040)	**-0.077** **(-0.160,-0.003)**	-0.013 (-0.109,0.077)	**-0.077** **(-0.161,-0.002)**	-0.030 (-0.109,0.034)	0.286 (0.225,0.366)	0.997 (0.708,1.432)	1.140 (0.859,1.524)	1.081 (0.841,1.395)	0.990 (0.712,1.391)	**1.332** **(1.014,1.755)**	1.089 (0.835,1.418)
MIL	-0.036 (-0.127,0.049)	-0.078 (-0.169,0.009)	-0.014 (-0.116,0.085)	-0.078 (-0.168,0.007)	-0.032 (-0.115,0.044)	-0.001 (-0.102,0.101)	0.289 (0.212,0.372)	1.146 (0.815,1.568)	1.087 (0.800,1.424)	0.994 (0.684,1.419)	1.338 (0.966,1.793)	1.095 (0.793,1.452)
MIR	0.001 (-0.062,0.061)	-0.042 (-0.101,0.019)	0.022 (-0.052,0.101)	-0.042 (-0.101,0.019)	0.005 (-0.046,0.053)	0.035 (-0.039,0.120)	0.037 (-0.051,0.130)	0.252 (0.208,0.301)	0.949 (0.774,1.155)	0.867 (0.647,1.175)	1.168 (0.929,1.463)	0.955 (0.770,1.177)
PAR	-0.013 (-0.063,0.037)	**-0.055*** **(-0.102,-0.004)**	0.008 (-0.052,0.077)	**-0.055** **(-0.100,-0.005)**	-0.008 (-0.042,0.024)	0.021 (-0.044,0.101)	0.023 (-0.054,0.108)	-0.014 (-0.063,0.039)	0.265 (0.235,0.298)	0.915 (0.709,1.201)	**1.231** **(1.035,1.473)**	1.007 (0.855,1.185)
REB	−0.038* (-0.122,0.039)	**-0.080** **(-0.161,-0.005)**	-0.016 (-0.110,0.077)	**-0.080** **(-0.162,-0.003)**	−0.033* (-0.109,0.034)	-0.003 (-0.099,0.096)	-0.002 (-0.107,0.103)	-0.038 (-0.124,0.042)	-0.025 (-0.104,0.046)	0.290 (0.225,0.366)	**1.346** **(1.014,1.764)**	1.101 (0.835,1.430)
SER	0.037 (-0.014,0.089)	-0.006 (-0.052,0.046)	0.058* (-0.007,0.132)	-0.005 (-0.052,0.046)	**0.041*** **(0.004,0.078)**	**0.071** **(0.003,0.153)**	0.073 (-0.008,0.160)	0.036 (-0.016,0.092)	**0.050** **(0.008,0.091)**	**0.074** **(0.003,0.154)**	0.215 (0.184,0.249)	**0.818*** **(0.677,0.984)**
VEN	-0.011 (-0.061,0.038)	-0.054 (-0.105,0.001)	0.010 (-0.057,0.085)	**-0.053** **(-0.102,-0.001)**	-0.007 (-0.043,0.029)	0.023 (-0.046,0.105)	0.025 (-0.057,0.113)	-0.012 (-0.065,0.044)	0.002 (-0.042,0.044)	0.026 (-0.046,0.107)	**-0.048*** **(-0.093,-0.004)**	0.263 (0.230,0.301)

Drugs are reported in alphabetical order. Diagonal panels are the population averaged dropout rate, upper triangular and lower triangular panels are the relative risks (RRs) and risk differences (RDs) of the first drug in alphabetical order compared with the second drug in alphabetical order, respectively. Drugs with lower dropout rate are more acceptable; RRs smaller than 1.0 or negative RDs favor the first drug in alphabetical order. To obtain comparisons in the opposite direction, reciprocals should be taken for RR and opposite sign should be used for RD. Statistically significant results are in bold and underlined. Comparisons statistically significant here but not in Cipriani et al[34]. or vice versa are noted with *. For all summaries, we report both the Bayesian posterior medians and the 95% credible intervals. BUR=bupropion, CIT=citalopram, DUL=duloxetine, ESC=escitalopram, FLU=fluoxetine, FVX=fluvoxamine (FVX), MIL=milnacipran, MIR=mirtazapine, PAR=paroxetine, REB=reboxetine, SER=sertraline, and VEN=venlafaxine.

ESC and sertraline (SER) were more effective and more acceptable as measured by the proportion responding and discontinuing treatment. MIR and VEN had good efficacy but low acceptability as measured by the proportion discontinuing treatment. CIT had high acceptability but low efficacy. To visually compare the efficacy and acceptability of the 12 antidepressant drugs, Figure 2.1 presents the treatment-specific posterior medians of response and dropout proportions, with their 95% posterior credible intervals.

As compared to the results of Cipriani et al. [34], for efficacy, we did not find significant differences between SER and DUL, FVX, and PAR, nor between VEN and DUL. REB was only less effective than BUP, ESC, MIR, SER, and VEN, but not other treatments. In terms of acceptability, both ESC and SER are better-tolerated than FVX, PAR, REB, and VEN. In addition, SER is better-tolerated than FLU. CIT is better-tolerated than not only FVX and REB, but also PAR. Lastly, we did not find significant differences comparing BUP versus REB, and DUL versus ESC and SER.

Figure 2.2 compares the ORs reported in Cipriani et al. [34] (y-axis) against the RRs estimated from our model (x-axis) of the 66 head-to-head comparisons of efficacy and treatment discontinuation. As expected, given how common the outcomes are, 81.1% (107/132) of the treatment effects are overestimated using the OR instead of the RR; only 18.9% (25/132) were underestimated. For efficacy, the overestimation can be as high as 57.4% (OR = 2.03 vs. RR = 1.29 comparing MIR vs. REB) while the underestimation is as high as 5.3% (OR = 1.00 vs. RR = 0.95 comparing MIL and PAR); for acceptability, the overestimation goes up to 28.7% (OR = 0.62 vs. RR = 0.87 comparing BUP vs. REB) while the underestimation can be as large as 19.2% (OR = 0.87 vs. RR = 0.73 comparing CIT and MIL). In addition, 7.6% (10/132) of the comparisons between ORs and RRs have opposite signs, for which both estimates are very close to the null (see red symbols in Figure 2). A direct comparison between the reported ORs in Cipriani et al6 and our marginal ORs is presented in the web appendix, and similar conclusions are shown.

Comparative efficacy and acceptability of antimanic drugs in acute mania

Cipriani et al. [35] comprehensively reviewed 68 randomized controlled trials (16,073 participants) from Jan 1, 1980 to Nov 25, 2010, which compared antimanic drugs at

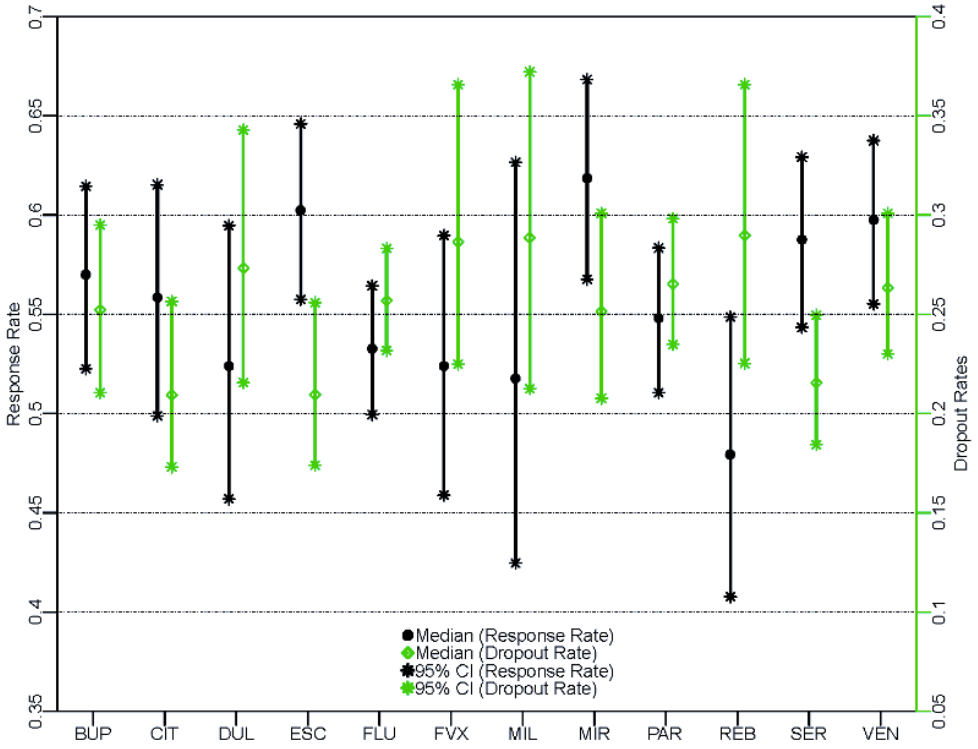

Figure 2.1: Response and dropout rates of the 12 antidepressants

therapeutic dose range for the treatment of acute mania in adults. The main outcomes were the mean change on mania rating scales and the proportion of patients who discontinued the assigned treatment at 3 weeks (dichotomous outcome for acceptability). The secondary outcome was response rate (response rate was defined as the proportion of the total number of patients who had a reduction of at least 50% on the total score between baseline and endpoint on a standardized rating scale for mania). Here, we only focus on the binary response for efficacy and the treatment discontinuation or dropout rate. Two treatments, gabapentin and asenapine that were only included in one or two trials were excluded.

Table 6 summarizes the efficacy results. The population-averaged treatment-specific response rates ranged from 0.22 (95% CI 0.08 to 0.48) for topiramate (TOP) to 0.56 for olanzapine (OLA) (95% CI 0.49 to 0.63) and haloperidol (HAL) (95% CI 0.48 to 0.64). Compared to placebo, RRs and RDs are significant for all antimanic treatments,

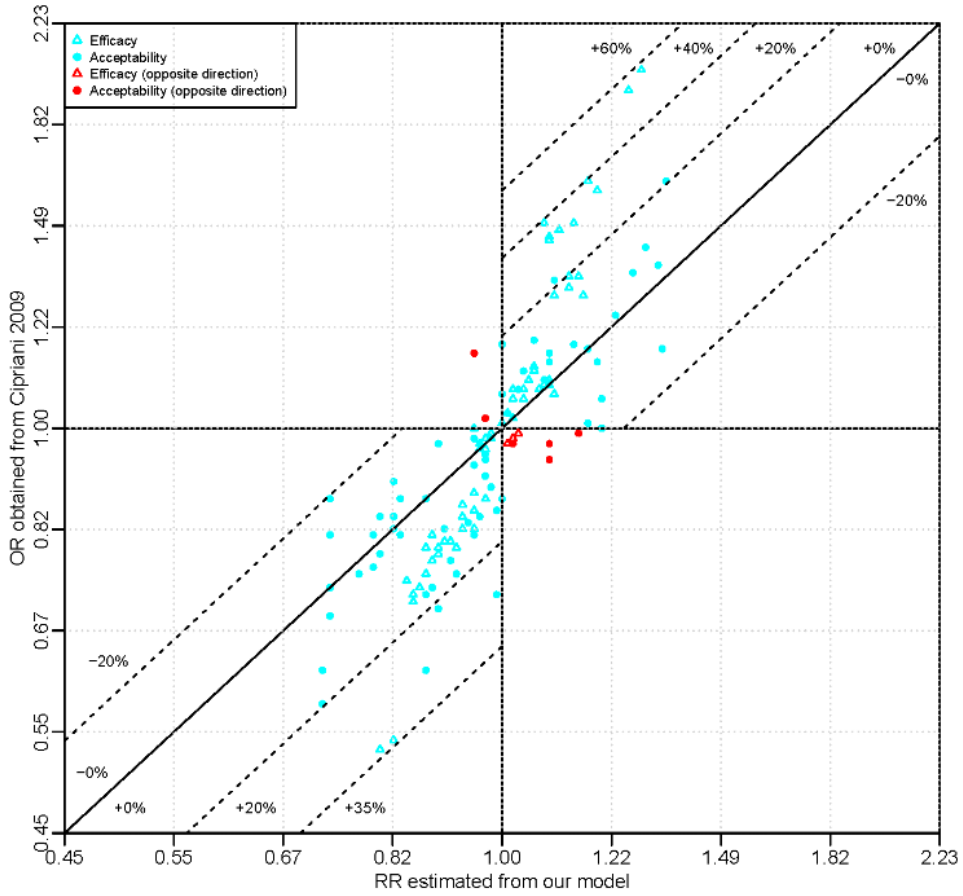

Figure 2.2: Comparison of the ORs versus the RRs for the 12 antidepressants

except lamotrigine (LAM) and TOP. In addition, all active treatments except LAM and ziprasidone (ZIP) are significantly more effective than TOP. Table 7 shows the results for acceptability (dropout). The population-averaged treatment-specific dropout proportions range from 0.30 for risperidone (RIS) (95% CI 0.24 to 0.37) and OLA (95% CI 0.25 to 0.36) to TOP at 0.48 (95% CI 0.32 to 0.65). The upper and lower triangular panels report the RRs and RDs of all pairwise comparisons.

Table 2.6: Population-averaged responses rates (proportions), relative risks, and risk differences of the 12 Antimanic drugs

	ARI	CAR	HAL	LAM	LIT	OLA	PLA	QUE	RIS	TOP	VAL	ZIP
ARI	0.504 (0.437,0.574)	0.956 (0.720,1.374)	0.903* (0.750,1.081)	0.949 (0.607,2.242)	0.918 (0.749,1.126)	0.897 (0.749,1.074)	**1.369** (1.167,1.586)	0.925 (0.766,1.117)	0.917 (0.758,1.106)	**2.332** (1.046,6.538)	0.955 (0.780,1.169)	1.055 (0.858,1.326)
CAR	-0.023 (-0.183,0.145)	0.528 (0.376,0.675)	0.944* (0.658,1.258)	0.994 (0.577,2.381)	0.959 (0.666,1.295)	0.939 (0.658,1.240)	**1.433** (1.006,1.863)	0.967 (0.677,1.286)	0.959 (0.667,1.274)	**2.427** (1.034,6.899)	0.998 (0.695,1.342)	1.105 (0.763,1.507)
HAL	−0.054* (-0.153,0.041)	−0.031* (-0.202,0.134)	0.559 (0.482,0.637)	1.053* (0.670,2.486)	1.017* (0.824,1.254)	0.995 (0.832,1.183)	**1.517** **(1.289,1.767)**	1.026* (0.845,1.239)	1.017 (0.840,1.221)	**2.587** **(1.156,7.262)**	1.059* (0.861,1.299)	1.171* (0.950,1.456)
LAM	-0.027 (-0.314,0.285)	-0.003 (-0.321,0.334)	0.028* (-0.262,0.341)	0.531 (0.226,0.810)	0.964 (0.409,1.526)	0.944* (0.401,1.478)	1.440 (0.612,2.236)	0.974 (0.412,1.528)	0.965* (0.408,1.516)	2.405 (0.762,7.371)	1.004 (0.424,1.589)	1.110 (0.469,1.778)
LIT	-0.045 (-0.154,0.061)	-0.023 (-0.196,0.148)	0.009* (-0.108,0.123)	-0.020 (-0.334,0.274)	0.550 (0.463,0.639)	0.978 (0.804,1.185)	**1.491** **(1.239,1.778)**	1.007 (0.825,1.239)	1.000 (0.806,1.233)	**2.544** **(1.131,7.142)**	1.041 (0.838,1.288)	1.149 (0.918,1.468)
OLA	-0.058 (-0.152,0.038)	-0.034 (-0.200,0.127)	-0.003 (-0.101,0.094)	−0.031* (-0.343,0.258)	-0.012 (-0.117,0.096)	0.563 (0.490,0.630)	**1.524** **(1.325,1.753)**	1.031 (0.861,1.240)	1.022 (0.855,1.221)	**2.598** **(1.165,7.300)**	1.065* (0.889,1.275)	1.177* (0.963,1.464)
PLA	**0.136** **(0.064,0.207)**	**0.160** **(0.002,0.308)**	**0.190** **(0.111,0.270)**	0.162 (-0.144,0.445)	**0.181** **(0.091,0.274)**	**0.194** **(0.124,0.264)**	0.368 (0.331,0.409)	**0.676** **(0.578,0.798)**	**0.669** **(0.573,0.789)**	1.702 (0.772,4.795)	**0.698** **(0.589,0.832)**	**0.771** **(0.650,0.936)**
QUE	-0.041 (-0.140,0.057)	-0.018 (-0.184,0.146)	0.014* (-0.092,0.117)	-0.014 (-0.328,0.275)	0.004 (-0.103,0.119)	0.017 (-0.083,0.117)	**-0.177** (-0.256,-0.097)	0.546 (0.470,0.620)	0.991 (0.820,1.201)	**2.527** **(1.129,7.062)**	1.033 (0.837,1.266)	1.142 (0.926,1.429)
RIS	-0.046 (-0.146,0.052)	-0.023 (-0.192,0.140)	0.009 (-0.096,0.110)	−0.019* (-0.333,0.271)	-0.000 (-0.116,0.116)	0.012 (-0.087,0.109)	**-0.182** (-0.261,-0.103)	-0.005 (-0.108,0.099)	0.550 (0.475,0.628)	**2.547** **(1.135,7.163)**	1.042* (0.847,1.279)	1.151* (0.934,1.441)
TOP	**0.287** **(0.022,0.445)**	**0.306** **(0.015,0.514)**	**0.341** **(0.073,0.506)**	0.305 (-0.087,0.630)	**0.332** **(0.061,0.501)**	**0.344** **(0.078,0.504)**	0.151 (-0.108,0.300)	**0.328** **(0.060,0.488)**	**0.332** **(0.064,0.496)**	0.216 (0.078,0.478)	**0.409** **(0.146,0.912)**	0.453* (0.161,1.020)
VAL	-0.024 (-0.131,0.079)	-0.001 (-0.172,0.166)	0.031* (-0.082,0.140)	0.002 (-0.314,0.294)	0.022 (-0.094,0.137)	0.034* (-0.065,0.129)	**-0.159** **(-0.247,-0.077)**	0.018 (-0.096,0.125)	0.022* (-0.090,0.131)	**-0.310** **(-0.477,-0.045)**	0.528 (0.449,0.615)	1.106 (0.885,1.403)
ZIP	0.026 (-0.076,0.134)	0.050 (-0.121,0.218)	0.082* (-0.027,0.189)	0.053 (-0.262,0.346)	0.071 (-0.044,0.195)	0.085* (-0.020,0.189)	**-0.109** **(-0.191,-0.026)**	0.068 (-0.040,0.176)	0.072* (-0.035,0.182)	−0.260* (-0.424,0.009)	0.050 (-0.061,0.167)	0.479 (0.394,0.557)

Drugs are reported in alphabetical order. Diagonal panels are the population-averaged response rate; upper triangular and lower triangular panels are the relative risks (RRs) and risk differences (RDs) of the first drug in alphabetical order compared with the second drug in alphabetical order, respectively. Drugs with higher response rate are more effective; RRs larger than 1.0 or positive RDs favor the first drug in alphabetical order. To obtain comparisons in the opposite direction, reciprocals should be taken for RR and opposite sign should be used for RD. Statistically significant results are in bold and underlined. Comparisons statistically significant here but not in Cipriani et al[35]. or vice versa are noted with *. For all summaries, we report both the Bayesian posterior medians and the 95% credible intervals. ARI=aripiprazole, CAR=carbamazepine, HAL=haloperidol, LAM=lamotrigine, LIT=lithium, OLA=olanzapine, PLA=placebo, QUE=quetiapine, RIS=risperidone, TOP=topiramate, VAL=valproate, and ZIP=ziprasidone.

Table 2.7: Population-averaged dropout rates (proportions), relative risks, and risk differences of the 12 Antimanic drugs

	ARI	CAR	HAL	LAM	LIT	OLA	PLA	QUE	RIS	TOP	VAL	ZIP
ARI	0.368 (0.295,0.443)	1.013 (0.714,1.507)	0.960 (0.746,1.254)	0.884 (0.617,1.341)	1.012 (0.786,1.304)	1.219 (0.950,1.564)	0.902 (0.737,1.078)	1.152 (0.813,1.687)	1.224 (0.928,1.613)	0.761* (0.538,1.177)	1.067 (0.807,1.416)	0.911 (0.702,1.176)
CAR	0.005 (-0.133,0.134)	0.363 (0.249,0.489)	0.949 (0.641,1.342)	0.870 (0.552,1.409)	0.997 (0.684,1.401)	1.202 (0.818,1.696)	0.889 (0.620,1.201)	1.132 (0.727,1.782)	1.206 (0.803,1.748)	0.753* (0.476,1.222)	1.051 (0.707,1.513)	0.898 (0.603,1.272)
HAL	-0.015 (-0.111,0.084)	-0.020 (-0.149,0.118)	0.381 (0.307,0.469)	0.918 (0.638,1.404)	1.050 (0.812,1.371)	1.267* (0.998,1.616)	0.934 (0.767,1.139)	1.192 (0.856,1.774)	1.269 (0.978,1.677)	0.792* (0.557,1.225)	1.108 (0.843,1.473)	0.945 (0.733,1.228)
LAM	-0.048 (-0.207,0.102)	-0.054 (-0.231,0.128)	-0.034 (-0.195,0.121)	0.416 (0.284,0.564)	1.145 (0.765,1.620)	1.382* (0.912,1.958)	1.020 (0.695,1.386)	1.302 (0.817,2.041)	1.384* (0.908,2.001)	0.864 (0.537,1.396)	1.207 (0.786,1.757)	1.030 (0.680,1.471)
LIT	0.004 (-0.086,0.097)	-0.001 (-0.122,0.134)	0.018 (-0.076,0.120)	0.052 (-0.092,0.207)	0.363 (0.296,0.437)	1.207* (0.939,1.532)	0.891 (0.729,1.065)	1.140* (0.811,1.639)	1.211* (0.916,1.589)	0.754 (0.532,1.146)	1.054 (0.801,1.387)	0.900 (0.699,1.159)
OLA	0.066 (-0.017,0.151)	0.061 (-0.058,0.194)	0.080* (-0.001,0.169)	0.115* (-0.028,0.267)	0.062* (-0.020,0.143)	0.301 (0.245,0.363)	**0.739 (0.621,0.869)**	0.943 (0.681,1.366)	1.003 (0.780,1.303)	**0.624 (0.447,0.959)**	0.875 (0.688,1.113)	**0.746* (0.586,0.955)**
PLA	-0.040 (-0.109,0.031)	-0.045 (-0.156,0.080)	-0.027 (-0.097,0.056)	0.008 (-0.127,0.155)	-0.044 (-0.114,0.026)	**-0.106 (-0.157,-0.053)**	0.408 (0.364,0.452)	1.277* (0.970,1.785)	**1.359 (1.113,1.680)**	0.844 (0.632,1.256)	1.183 (0.974,1.470)	1.010 (0.850,1.221)
QUE	0.048 (-0.076,0.165)	0.042 (-0.105,0.197)	0.061 (-0.058,0.187)	0.096 (-0.070,0.268)	0.045* (-0.078,0.154)	-0.018 (-0.132,0.087)	0.088* (-0.013,0.181)	0.318 (0.226,0.433)	1.064 (0.733,1.477)	0.662* (0.434,1.056)	0.927 (0.626,1.321)	0.790 (0.541,1.109)
RIS	0.067 (-0.025,0.158)	0.062 (-0.065,0.201)	0.081 (-0.007,0.178)	0.115* (-0.030,0.271)	0.063* (-0.029,0.152)	0.001 (-0.077,0.078)	**0.107 (0.041,0.168)**	0.019 (-0.088,0.131)	0.300 (0.238,0.374)	**0.621 (0.436,0.968)**	0.872 (0.655,1.164)	**0.743* (0.566,0.977)**
TOP	−0.115* (-0.291,0.060)	−0.119* (-0.316,0.078)	−0.100* (-0.279,0.077)	-0.065 (-0.278,0.142)	-0.118 (-0.296,0.049)	**-0.181 (-0.352,-0.014)**	-0.075 (-0.238,0.084)	−0.162* (-0.342,0.020)	**-0.182 (-0.355,-0.011)**	0.482 (0.324,0.654)	1.401* (0.898,2.010)	1.193 (0.778,1.696)
VAL	0.023 (-0.077,0.121)	0.018 (-0.110,0.156)	0.037 (-0.062,0.140)	0.071 (-0.081,0.231)	0.019 (-0.080,0.112)	-0.043 (-0.125,0.033)	0.063 (-0.011,0.131)	-0.025 (-0.145,0.098)	-0.044 (-0.138,0.048)	0.138* (-0.038,0.314)	0.345 (0.269,0.426)	0.854 (0.645,1.116)
ZIP	-0.036 (-0.134,0.062)	-0.041 (-0.172,0.100)	-0.022 (-0.119,0.082)	0.012 (-0.139,0.173)	-0.041 (-0.137,0.056)	**-0.102* (-0.188,-0.016)**	0.004 (-0.070,0.076)	-0.084 (-0.201,0.040)	**-0.103* (-0.197,-0.008)**	0.078 (-0.096,0.260)	-0.059 (-0.158,0.042)	0.404 (0.331,0.480)

Drugs are reported in alphabetical order. Diagonal panels are the population-averaged dropout rate; upper triangular and lower triangular panels are the relative risks (RRs) and risk differences (RDs) of the first drug in alphabetical order compared with the second drug in alphabetical order, respectively. Drugs with lower dropout rate are more acceptable; RRs lower than 1.0 or negative RDs favor the first drug in alphabetical order. To obtain comparisons in the opposite direction, reciprocals should be taken for RR and opposite sign should be used for RD. Statistically significant results are in bold and underlined. Comparisons statistically significant here but not in Cipriani et al.[35]. or vice versa are noted with *. For all summaries, we report both the Bayesian posterior medians and the 95% credible intervals. ARI=aripiprazole, CAR=carbamazepine, HAL=haloperidol, LAM=lamotrigine, LIT=lithium, OLA=olanzapine, PLA=placebo, QUE=quetiapine, RIS=risperidone, TOP=topiramate, VAL=valproate, and ZIP=ziprasidone.

To visually compare the efficacy and acceptability of the 12 antimanic drugs, Figure 2.3 plots the treatment-specific posterior medians of the response and dropout proportions, with their 95% posterior credible intervals. The 95% credible intervals of LAM and TOP are extremely wide because they are studied in only 3 and 5 trials respectively, much fewer than the others. TOP is less effective and less well tolerated than placebo.

Our results differ from Cipriani et al. [35] in some aspects. For efficacy, we do not find significant differences between HAL, RIS, and OLA with the other treatments, while in Cipriani et als paper.7, HAL, RIS, and OLA showed significant efficacy compared with some other treatments. For acceptability, except that OLA and RIS have significantly lower proportions of discontinuation compared to placebo, TOP, and ZIP, we do not find any other statistically significant head-to-head comparisons. In contrast, Cipriani et al. [35] found that OLA, RIS, and quetiapine (QUE) led to significantly fewer discontinuations than did lithium (LIT), LAM, placebo, and TOP.

Figure 2.4 compares the ORs reported in Cipriani et al. [35] (y-axis) against the RRs estimated from our model (x-axis) of the 66 head-to-head comparisons for treatment discontinuation (acceptability) and the 11 comparisons with placebo for efficacy. Overall, 90.9% (70/77) of the treatment effects are overestimated, and 9.1% (7/77) of them are underestimated. Specifically, for efficacy, the overestimation is as high as 74.8% (OR = 1/0.40 = 2.50 vs. RR = 1.43 comparing CAR vs. placebo) while the underestimation is as high as 30.5% (OR = 1/1.30 = 0.77 vs. RR =1/1.70 = 0.59 comparing TOP and placebo). For acceptability, the overestimation is as large as 54.3% (OR = 1/0.47 = 2.13 vs. RR = 1.38 comparing LAM vs. OLA), while the underestimation is as large as 18.0% (OR=1.05 vs. RR=0.89 comparing LIT and placebo). In addition, 6.1% (4/66) of the comparisons between the RRs and the ORs for acceptability are in the opposite direction of the null (red plotting symbols in Figure 4). A direct comparison between the reported ORs in Cipriani et al. [34] and our marginal ORs is presented in the web appendix, and similar conclusions are shown.

2.4 Discussion

Network meta-analysis is increasingly utilized to synthesize direct and indirect evidence for different treatments. However, many current network meta-analyses focus on

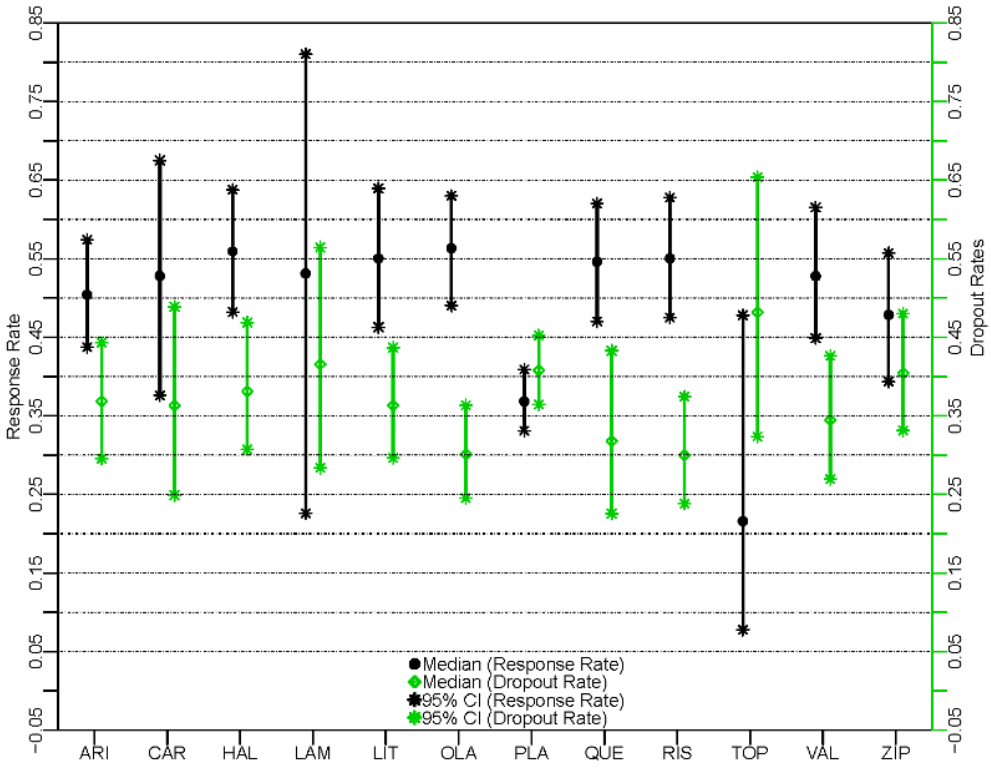

Figure 2.3: Response and dropout rates of the 12 antimanic drugs

treatment contrasts, in which one of the arms of each study is chosen as "baseline". Since different studies may have different "baselines", as a consequence of changing standards of care or changes in the underlying risks of study populations (e.g., initial trial may include more severely ill patients), specifying a common distribution for baseline groups is generally not interpretable. Although one may prefer to leave the baseline treatment as a fixed, study-specific parameter with the argument that they are fundamentally different from each other. However, while we make a relatively strong assumption on exchangeability of the probability of events within each treatment group across studies, our model is valid under the missing at random (MAR) assumption. The contrast-based Lu and Ades approach is valid only under a missing completely at random (MCAR) assumption, as shown in a recent AHRQ report (http://www.ncbi.nlm.nih.gov/books/NBK116689/pdf/TOC.pdf) and a corresponding technical report [78]. In addition, many current NMA methods only report the relative

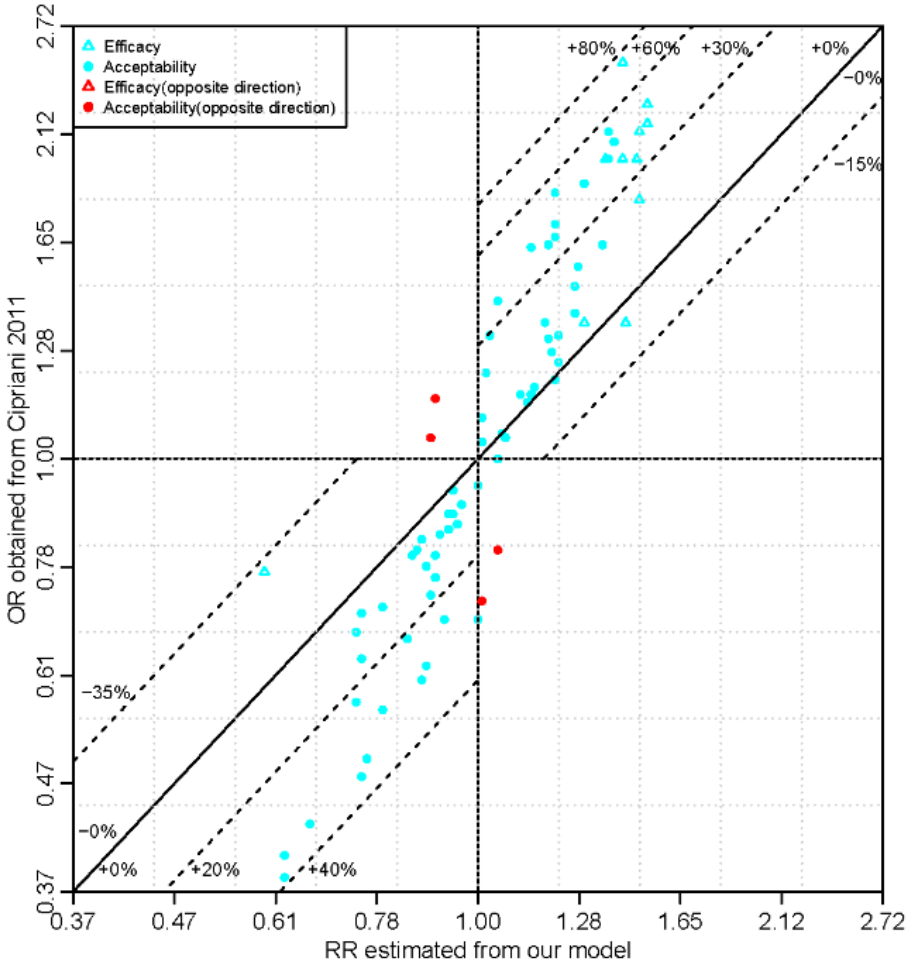

Figure 2.4: Comparison of the ORs versus the RRs for the 12 antimanic drugs

treatment effect on an OR scale [47][48][49][16][17][79][80][55]. Although they do offer valid statistical significance testing concerning the OR and can incorporate data from studies that only report relative treatment effects, without making strong assumptions on the event rate in a reference group, they fail to report treatment-specific event rates, risk differences and relative risks, which should be considered in making treatment recommendations. Although in some cases, it is unfortunate that some people tend to misspecify the distribution for the reference group and sometimes can lead to incorrect inference and interpretation, it should not construed to against our effort to estimate and report treatment-specific event rates. With the two comprehensive overviews, we

illustrate how this novel arm-based Bayesian hierarchical model can be used to estimate these key statistics, and in some circumstances lead to different conclusions.

For the two NMAs [34][35] considered, relatively high response proportions (up to 0.62) were observed. The differences between ORs and RRs that we illustrate can be explained in large part by the theoretical difference between the OR and the RR for common events [81]. The limitation of only reporting the ORs is discussed in detail in the web appendix. There is also a theoretical difference between the marginal treatment effects averaged over all studies by our approach, and the conditional treatment effects reported for a typical NMA by the contrast-based approaches such as used by Cipriani et al. [34][35]. Marginal treatment effects are generally smaller than the conditional treatment effects estimated from random effects models [82]. Finally, our differing ORs and RRs may partially be the result of the potential difference between model assumptions (e.g., the assumed variance and correlation structure) and the potential bias using current contrast-based models as illustrated in the hypothetical data analyses.

To compare the performance of the proposed arm-based versus current contrast-based Bayesian hierarchical models, we create two hypothetical network meta-analysis data sets including 11 trials and 3 treatment arms under either a homogenous RR or a homogenous RD assumption, in which the full data sets (i.e., assuming each trial compares all treatment arms) are available to estimate the true parameters (see details in the Web appendix). We found that the proposed arm-based NMA method outperformed the current contrast-based NMA methods.

In addition to some common concerns of network meta-analysis [83][38][67], there are some additional limitations for the proposed network meta-analysis approaches. First, to facilitate the estimation of treatment-specific population-averaged event proportions, we assume that each study hypothetically compares all treatments, with unstudied arms being missing at random conditional on the observed arms. Such models allow us to borrow information across multiple treatments within studies to reduce potential bias. However, it is plausible that investigators may have selected treatment arms on purpose based on the results of previous trials, which may lead to nonignorable missingness and potentially bias our event rate estimation. In addition, to robustly estimate event rates for each treatment, it is very important to have adequate number of trials with

adequate samples for each treatment in a network meta-analysis. Different model assumptions may lead to different results in poorly connected networks. Second, in this article, we only considered a saturated multivariate Bayesian hierarchical mixed model with unstructured variance-covariance matrix. Although various model simplifications gave similar results (not presented), we did not perform analysis over all possible reduced models (e.g. models with equal variances, and/or equal correlations among all treatments), a number of which may further improve statistical efficiency. Arguably, the unstructured variance-covariance matrix allows us to better summarize the evidence contained in the data without enforcing an artificial structure, such as equal variances or equal correlations. Third, in addition to the evaluation of heterogeneity of treatment effects, inconsistency is a major concern in network meta-analysis. Much ongoing debate over the value of network meta-analysis concerns the agreement between the direct and indirect evidence. In addition, inconsistency and its trade-off with heterogeneity can be very important when selecting the scale for NMA [83]. Achana et al.[84] has proposed an important method to adjust for baseline imbalance in order to possibly reduce heterogeneity and inconsistency for the CB methods. Some statistical methods have been proposed for identifying this disagreement when using contrast-based approaches with the odds ratio as the main effect measure [17][27][85][26][86], statistical methods for identifying and accounting for potential inconsistency based on our proposed models, formulated from the missing data perspective, await further development. Finally, in this paper, we do not consider individual-level or study-level covariates, which has already been briefly discussed elsewhere [87][88].

In summary, we have proposed and implemented a novel arm-based multiple-treatments meta-analysis in a Bayesian framework, which is different than the methods used by Cipriani in two NMAs [34][35]. With this arm-based approach, estimates of treatment-specific event rates or proportions, RDs and RRs are provided. Using two hypothetical data sets, we show that our method provides more accurate estimates than the methods used by Cipriani et al. [34][35]. Such differences could lead to different treatment recommendations.

Chapter 3

Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness

In this chapter, we extend the missing at random (MAR) model proposed in Chapter 2 to incorporate nonignorable missingness using *selection models* approach. Section 3.1 introduces the proposed method which jointly models the observed data and missingness indicator. We also show various important model simplifications. In Section 3.2, we apply our proposed model to a smoking cessation data set, and compare the results with those obtained from models assuming ignorable missingness. We then conduct various simulation studies in Section 3.3 to investigate the performance of our proposed method in handling the nonignorable missingness. We also compare comprehensively the performance of the AB method and the CB method using extensive simulation studies in this section. Finally, Section 3.4 summarizes our findings and conclusions.

3.1 Statistical methods

Nonignorable missingness (or MNAR) is inevitable when selectively choosing treatments to include in trials or selectively choosing trials to include in an NMA. Unfortunately, one can never tell from the data at hand whether the missing values are MAR or MNAR [59]. The fundamental difficulty is that potential "lurking variables" controlling the missingness are unobserved—by definition—and so we can never rule them out. Rather than trying to test whether the missing values are MAR, we develop a *sensitivity* analysis tool to explore how inferences may change if the assumption of MAR is violated. Note that even if an assumption of MAR seems reasonable, it is worthwhile to investigate how the results may change under nonignorable missingness.

We now introduce a partition of the complete data Y into observed values, Y_{obs}, and missing values, Y_{mis}, i.e., $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. Let R be a $I \times K$ indicator matrix for missingness in a NMA containing I trials and K treatments. Then $(Y_{\text{obs}}, Y_{\text{mis}}, R)$ and (Y_{obs}, R) are referred to as the *complete* data and *observed* data, respectively. The practical implication of MNAR is that the likelihood requires an explicit model for R. *Selection* models, which were introduced by the econometrician Heckman [89], provide this explicit form. They factor the joint distribution of the complete data and the missingness indicators into a marginal density for the complete data and a conditional density for the missingness indicators given the complete data, i.e., $f(Y_{\text{obs}}, Y_{\text{mis}}, R|\theta, \alpha) = f(Y_{\text{obs}}, Y_{\text{mis}}|\theta, \alpha) f(R|Y_{\text{obs}}, Y_{\text{mis}}, \theta, \alpha)$, where θ is the parameter for the complete data and α is the parameter for the missingness mechanism. This factorization can usually be simplified to $f(Y_{\text{obs}}, Y_{\text{mis}}, R|\theta, \alpha) = f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) f(R|Y_{\text{obs}}, Y_{\text{mis}}, \alpha)$, if we assume that $Y|\theta$ is conditionally independent of α, and $R|Y_{\text{obs}}, Y_{\text{mis}}, \alpha$ is conditionally independent of θ, which is usually reasonable in practice. We in further call $f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$ the *model of interest* (MOI) and $f(R|Y_{\text{obs}}, Y_{\text{mis}}, \alpha)$ the *model of missingness* (MOM) [90].

Consider a collection of RCTs $i = 1, 2, \ldots, I$, each of which only includes a subset of the complete collection of K treatments. Let k_i be the number of treatments, and S_i be the set of treatments that are compared in the i^{th} trial. Trials with $k_i \geq 3$ are called "multi-arm" trials, in contrast to $k_i = 2$ for "two-arm" trials. For our binary data, let $D_i = (y_{ik}, n_{ik}), k \in S_i, i = 1, 2, \ldots, I$ denote the available data from the i^{th} trial, where n_{ik} is the total number of subjects and y_{ik} is the total number of responses for the k^{th}

arm in the i^{th} trial. The corresponding probability of response is denoted by p_{ik}.

For the MOI, we consider the multivariate generalized linear mixed model (MGLMM) as in Chapter 2. $(Y_{\text{obs}}, Y_{\text{mis}})$ can be written as a collection of vectors $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I)$ for I trials, where each vector $\boldsymbol{y}_i$ contains two parts, $\boldsymbol{y}_{i,\text{obs}}$ and $\boldsymbol{y}_{i,\text{mis}}$. The elements y_{ik} of $\boldsymbol{y}_i = \{y_{ik}\}$ are assumed to be independently binomially distributed with probability p_{ik}. A multivariate normal distribution (MVN) for p_{ik} on a transformed scale, similar to that in Chapter 2, is then assumed as follows:

$$\text{g}(p_{ik}) = \mu_k + \nu_{ik}, \;\; (\nu_{i1}, \ldots, \nu_{iK})^T \sim MVN(0, \Sigma_k), \tag{3.1}$$

where $g()$ is some appropriate link function, μ_k is the fixed effect for the k^{th} treatment, ν_{ik} is the random effect for the k^{th} treatment in the i^{th} trial, and $(\nu_{i1}, \ldots, \nu_{iK})$ is a vector of random effects whose covariance matrix is Σ_k. A possible factorization of Σ_K is $\Sigma_K = \text{diag}(\sigma_1, \ldots, \sigma_K) \times \Omega_K \times \text{diag}(\sigma_1, \ldots, \sigma_K)$, where Ω_K is a positive definite correlation matrix with off-diagonal elements $\rho_{k_1 k_2}$, σ_k is the standard deviation for random effect ν_{ik}, and $\text{diag}(\sigma_1, \ldots, \sigma_K)$ is a diagonal matrix with elements σ_i. In (3.1), the trial-level heterogeneity in response to treatment k is captured by σ_k, and the within-trial dependence among treatments is captured by Ω_K. An advantage of our AB method over the CB method is that the population-averaged event rate can then be calculated, for example with the probit link, in a closed form as $\pi_k = E(p_{ik}|\mu_k, \sigma_k) = \int_{-\infty}^{\infty} \Phi(\mu_k + \sigma_k z)\phi(z)dz = \Phi(\mu_k/\sqrt{1+\sigma_k^2}), k = 1, \ldots, K$, where $\Phi()$ is the standard normal cumulative density function and $\phi()$ is the standard normal density function. As such, marginal measures may be calculated as $\text{RR}_{kl} = \pi_k/\pi_l$, $\text{RD}_{kl} = \pi_k - \pi_l$, and $\text{OR}_{kl} = \frac{\pi_k/(1-\pi_k)}{\pi_l/(1-\pi_l)}$. We can also use an other link function, for example *logit* link, under which condition, the population-averaged event rate is $\pi_k \approx expit(\mu_k/\sqrt{1+C^2\sigma_k^2})$, where $C = 16\sqrt{3}/(15\pi)$ and $expit(x) = e^x/(1+e^x)$ [82].

For the MOM, let R_{ik} be an element of the matrix $R = (R_{ik})$, taking values $R_{ik} = 0$ if y_{ik} is observed and $R_{ik} = 1$ if y_{ik} is missing. If $f(R|Y_{\text{obs}}, Y_{\text{mis}}, \alpha) = f(R|\alpha)$, the mechanism is MCAR; if $f(R|Y_{\text{obs}}, Y_{\text{mis}}, \alpha) = f(R|Y_{\text{obs}}, \alpha)$, it is MAR; if there is no simplification for the conditional distribution $f(R|Y_{\text{obs}}, Y_{\text{mis}}, \alpha)$, it is MNAR. We assume R_{ik} has a Bernoulli distribution with a probability of missingness p_{ik}^{mis}, i.e., $R_{ik} \sim Ber(p_{ik}^{\text{mis}})$, where p_{ik}^{mis} may well depends on y_{ik}. A common way to realize the above is through linking p_{ik}^{mis} with the estimated p_{ik} (which is an approximation of $\frac{y_{ik}}{n_{ik}}$)

instead of y_{ik}, i.e., let $g(p_{ik}^{\text{mis}}) = f(p_{ik})$, where $g(p_{ik}^{\text{mis}})$ and $f(p_{ik})$ are some prespecified functions of p_{ik}^{mis} and p_{ik}.

Finally the joint distribution of Y_{obs} and R can be derived by integrating out the Y_{mis} as follows

$$P(Y_{\text{obs}}, R) = \Pi_{i=1}^{I} \left\{ \int \left\{ \Pi_{k=1}^{K} (p_{ik})^{y_{ik}} (1-p_{ik})^{n_{ik}-y_{ik}} \times \Pi_{k=1}^{K} (p_{ik}^{\text{mis}})^{R_{ik}} (1-p_{ik}^{\text{mis}})^{1-R_{ik}} \right\} \mathrm{d}Y_{\text{mis}} \right\}. \tag{3.2}$$

3.1.1 MOIs incorporating heterogeneity

Heterogeneity refers to the between-trial variation, i.e., multiple studies of the same research question may have different underlying values of the effect measure being estimated. Specifically, heterogeneity may be said to be present for treatment k if $p_{ik} \neq p_{jk}$ for some pair of trials i and j. A common solution to heterogeneity is through a random-effects model, which assumes that the underlying effects in trials of the same treatment come from a common distribution, usually normal.

In this paper we use the probit link for all models of interest and talk about various simplifications of (3.1). The simplest model incorporating heterogeneity can be specified as $\Phi^{-1}(p_{ik}) = \mu_k + \nu_i$ (MOI i), where $\nu_i \sim N(0, \sigma^2)$, corresponding to a covariance matrix $\Sigma_K = \sigma^2 I_K$. MOI i can be expanded to allow heterogeneous variances σ_k^2 instead of σ^2 as $\Phi^{-1}(p_{ik}) = \mu_k + \nu_{ik}$ with $\nu_{ik} \sim N(0, \sigma_k^2)$. We call this MOI ii. This special structure allows each treatment group to have its own heterogeneity parameter σ_k and a random treatment effect of $(\sigma_k - \sigma_l)\eta_i$ comparing treatment k versus l ($k \neq l$) in the i^{th} trial where $\eta_i \overset{\text{iid}}{\sim} N(0,1)$. Another way to expand model MOI i is to assume $\Phi^{-1}(p_{ik}) = \mu_k + \nu_{ik}$, where $(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})$ has an exchangeable correlation matrix with parameter ρ and ν_{ik} has the same variance σ^2. We call this MOI iii. An even more general model with exchangeable correlations is $\Phi^{-1}(p_{ik}) = \mu_k + \nu_{ik}$ where $(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})$ has an exchangeable correlation matrix with parameter ρ but different variances σ_k^2 for different treatment k. We call this MOI iv. The most general model has an unstructured covariance matrix Σ_K, which is an arbitrary $K \times K$ positive definite covariance matrix; this is model MOI v. This model assumes that the variances of different treatments are different, and correlations between different pairs of treatments are also different.

In general, MOI i is simple but may not be practical in most cases, and there may not be enough information contained in the data to accurately estimate all the parameters in MOI v when the number of treatments K is large and the number of studies I is small. Thus the exchangeable correlation models MOI iii and MOI iv appear to offer sensible yet practical alternatives.

3.1.2 MOM specification

Turning to the problem of specifying the model of missingness (MOM), we use the formula $\text{logit}(p_{ik}^{\text{mis}}) = \alpha_{0k} + \alpha_{1k} \times \text{logit}(p_{ik})$ (denoted MOM i) to link the probability of missingness p_{ik}^{mis} to the estimated p_{ik}, where α_{0k} is an unknown scalar parameter, and α_{1k} determines the missing mechanism, i.e., nonignorable missingness if $\alpha_{1k} \neq 0$ and ignorable missingness if $\alpha_{1k} = 0$. In this model, the probabilities of missingness for different treatments have different missingness parameters α_{1k}. A simpler model can be specified as $\text{logit}(p_{ik}^{\text{mis}}) = \alpha_{0k} + \alpha_1 \times \text{logit}(p_{ik})$ (denoted MOM ii), where all treatments share the same missingness parameter α_1.

3.1.3 Prior distributions, computation, and model selection

Since improper prior distributions may lead to improper posteriors in some complex models [69][70][91][92], we select minimally informative but proper priors. Specifically, we chose a very weakly informative prior $N(0, 1000)$ for μ_k, and a moderately informative $Gamma(1, 1)$ prior for the precisions $\tau = 1/\sigma^2$ in MOI i and $\tau_i = 1/\sigma_i^2$ in MOI ii, corresponding to a 95% Bayesian CI for variance parameters ranging from 0.27 to 39.5. In MOI iii and MOI iv, σ and σ_k have a uniform prior $U(0, 5)$ and ρ has a uniform prior $U(0.0001, 1)$. A vague Wishart prior is set for the precision matrix in the unstructured model MOI v, i.e., $\Sigma_K^{-1} \sim W(V, n)$, where $n = K$ is the degrees of freedom and V is a known $K \times K$ matrix with diagonal elements equal to 1.0 and off-diagonal elements equal to 0.005 [8]. It turns out that the above prior corresponds to a 95% Bayesian CI for the variance parameters ranging from 0.45 to 32.1, and a 95% Bayesian CI for the correlation parameters ranging from -1.00 to 1.00 (i.e., fully noninformative). Finally we specify a $logistic(0, 1)$ prior for α_{0k} and a weakly informative $N(0, 0.68)$ prior for α_1 and α_{1k} [93][94], which correspond to an approximately flat prior on the scale of p_{ik}^{mis}.

All models were implemented via Markov chain Monte Carlo (MCMC) methods using the WinBUGS software. We employed a burn-in of 1,000,000 iterations and followed by 1,000,000 iterations to calculate posterior estimates of the parameters of interest. The convergence of MCMC chains is assessed by the Gelman-Rubin convergence statistic and a visual inspection of the chains.

The Deviance Information Criterion (DIC) [95] was used as the model selection criterion. The deviance, up to an additive quantity not depending upon θ, is $D(\theta) = -2\log L(\theta; \text{Data})$, where $L(\theta; \text{Data})$ is the likelihood for the respective model. The DIC is given by $\overline{D(\theta)} + p_D$, where $\overline{D(\theta)} = E_{\theta|\text{Data}}[D(\theta)]$ is the Bayesian deviance, and $p_D = \overline{D(\theta)} - D(\overline{\theta})$ is the effective number of model parameters. It rewards better fitting models through the first term and penalizes more complex models through the second term. A model with smaller overall DIC value is preferred. WinBUGS provides DIC estimates for the MOIs and MOMs separately, and we call them *DIC for model of Interest* (DIC_I) and *DIC for model of Missingness* (DIC_M) respectively. Y_{mis} are treated as extra parameters in the MOMs, with the MOI acting as their prior distribution. Mason et al. [90] have suggested the use of DIC_M to compare the fit of MOMs with the same MOI.

3.2 *Smoking cessation* data application

We apply our Section 2 method to the a smoking cessation data set [96][97] shown in Table 3.1. This data set comprises 24 trials (22 two-arm and 2 three-arm trials) and 18,822 participants trying to quit smoking using one of the 4 treatments: (A) no contact, (B) self-help, (C) individual counseling, and (D) group counseling.

Table 3.2 shows the posterior medians of population-averaged event rates (π_A, π_B, π_C, and π_D) and their 95% posterior credible intervals for various models. MOI iii, MOI iv, and MOI v, which assume random effects $(\nu_{i1}, \nu_{i2}, \ldots, \nu_{iK})$ in the same trial have a multivariate normal distribution, show smaller DIC_I than MOI i and MOI ii, which assume the random effects to be independent. This suggests treatments in the same trial are correlated. While DIC_I of MOI iii, MOI iv, and MOI v are quite similar to each other, we select MOI iii as the best model of interest because it is the simplest model among these three and is easiest to implement.

Now let us take a look at the results of a sensitivity analysis. We call the joint modeling of MOI iii and MOM i "JM i" and the joint modeling of MOI iii and MOM ii "JM ii". Table 3.2 shows that the estimates for the population-averaged event rates π_A and π_C from JM i and JM ii are exactly the same as MOI iii, while those for π_B and π_D are slightly different from MOI iii. This phenomenon suggests that estimates for treatments A (No Contact) and C (Individual Counseling) are more robust to the nonignorable missingness, while estimates for B (Self Help) and D (Group Counseling) are a little more sensitive. However, the differences are quite small, even for the 95% credible intervals. Note that DIC_M for "JM i" is smaller than DIC_M for "JM ii", thus MOM i is more suitable for these data than MOM ii. As such, we adopt "JM i" for all further analysis.

Since the values of α_{1k} control the degree of departure from MAR missingness, we envisage further sensitivity analyses in which the changes in the estimated parameters of interest are studied for different values of the missingness parameters α_{1k}. In other words, we carry out a sensitivity analysis in which a series of models are run with a set of fixed values for the α_{1k}. More specifically, we use 15 values uniformly distributed between -1 and 1, namely -1.00, -0.86, -0.71, -0.57, -0.43, -0.29, -0.14, 0.00, 0.14, 0.29, 0.43, 0.57, 0.71, 0.86, and 1.00 for α_{1k}. Figure 3.1 presents the posterior medians and their 95% credible intervals for the population-averaged response rates versus different values of α_{1k}. Note that posterior medians of π_A and π_C versus α_{1k} in the left part of Figure 3.1 are horizontal lines, while posterior medians of π_B and π_D versus α_{1k} in the right part of Figure 3.1 have slight slopes. Thus the conclusion regarding treatments B and D is that they are slightly dependent on the missingness parameter α_{1k}, but treatments A and C are more robust to change here, similar to the conclusion from Table 3.2. Note that in the smoking cessation data, the numbers of trials that contain treatments A (19) and C (19) are larger than B (6) and D (6), which probably explains at least part the reason why treatments B and D are more sensitive to the missingness mechanism choice.

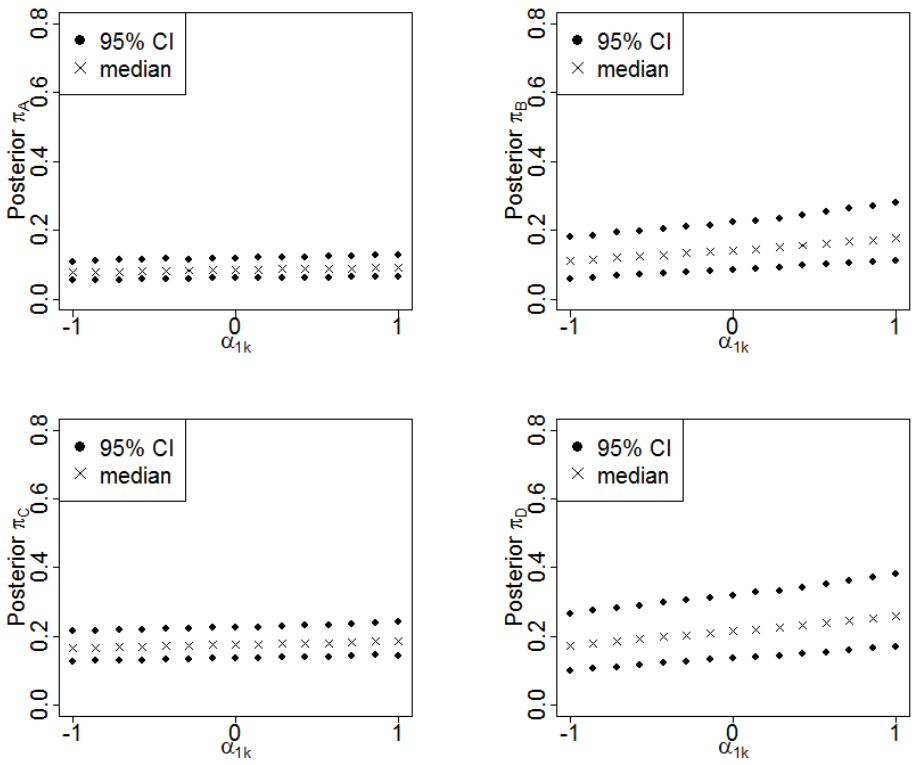

Figure 3.1: Population event rate variation with changes in α_{1k}

3.3 Simulations

3.3.1 Simulation setups

We compare the performance of the existing CB method (see details in the next paragraph) and our AB method under MCAR and MAR mechanisms in Simulation 1, and investigate the influence of nonignorable missingness to our AB method under MNAR mechanism in Simulation 2. We simulated 1000 replicate data sets. For each replicate, an NMA comprising 30 trials and 3 treatments (1, 2, and 3) was generated with 100 patients assigned to each arm in each trial for convenience. The bias of parameter estimates, which is the difference between true value and the mean of 1000 posterior median estimates, and the empirical mean squared error (MSE) were calculated as measures of performance.

The most popular CB method [17][80] uses the following Bayesian hierarchical model,

$$y_{ik} \overset{\text{ind}}{\sim} \text{Bin}(n_{ik}, p_{ik}),\ i = 1, \ldots, I,\ k \in S_i,$$
$$\text{logit}(p_{ik}) = \mu_i + X_{ik}\delta_{ibk}, \delta_{ibk} \overset{\text{ind}}{\sim} N(d_{bk}, \sigma^2_{bk}),$$

where μ_i is the specified baseline effect that is commonly regarded as a nuisance parameter; X_{ik} is the indicator for baseline, taking value 0 when $k = b$ and 1 when $k \neq b$; $b(i)$ is the specified baseline treatment in trial i, commonly denoted as b for simplicity as above; and δ_{ibk} represents the contrast between treatment k and b for the i^{th} trial and is assumed to be a random effect with a normal distribution with mean d_{bk} and variance σ^2_{bk}. This method assumes that $d_{hk} = d_{bk} - d_{bh}$ and $\text{Corr}(\delta_{ibh}, \delta_{ibk}) = \gamma^{(b)}_{hk}$ for $h \neq k \neq b$.

Simulation 1 The unstructured heterogeneous-variance model MOI v is used to generate the complete data set which contains 30 trials and 3 arms. We let the mean parameters have values $\mu_1 = -1.0$, $\mu_2 = -0.5$, and $\mu_3 = -0.8$, standard deviation parameters have values $\sigma_1 = 0.3$, $\sigma_2 = 0.4$, and $\sigma_3 = 0.5$, and correlation coefficients have values $\rho_{12} = 0.4$, $\rho_{13} = 0.5$, and $\rho_{23} = 0.6$. The response rate p_{ik} for the i^{th} trial and k^{th} treatment can be calculated with the above parameters according to formula $p_{ik} = \Phi(\mu_k + \nu_{ik})$, where $(\nu_{i1}, \ldots, \nu_{iK})^T \sim \text{MVN}(0, \Sigma_k)$ and Σ_k is the covariance matrix determined by $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23})$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$. Then the number of binary responses $\{y_{ik}\}$ are randomly generated by a binomial distribution with $n = 100$ and probabilities p_{ik}. Note that the above parameters correspond to the true population-averaged treatment-specific response rates $\pi_1 = 0.17$, $\pi_2 = 0.32$, and $\pi_3 = 0.24$, and thus the true odds ratios are $\text{OR}_{21} = 2.33$, $\text{OR}_{31} = 1.53$, and $\text{OR}_{32} = 0.66$.

R_{ik} represents the missingness, taking value 1 when the record for the k^{th} treatment in the i^{th} trial is missing and 0 when the record is present, and p^{mis}_{ik} is the corresponding probability of missingness. We consider simulated scenarios of missingness mimicking the characteristics of the real smoking cessation data, where most trials (22 trials) are two-arm trials and only a few (2 trials) are three-arm trials. Arm 2 which records treatment 2 is assumed to be completely observed, while Arms 1 and 3 have missing values. We let $n_{\text{mis}} = 10$, 11, 12, 13, 14 trials be missing for Arm 3, and then another n_{mis} trials be missing for Arm 1 selected from the remaining $30 - n_{\text{mis}}$, ensuring

each trial contains at least 2 treatments (while $30 - 2n_{\text{mis}}$ have 3). For the MCAR situation, the missingness of Arm 3 and Arm 1 are determined by $\text{logit}(p_{i3}^{\text{mis}}) = 1$ and $\text{logit}(p_{i1}^{\text{mis}}) = -1$ respectively. For the MAR situation, the missingness of Arm 3 is determined by $\text{logit}(p_{i3}^{\text{mis}}) = 1 + \text{logit}(\frac{y_{i2}}{n})$, whereas the missingness of Arm 1 is determined by $\text{logit}(p_{i3}^{\text{mis}}) = -1 - \text{logit}(\frac{y_{i2}}{n})$. Note that in missing data simulations it is a common approach to generate complete data first and then randomly delete some, leaving the missing percentage unknown. We instead first decide how many trials will be deleted and then select these trials according to p_{i3}^{mis} and p_{i1}^{mis} until the pre-determined number of missing trials is satisfied. In this way, we are able to track the model performance under various missing percentages.

Simulation 2 We now assess the influence of nonignorable missingness (MNAR). For simplicity, we generate data according to MOI iii, which is the best model of interest for the smoking cessation data set. We let the true mean parameters be $\mu_1 = -1.4$, $\mu_2 = -1.0$, and $\mu_3 = -0.8$, standard deviance $\sigma = 0.4$, and correlation coefficient $\rho = 0.5$; p_{ik} and y_{ik} are then generated accordingly. The true odds ratios are $\text{OR}_{21} = 2.00$, $\text{OR}_{31} = 2.77$, and $\text{OR}_{32} = 1.38$ correspondingly. We assume Arm 1 and Arm 3 are observed and the probability of missingness for Arm 2 is dependent on the unobserved values themselves y_{i2} through formula $\text{logit}(p_{i2}^{\text{mis}}) = 1 + \text{logit}(\frac{y_{i2}}{n})$, which leads to nonignorable missingness.

3.3.2 Simulation results

Simulation 1 Figure 3.2 presents bias and MSE of ORs (OR_{21}, OR_{31}, and OR_{32}) obtained from the AB method and the CB method under both MCAR and MAR mechanisms. Bias from our proposed AB method is consistently smaller than 0.05 for all $n_{\text{mis}} = 10$, 11, 12, 13, or 14 under both mechanisms, where bias from the CB method sometimes is bigger than 0.05 (see the two plots in the left column of Figure 3.2). Under both MCAR and MAR mechanisms, MSE from our AB method (the right column of Figure 3.2) is consistently smaller than from that of CB method for different n_{mis} values. This suggests that our method is less biased than the CB method. Another phenomenon we observe in Figure 3.2 is that the difference between two methods is smaller when the number of missing trials is small, e.g. when $n_{\text{mis}} = 10$ and 11. However, when n_{mis} becomes bigger, e.g. 12, 13, and 14, which mimics the real smoking cessation data, our

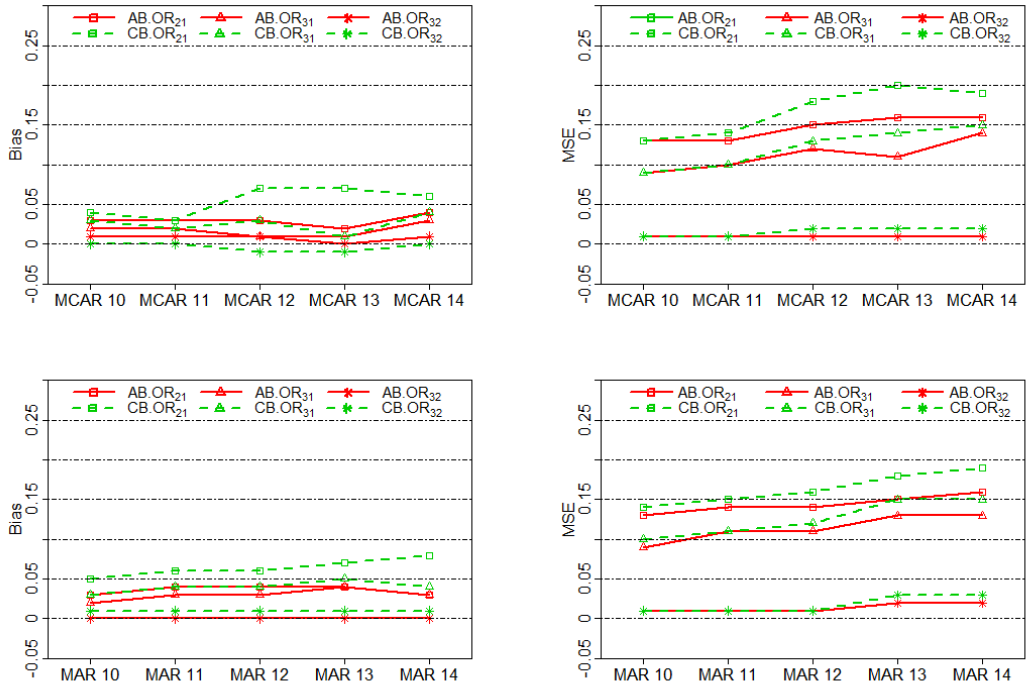

Figure 3.2: Bias and MSE under MCAR and MAR mechanisms

AB method is more robust to the missing data. In general, our proposed AB method outperforms the existing CB method in terms of bias and MSE under both MCAR and MAR assumptions.

Simulation 2 Our investigation of the influence of nonignorable missingness is based on fitting 5 models (M_{MAR}, M_{MNAR1}, M_{MNAR2}, M_{MNAR3}, and M_{MNAR4}) to the simulated dataset which contains nonignorable missingness. M_{MAR} is set to be exactly MOI iii, which ignores the nonignorable missingness. M_{MNAR1} uses MOI iii as the model of interest and sets the parameter of missingness equal to 0, i.e., $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + 0 * \text{logit}(p_{i2})$, thus it is actually equivalent to the M_{MAR} model. For M_{MNAR2}, both parts of the model are correctly specified, i.e., MOI iii for model of interest, and $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + \alpha_1 * \text{logit}(p_{i2})$ for model of missingness (Note that the estimated posterior p_{i2} is an approximation for $\frac{y_{i2}}{n}$). M_{MNAR3} and M_{MNAR4} have overly complex forms for missingness, i.e., $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + \alpha_1 * \text{logit}(p_{i2}) + \alpha_2 * p_{i2}^2$ and $\text{logit}(p_{i2}^{\text{mis}}) =$

$\alpha_0 + \alpha_1 * \text{logit}(p_{i2}) + \alpha_2 * \text{logit}(p_{i1})$ respectively, jointly modeled with MOI iii. In general, M_{MNAR2}, M_{MNAR3}, and M_{MNAR4} take nonignorable missingness into account, while M_{MAR} and M_{MNAR1} disregard nonignorable missingness.

Table 3.3 provides evidence that ignoring nonignorable missingness will lead to bias. M_{MAR} and M_{MNAR1} produce both larger relative bias and MSE, for the estimation of OR_{21} and OR_{32}, compared with the true M_{MNAR2} and overly complex M_{MNAR3} and M_{MNAR4}. Relative bias is defined as bias divided by the true value. For example, the relative bias for OR_{21} is -0.11 from both M_{MAR} and M_{MNAR1}, while it is only -0.04 from M_{MNAR2} and M_{MNAR3}, and -0.08 from M_{MNAR4}. Now let us take the OR_{32} as an example, the MSE is 0.11 from M_{MAR} and M_{MNAR1}, in contrast to 0.07 from M_{MNAR2} and M_{MNAR3} and 0.10 from M_{MNAR4}. In a nutshell, joint models M_{MNAR2}, M_{MNAR3}, and M_{MNAR4} incorporating nonignorable missingness do outperform M_{MNAR1} and M_{MAR}, though misspecification of the missingness may slightly affect the model performance as M_{MNAR4}.

Another phenomenon we observe in Table 3.3 is that the relative biases and MSEs for OR_{31} from all models are the same, which is because Arm 1 and Arm 3 are fully observed. Therefore though our proposed method aims to tackle missing data in NMA, especially nonignorable missingness, it is robust even if there is no missingness. Note that in all joint models the relative biases for OR_{31} are all around 0, while relative biases for OR_{21} and OR_{32} are slightly bigger than 0, partly due to the noninformative priors for the missingness parameters.

In summary, when nonignorable missingness is present, our selection model method provides an effective solution.

3.4 Discussion

Although clinical and policy-making interest often lies in comparing active agents, new drugs are often compared with placebo or standard treatments in order to obtain approval for drug licensing [98]. Thus it is unrealistic to expect that comparisons of all treatments of interest will be provided from any single trial as the licensed and reference treatments can differ across countries and over time. Network meta-analysis, if properly applied, can serve decision-making as a better tool than pairwise meta-analysis [99] and

bring tremendous changes to the practice of evidence-based medicine.

Selective choices of treatments and trials may lead to MNAR. Neither the current CB method, assuming MCAR, nor the AB method [8], assuming MAR, can handle this thorny situation. We extended the AB method in Chapter 2 to incorporate nonignorable missingness using sensitivity analysis, and evaluated its performance through simulations. In addition, we have also shown with simulation studies that the AB method outperforms the current CB method in terms of both bias and MSE.

Though other methods exist for sensitivity analysis, selection models are intuitively appealing because they show how the probabilities of missingness depend directly on the data values and their factorization specifies $f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$ directly, which is the distribution that analysts are usually interested in. However, selection models rely heavily on the correct specification of the model form[59]. Alternatively, pattern-mixture models, which do not require correct specification of the precise model form (albeit with their own limitations), should be considered as well to handle nonignorable missingness.

Important limitations for all NMA methods include inconsistency and publication bias. Inconsistency, defined as apparent discrepancy between direct and indirect comparisons of two treatments, is one of the major issues in NMA. The extensive criticism of network meta-analysis is associated with the difficulty in evaluating the assumption underlying the statistical synthesis of direct and indirect evidence. Methods assessing inconsistency have their own drawbacks and some of them are cumbersome to apply. Publication bias, the concern that studies with significant results are more likely to be published and published studies are more likely to be included in meta-analysis, is another potential source of bias[28] in NMA. Approaches that test as well as account for inconsistency and publication bias await further exploration.

Although researchers recognize the benefits of NMA, they often use indirect evidence as a second choice, giving priority to direct evidence to inform decision making. Overall, many health care practitioners remain skeptical of this emerging statistical technique. The assumption underlying the models, the statistical expertise required to fit them, the issues of inconsistency and publication bias, and the lack of an interpretable and simple measure to summarize the results and to evaluate the risk of bias contribute to this skepticism. Future research should focus on interpretation and applicability in addition to more imaginative statistical modeling.

Table 3.1: Smoking Cessation Data (y_{ik}/n_{ik})

Baseline	Study	A	B	C	D
A	1	9/140		23/140	10/138
	2	79/702	77/694		
	3	18/671	21/535		
	4	8/116	19/146		
	5	75/731		363/714	
	6	2/106		9/205	
	7	58/549		237/1561	
	8	0/33		9/48	
	9	3/100		31/98	
	10	1/31		26/95	
	11	6/39		17/77	
	12	95/1107		134/1031	
	13	15/187		35/504	
	14	78/584		73/675	
	15	69/1177		54/888	
	16	64/642		107/761	
	17	5/62		8/90	
	18	20/234		34/237	
	19	0/20			9/20
B	20		20/49	16/43	
	21		11/78	12/85	29/170
	22		7/66		32/127
C	23			12/76	20/74
	24			9/55	3/26

Note: y_{ik} is the number of events and n_{ik} is the total number of subjects.

Table 3.2: Posterior summaries of population-averaged event rates for *smoking cessation* data

	MOI i	MOI ii	MOI iii	MOI iv	MOI v	JM i	JM ii
π_A	0.11	0.10	0.09	0.08	0.08	0.09	0.09
	(0.08,0.15)	(0.08,0.13)	(0.06,0.12)	(0.06,0.11)	(0.06,0.10)	(0.06,0.13)	(0.06,0.12)
π_B	0.13	0.17	0.14	0.16	0.15	0.13	0.15
	(0.09,0.18)	(0.11,0.27)	(0.09,0.22)	(0.09,0.33)	(0.09,0.24)	(0.07,0.22)	(0.09,0.24)
π_C	0.18	0.25	0.18	0.18	0.17	0.18	0.18
	(0.14,0.24)	(0.16,0.39)	(0.14,0.23)	(0.13,0.24)	(0.13,0.23)	(0.14,0.24)	(0.14,0.23)
π_D	0.20	0.31	0.21	0.23	0.21	0.20	0.22
	(0.14,0.27)	(0.18,0.50)	(0.14,0.32)	(0.13,0.45)	(0.13,0.33)	(0.11,0.33)	(0.14,0.34)
DIC_M	—	—	—	—	—	***104.1***	109.2
DIC_I	545.50	374.9	***323.1***	325.5	324.4	323.2	323.1

Posterior medians and their 95% credible intervals.

The bold and italic cell shows the smaller DIC.

Table 3.3: Performance of joint modeling when MNAR is present

		OR_{21}	OR_{31}	OR_{32}
ReBias	M_{MAR}	-0.11	0.01	0.16
	M_{MNAR1}	-0.11	0.01	0.16
	M_{MNAR2}	-0.04	0.01	0.08
	M_{MNAR3}	-0.04	0.01	0.08
	M_{MNAR4}	-0.08	0.01	0.11
MSE	M_{MAR}	0.14	0.18	0.11
	M_{MNAR1}	0.14	0.18	0.11
	M_{MNAR2}	0.13	0.18	0.07
	M_{MNAR3}	0.13	0.18	0.07
	M_{MNAR4}	0.13	0.18	0.10

M_{MAR}: MOI iii & NA

M_{MNAR1}: MOI iii & $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + 0 * \text{logit}(p_{i2})$

M_{MNAR2}: MOI iii & $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + \alpha_1 * \text{logit}(p_{i2})$

M_{MNAR3}: MOI iii & $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + \alpha_1 * \text{logit}(p_{i2}) + \alpha_2 * p_{i2}^2$

M_{MNAR4}: MOI iii & $\text{logit}(p_{i2}^{\text{mis}}) = \alpha_0 + \alpha_1 * \text{logit}(p_{i2}) + \alpha_2 * \text{logit}(p_{i1})$

$\text{ReBias} = \frac{\text{Bias}}{\text{True Value}}$

Chapter 4

The effects of excluding trials from network meta-analyses

Mills et al. [58] showed empirically that excluding treatments in NMA sometimes can have important effects on treatment effect estimates. In this chapter, we instead obtain empirical evidence on whether selective inclusion of trials can impact treatment effect estimates in an NMA setting, using both the AB and CB methods. Section 4.1 describes the source of data as well as their extraction, and the reanalysis strategy. Section 4.2 presents the analyzed networks, shows the impact of removing trials, and compares the performance of the AB and CB methods in terms of impact. We conclude this chapter with discussions of limitations and future work in Section 4.3.

4.1 Materials and Methods

Data source and data extraction

The data sets come from a paper by Veroniki et al. [1] evaluating inconsistency in NMAs. The authors searched in PubMed for research articles published between March 1997 and February 2011 including networks of at least four treatments, one closed loop and dichotomous primary outcomes. After some screening process they ended up with 40 networks. Then they extracted data regarding the year of publication, the methods applied for indirect comparison, the number of studies, and the number of arms the

studies included, as well as the total number of interventions involved in each network. The extracted trial data include the name of each trial, the number of events, the sample size and the treatment in every arm.

We gained the data of these 40 networks from the authors of this paper [1] and now present them in Table 4.1. n_t represents the total number of treatments of interest in each NMA. Due to identifiability issue, we require that each treatment has been compared in at least 3 trials in each network, which will be explained in the next section. n_{t_3} records the number of treatments that are compared in less than 3 trials. $n_{t_3} = 0$ indicates all treatments included in the NMA have been compared in at least 3 trials; while $n_{t_3} \neq 0$ suggests some treatments appear less than 3 times in this NMA. We delete all NMAs whose $n_{t_3} \neq 0$, and finally end up with 14 networks showed in bold cells in Table 4.1.

Statistical analysis

For each of the 14 NMAs, we firstly analyzed the complete available data with both the AB and the CB methods and recorded the treatment effect estimates for all pairwise comparisons. Then we performed reanalysis excluding one trial every time and estimating corresponding estimates with the remaining trials. To evaluate the impact of exclusion of trials on estimation, we calculated the *absolute relative changes* (ARC) of the estimates. For example, if an estimated OR is 0.90 in the full network and 0.75 in the network with one trial excluded, then the ARC is $|(0.75-0.90)/0.90| = 0.17$. We applied both AB and CB methods to investigate and compare relative changes regardless of the analyses chosen in the original publications.

We now present the two categories of model parameterizations. We first briefly show the AB method proposed in Chapter 2 Section 2.2.2, but in a simpler format, as $\Phi^{-1}(p_{ik}) = \mu_k + \sigma_k \nu_{ik}$, where $\nu_{ik} \sim N(0,1)$. Then the population-averaged event rate for treatment k has a closed form as $\pi_k = E(p_{ik}|\mu_k, \sigma_k) = \int_{-\infty}^{\infty} \Phi(\mu_k + \sigma_k z)\phi(z)dz = \Phi(\mu_k/\sqrt{1+\sigma_k^2}), k = 1, \ldots, K$. The marginal ORs are then defined as $\text{OR}_{kl} = [\pi_k/(1-\pi_k)]/[\pi_l/(1-\pi_l)]$ for a pairwise comparison between treatments k and l ($k \neq l$). In this model, μ_k and σ_k are the two parameters of interest for each treatment k. Thus in order to be identifiable, at least 2 trials are required for each treatment. In other words, we need to make sure each treatment in an NMA has been compared in at least 2 trials.

And since we delete a trial in each reanalysis, we select NMAs whose treatments have been compared in at least 3 trials as is shown in Table 4.1. Though the AB method focuses on estimating event rate for each treatment arm, we use OR as the reporting scale in this paper in order to be consistent with the CB method.

We then present the CB method proposed by Lu & Ades [17] on the following hierarchical structure, which is in a more general form than the method in Section 2.2.1:

$$\begin{aligned}
&\text{logit}(p_{ik}) = \mu_k + X_{ik}\delta_{ib(i)k}, \\
&\boldsymbol{\delta}_{ib(i)k,k\in S_i}|\boldsymbol{d}, V \sim MVN(\boldsymbol{d}, V_{|S_i|-1}),
\end{aligned}$$

where vector $\boldsymbol{\delta}_{ib(i)k,k\in S_i}$ has a multivariate normal distribution with mean vector $\boldsymbol{d} = (d_{b(i)k})$ and covariance matrix $V_{|S_i|-1}$ if the ith trial is a multi-arm comparison or an univariate normal distribution if the ith trial is a two-arm comparison. S_i is the cardinality of trial i. A very common $V_{|S_i|-1}$ is a homogeneous-variance exchangeable matrix with correlation 1/2, i.e., $\delta_{ib(i)k} \sim N(d_{b(i)k}, \sigma^2)$ and $cov(d_{b(i)k}, (d_{b(i)h}) = 1/2\sigma^2$. The model also assumes exchangeability, i.e., $d_{hk} = d_{bh} - d_{bk}$. Finally $\text{OR}_{hk} = e^{d_{hk}}$.

Analyses were conducted using WinBUGS version 1.4 (Medical Research Council Biostatistics Unit, Cambridge) and R version 3.0.2 (www.r-project.org/).

4.2 Results

Analyzed networks

The 14 networks with individual trial data involving 567 randomized controlled trials with 389361 patients are shown in more details in Table 4.2. It presents the names of the 14 NMAs, the total number of studies and nodes (i.e., treatments of interest) in each NMA and the outcome investigated in each evidence synthesis. Networks ranged in size from 9 trials on 4 treatments to 111 trials on 12 treatments. Note that here 9 trials on 4 nodes does not conflict with the ≥ 3 criteria since some trials are multi-arm where multiple treatments (> 2) are compared. Thus the condition that each treatment is compared in at least 3 trials is still satisfied.

Changes after removal of trials

$\binom{K}{2} = \frac{K(K-1)}{2}$ ORs were estimated in an NMA involving K treatments, thus $\frac{K(K-1)}{2}$ ARCs were calculated with the estimates obtained before and after removal of a trial. We record the maximum and mean of the ARCs after removing of each trial. Suppose we have I trials in an NMA, then I maximums and I means are recorded.

Figure 4.1 presents these maximums and means. The left plot shows the maximum ARCs obtained from the CB method (y-axis) versus these obtained from the AB method (x-axis). Dots of different colors are from different networks. Maximum ARCs are mostly within 0.2, but up to 0.64 for the CB method and 0.53 for the AB method. The red fitted regression line is slightly above $y = x$, which suggests removal of trials causes slightly larger maximum ARCs from the CB method than the AB method. The right plot is for mean ARCs. Most points are within 0.1, while they are up to 0.31 for the CB method and 0.26 for the AB method. Though the red regression line is still slightly above the $y = x$ line, difference between these two lines are very small. It seems the CB method is well-matched with the AB method using mean ARC as the standard measure. In short, exclusion of trials can sometimes have large impact on results.

Comparison of AB versus CB method

We investigated further the difference in robustness of the two methods in terms of impact when deleting a trial. A paired t-test comparing the maximum ARCs from the AB method with those from the CB method produced a p-value 0.56, showing that there were no statistically significant differences between the two methods. Similar analysis of mean absolute relative changes with p-value 0.69 drew the same conclusion.

Figure 4.2 is a Bland-Altman plot [109][110][111], which is often used to analyze the agreement between two different methods. In this plot, instead of using relative changes, we used differences of ORs in the log scale. The log OR differences obtained from the AB and CB methods are denoted as $\log\frac{\mathrm{OR}_{AB}}{\mathrm{OR}_{AB}^{(i)}}$ and $\log\frac{\mathrm{OR}_{CB}}{\mathrm{OR}_{CB}^{(i)}}$, where i represents the deleted trial. In Figure 4.2, the x-axis shows the mean of the log OR changes obtained from the two methods $\frac{1}{2}(\log\frac{\mathrm{OR}_{AB}}{\mathrm{OR}_{AB}^{(i)}} + \log\frac{\mathrm{OR}_{CB}}{\mathrm{OR}_{CB}^{(i)}})$, while the y-axis shows the difference of the log OR changes from the two methods $\log\frac{\mathrm{OR}_{AB}}{\mathrm{OR}_{AB}^{(i)}} - \log\frac{\mathrm{OR}_{CB}}{\mathrm{OR}_{CB}^{(i)}}$, which are defined as bias in Bland-Altman plot. The dashed gray line Figure 4.2 shows that the mean

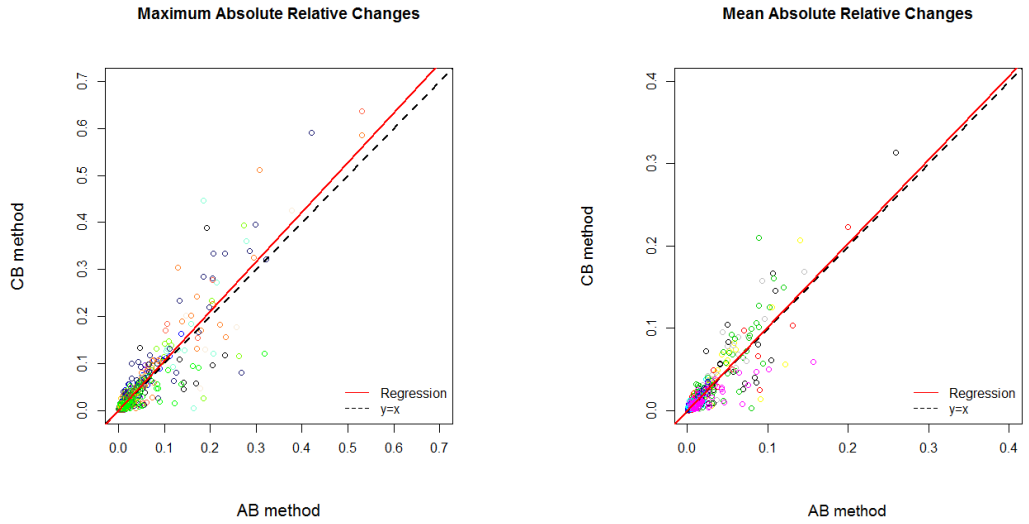

Figure 4.1: Scatter plot for maxmium and mean absolute relative changes in ORs comparing AB method with CB method. Different colors represent different networks. The red lines are the regression lines, and the black dash lines are the identical lines $y = x$.

bias is very close to 0, indicating the average discrepancy between methods is not large enough to be important. The solid gray lines representing the limits of agreements are relatively narrow, suggesting the two methods are essentially equivalent. In addition, the scatter points of difference do not show any particular pattern, suggesting that there is no systematic difference between the two methods.

4.3 Discussion

It is common for network meta-analysis to exclude trials and specific treatment arms based on widely diverse criteria [58]. The empirical impact in exclusion of treatments arms was investigated in Mills et al. [58], while the impact in exclusion of trials has not been previously discussed. In this chapter, we documented that exclusion of trials can have significant impact on results. Although exclusion of a particular trial does not affect the treatment effect estimates much in traditional meta-analyses [112][113], our

results do find that some exclusions can affect substantially the estimated treatment effects. We also find the performance of the AB and CB methods are similar.

Of course, caveats to our analysis exist. Firstly, it is not practical for us to retrieve the trials that had been excluded up front and evaluate the impact of their inclusion. Instead, the potential impact that further exclusions may bring was investigated in this paper. Secondly, although we suggest inclusion of all trials possible, we acknowledge that this is daunting in practice due to unavailability and high expense. Thirdly, sometimes certain exclusions are clearly justifiable; e.g., inclusion of poorly designed trials may instead lead to bias, which is broadly related to the outlying trials.

Turning to future work, we are interested in developing methods for handling publication bias, the concern that studies with significant results are more likely to be published, and published studies (especially those in the meta-analysts own language) are more likely to be included in an NMA. There have been various methods developed for publication bias in traditional meta-analyses, but few previous papers have considered methods for handling publication bias in NMA framework. Our future interest also lies in quantifying characteristics of the trials whose removal would bring more significant impact on the estimation. This research has a potential to serve as guidance for future network meta-analysis.

Table 4.1: 40 network meta-analyses from Veroniki et al. [1]

Network number	$\frac{n_{t_3}}{n_t}$	Network number	$\frac{n_{t_3}}{n_t}$
1	5/9	21	1/4
2	**0/5**	22	**0/4**
3	2/8	23	**0/4**
4	**0/4**	24	3/11
5	1/8	25	**0/8**
6	2/8	26	2/5
7	3/7	27	2/7
8	1/6	28	**0/5**
9	**0/4**	29	8/11
10	**0/12**	30	6/10
11	2/9	31	1/5
12	**0/5**	32	1/8
13	**0/6**	33	5/8
14	1/5	34	**0/5**
15	1/4	35	**0/4**
16	4/13	36	2/5
17	1/5	37	2/9
18	4/9	38	9/16
19	**0/4**	39	4/10
20	**0/6**	40	1/6

Note: n_t represents the number of treatments in each NMA, while n_{t_3} represents the number of treatments that have been compared in less than 3 trials. Bold cells have $n_{t_3} = 0$.

Table 4.2: 14 network meta-analyses we analyzed

Author name	# studies	# nodes	Condition/outcome
Ara 2009 [100]	11	5	Hypercholesterolaemia/Discontinuation due to adverse event
Ballesteros 2005 [101]	9	4	Dysthymia
Bucher 1997 [102]	18	4	Pneumocystis carinii
Cipriani 2009 [34]	111	12	Unipolar major depression in adults/ response to treatment
Eisenberg 2008 [103]	61	5	Smoking
Elliott 2007 [36]	22	6	Hypertension, high risk patients/patients who developed diabetes
Lu.1 2009 [80]	24	4	Smoking cessation
Lu.2 2007 [79]	40	6	Gastroesophageal reflux disease
Middleton 2010 [104]	20	4	Heavy menstrual bleeding
Mills 2009 [105]	89	4	Short-term smoking abstinence
Picard 2000 [106]	43	8	Pain on injection
Puhan 2009 [107]	34	5	COPD
Thijs 2008 [108]	23	5	Serious vascular events after transient ischaemic attack (TIA) or stroke
Trikalinos 2009 [46]	62	4	Non-acute coronary artery disease

Note: The author names of the 14 NMAs, the number of studies and nodes in each network and the outcomes.

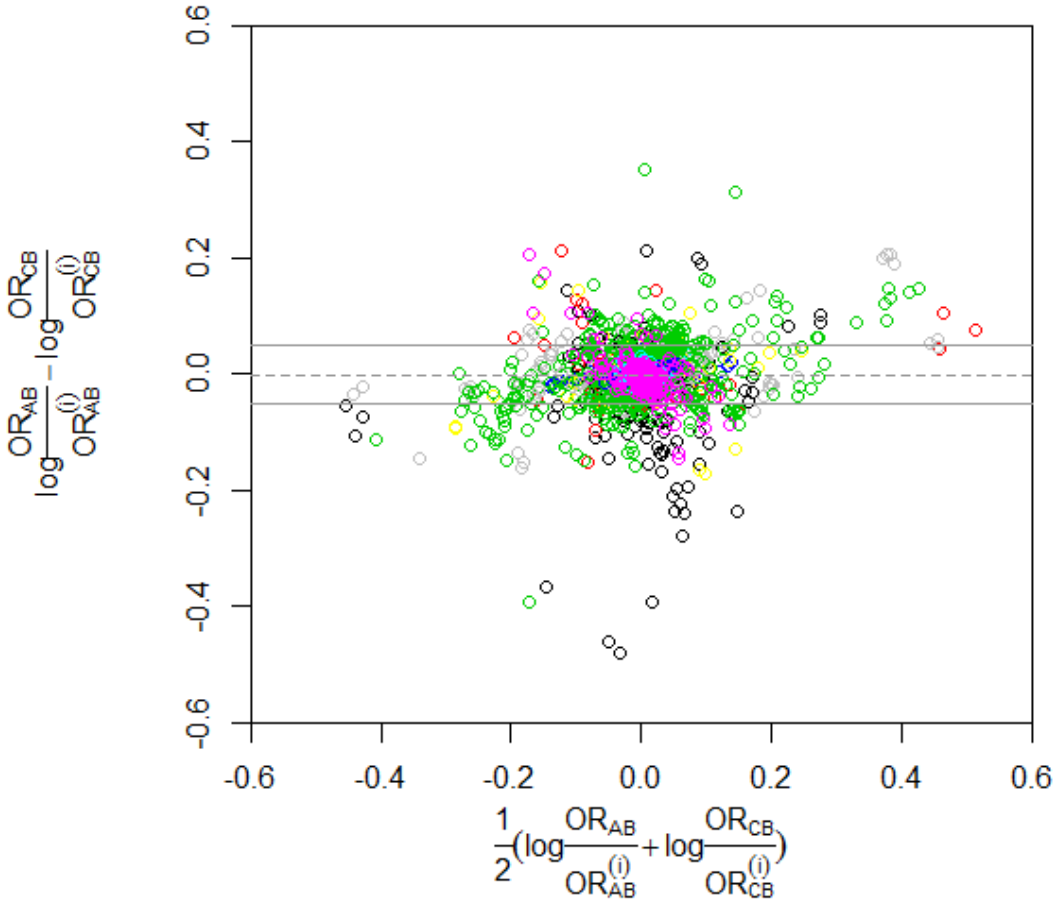

Figure 4.2: Bland-Altman Plot. The difference between log OR changes obtained from AB method and CB method is drawn against the mean of the log OR changes obtained from the two methods. Dash line represent the mean of bias, and the solid lines show the limits of agreement.

Chapter 5

Detecting outlying trials in network meta-analysis

Some trials may appear to deviate markedly from the others, and thus be inappropriate to be synthesized in an NMA. In addition, the inclusion of these trials in evidence synthesis may lead to bias in estimation. We call such trials *trial-level outliers*. In this chapter, we introduce two NMA model frameworks for Gaussian data (contrast-based and arm-based) in Section 5.2 and propose four Bayesian outlier detection measures in Section 5.3, which are then applied to a diabetes data set introduced in Section 5.1. The corresponding data analysis results are shown in Section 5.4, and the performance of the method is evaluated through simulation studies in Section 5.5. We end this chapter with a discussion and further model extension suggestions in Section 5.6.

5.1 Illustrative diabetes data

The diabetes network meta-analysis shown in Table 5.1 comprises efficacy responses over 12 internal industry-sponsored trials of 5 potential diabetes treatments (1: PIO (pioglitazone), 2: Placebo, 3: MET (metformin), 4: SU (sulfonylurea), and 5: ROSI (rosiglitazone)). The major efficacy outcome is the mean change in HbA1C (denoted by "mean" in Table 5.1), which is a lab measurement indicating the average level of blood sugar (glucose) over the previous 3 months, and is thought of as a measure of how well a patient is controlling his or her diabetes. The columns "n" and "sd" in Table 5.1

Table 5.1: Diabetes dataset

	1 (PIO)			2 (Placebo)			3 (MET)			4 (SU)			5 (ROSI)		
Trials	n	mean	sd	n	mean	sd	n	mean	sd	n	mean	sd	n	mean	sd
1	103	-0.76	0.97	115	-0.16	0.92									
2	248	-0.91	1.53	56	0.65	1.27									
3	73	-1.02	1.46	13	-1.10	1.63									
4	131	-1.03	1.57	65	0.66	1.38									
5	285	-1.25	1.15	138	-0.38	1.05									
6	379	-0.17	1.31	193	-1.06	1.67									
7	124	-0.26	0.58	441	-0.41	0.54							145	-0.35	0.49
8	533	-1.59	1.15				539	-1.79	1.13						
9	551	-1.66	1.01							541	-1.84	1.12			
10	283	-1.28	0.97							275	-1.4	0.98			
11	51	-1.07	1.36							56	-1	1.03			
12	41	-1.11	1.32							38	-0.78	1.21			

Note: n denotes sample size; mean denotes sample mean; sd denotes standard sample deviation.

represent sample size and sample standard deviation, respectively.

Figure 5.1 is an undirected graph elucidating the network of comparative relations for the 5 drugs in our NMA. Every node has different size, indicating the total sample size randomized to each treatment, and every edge has different thickness, indicating the frequency of each comparison (i.e., the number of studies including this pair of treatments). Seven trials (Trials 1-7) include comparisons of PIO versus Placebo, 4 trials (Trials 9-12) include comparisons of PIO versus SU, 1 trial (Trial 8) compares MET and PIO, and 1 trial (Trial 7) compares PIO, ROSI, and Placebo.

5.2 Statistical models for NMA of continuous data

In this section, we present two NMA model parameterizations (AB method and CB method) for Gaussian data.

Arm-based method: Chapter 2 proposed an arm-based NMA model for binary

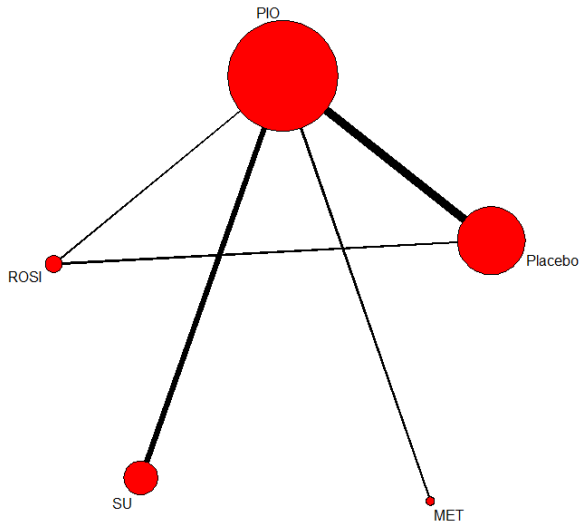

Figure 5.1: Graphical representation for the network of the diabetes dataset. The size of each node is proportional to the sample size randomized in each treatment, and the thickness of the link is proportional to the numbers of trials investigating the relation

data. Here we modify this model to adapt to the Gaussian diabetes data as follows:

$$\begin{aligned} y_{ik} &\sim N\left(\Delta_{ik}, \frac{\sigma^2}{n_{ik}}\right) \\ Sd_{ik}^2 &\sim Gamma\left(\frac{n_{ik}-1}{2}, \frac{n_{ik}-1}{2\sigma^2}\right) \\ \text{and } \Delta_{ik} &= \mu_k + \gamma\nu_{ik}, \end{aligned} \tag{5.1}$$

where the observations y_{ik} (most often thought of as group means $\overline{y}_{ik} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} y_{ikj}$) represent the mean response change of HbA1C over n_{ik} patients assigned to the k^{th} treatment in the i^{th} trial. This response is assumed to have a normal distribution with mean Δ_{ik} and variance $\frac{\sigma^2}{n_{ik}}$, where Δ_{ik} and σ^2 are the true population mean and variance. Sd_{ik}^2 is the group-level sample variance, which then has a $Gamma$ distribution with scale parameter $\frac{n_{ik}-1}{2}$ and rate parameter $\frac{n_{ik}-1}{2\sigma^2}$. Finally, μ_k is the treatment-specific fixed effect and γ is the standard deviation of Δ_{ik}, implemented via a random effects ν_{ik} independently distributed as standard normal. Since γ is the same across k, (5.1) is a

homogeneous-variance model. If we instead use γ_k, we obtain a heterogeneous-variance model.

Note that in (5.1), $Sd_{ik}^2 = \frac{1}{n_{ik}-1}\sum_{j=1}^{n_{ik}}(y_{ikj} - y_{ik})^2$ is the sample variance of n_{ik} observations, where j represents the subject. Thus by Basu's Theorem, y_{ik} and Sd_{ik}^2 are statistically independent, and $\frac{(n_{ik}-1)Sd_{ik}^2}{\sigma^2} \sim \chi^2(n_{ik}-1) \equiv Gamma(\frac{n_{ik}-1}{2}, \frac{1}{2})$. This in turn implies that $Sd_{ik}^2 \sim Gamma(\frac{n_{ik}-1}{2}, \frac{n_{ik}-1}{2\sigma^2})$.

Contrast-based method: Following Spiegelhalter et al. [114] and Ding and Fu [115], a contrast-based network meta-analysis model for Gaussian data can be written as $\Delta_{ik} = \Delta_i + X_{ik}\delta_{ibk}$ with $\delta_{ibk} \overset{\text{ind}}{\sim} N(d_{bk}, \epsilon^2)$. Here, X_{ik} is an indicator taking value 0 when $k = b$ and value 1 when $k \neq b$, Δ_i is the baseline mean response for the i^{th} trial, and δ_{ibk} measures the effect of treatment k relative to the baseline treatment b, which is permitted to change across studies. Note that when $X_{ik} = 0$, $\Delta_{ik} = \Delta_i$ represents the response in the baseline group in the i^{th} trial. d_{bk} is the mean contrast effect of treatment k versus b, and ϵ^2 is the variance. Here y_{ik} and Sd_{ik}^2 have the same distribution as in (5.1), i.e., $y_{ik} \sim N(\Delta_{ik}, \frac{\sigma^2}{n_{ik}})$ and $Sd_{ik}^2 \sim Gamma(\frac{n_{ik}-1}{2}, \frac{n_{ik}-1}{2\sigma^2})$. Again, this model is a homogeneous-variance model, while $\delta_{ibk} \sim N(d_{bk}, \epsilon_{bk}^2)$ corresponds to a heterogeneous-variance model.

The comparison between the arm-based method and the contrast-based method has been previously discussed [8][116]. In this paper, we will only focus on the arm-based method, though our methods could apply equally well to the contrast-based setting (and might well identify different outlying studies in that case).

5.3 Outlier detection measures

5.3.1 Relative distance

We define cross-validatory (or "leave one out") *relative distance* statistics, RD_i^k, to measure the effect of deleting trial i on our NMA estimate for a particular treatment k as

$$\text{RD}_i^k = \left| \frac{\hat{\eta}_k - \hat{\eta}_{k(i)}}{\hat{\eta}_k} \right|, \tag{5.2}$$

where $\hat{\eta}_k$ is some estimate of interest (e.g., the posterior mean treatment effect) for treatment k from the full data, and $\hat{\eta}_{k(i)}$ is the estimate for treatment k from the data

where trial i has been omitted. The bigger RD_i^k is, the higher the relative effect of deleting trial i is, and thus the higher the likelihood that trial i is influential and may be a "trial-level outlier" in this sense. We can also define an *average relative distance* (ARD) to measure the *average* effect of deleting trial i as:

$$\mathrm{ARD}_i = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{\hat{\eta}_k - \hat{\eta}_{k(i)}}{\hat{\eta}_k}\right|, \tag{5.3}$$

where $\hat{\eta}_k$ and $\hat{\eta}_{k(i)}$ have the same representations as in (5.2). The bigger the ARD_i is, the greater the average effect of deleting trial i is. We may define trial i as an outlier if RD_i^k or ARD_i is large relative to the full collection of values. Formal "probabilities of being an outlier" could also be computed; say, $P(\mathrm{RD}_i^k > \mathrm{T}|\mathrm{data})$ for some preselected threshold T, say $\mathrm{T} = 0.1$.

5.3.2 Standardized trial residuals

Again, based on cross-validatory thinking, we might calculate the fitted value for y_{ik} by conditioning on all data except $\mathbf{y}_i$, namely, $\mathbf{y}_{(i)} = (\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, \cdots, \mathbf{y}_I)'$, where $\mathbf{y}_i = \{y_{i,k}, k \in S_i\}$ and S_i represents the set of treatments compared in trial i. We then compute the difference between the observed and fitted values for y_{ik} and standardize it as follows:

$$\mathrm{STR}_i^k = \frac{y_{ik} - E(y_{ik}|\mathbf{y}_{(i)})}{\sqrt{Var(y_{ik}|\mathbf{y}_{(i)})}}, \tag{5.4}$$

where the STR_i^k stands for the Bayesian *standardized trial residual* for the k^{th} treatment in the i^{th} trial. The *average absolute standardized trial residual* (ASTR) can be defined correspondingly as:

$$\mathrm{ASTR}_i = \frac{1}{n_{S_i}}\sum_{k\in S_i}\left|\frac{y_{ik} - E(y_{ik}|\mathbf{y}_{(i)})}{\sqrt{Var(y_{ik}|\mathbf{y}_{(i)})}}\right|, \tag{5.5}$$

where S_i has the same representation as before, and n_{S_i} represents its cardinality. Large STR_i^k and ASTR_i, say larger than 1.5, suggest observation y_{ik} and trial i may be outliers respectively. Note that in formulas (5.4) and (5.5), we compute the posterior mean and variance with respect to the *conditional predictive distribution*,

$$f(y_{ik}|\boldsymbol{y}_{(i)}) = \frac{f(\boldsymbol{y})}{f(\boldsymbol{y}_{(i)})} = \int f(y_{ik}|\boldsymbol{\theta}, \boldsymbol{y}_{(i)})p(\boldsymbol{\theta}|\boldsymbol{y}_{(i)})\,\mathrm{d}\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is the entire parameter collection. In other words, $f(y_{ik}|\boldsymbol{y}_{(i)})$ is the posterior predictive density of y_{ik} given the remainder of the data except that concerning trial i.

5.3.3 Bayesian p-value

An alternative to cross-validatory approaches as described in Sections 5.3.1 and 5.3.2 is the use of *posterior predictive* model checks, an approach initially promoted by Rubin [117] and popularized by Gelman et al. [118]. The key idea is to construct some "discrepancy measures" that capture departures of the observed data from the assumed model (likelihood and prior distribution). Note that though such measures must be functions of observed data alone in the classical frequentist framework, Bayesian model checking based on posterior predictive distributions allows more general measures that depend on *both* data and parameters. Gelman et al. (1996) suggest an omnibus goodness of fit discrepancy measure $D_i^k(y_{ik}, \boldsymbol{\theta})$ that depends on the parameters $\boldsymbol{\theta}$ and the data y_{ik},

$$D_i^k = \frac{[y_{ik} - E(Y_{ik}|\boldsymbol{\theta})]^2}{Var(Y_{ik}|\boldsymbol{\theta})}.$$

We subsequently define an *average discrepancy* measure $AD_i(\boldsymbol{y}_i, \boldsymbol{\theta})$ as

$$AD_i = \sum_{k \in S_i} \frac{[y_{ik} - E(Y_{ik}|\boldsymbol{\theta})]^2}{Var(Y_{ik}|\boldsymbol{\theta})},$$

where of course $\boldsymbol{\theta}$ varies according to its posterior distribution, and S_i is the set of treatments that are compared in trial i. We now can compare the distribution of $D_i^k(y_{ik}, \boldsymbol{\theta})$ and $AD_i(\boldsymbol{y}_i, \boldsymbol{\theta})$ for the observed data y_{ik} and $\boldsymbol{y}_i$ with that of $D_i^k(y_{ik}^*, \boldsymbol{\theta})$ and $AD_i(\boldsymbol{y}_i^*, \boldsymbol{\theta})$ for hypothetical future values y_{ik}^* and $\boldsymbol{y}_i^*$. Note that y_{ik}^* and $\boldsymbol{y}_i^*$ are defined as another "copy" of the observed data point y_{ik} and vector $\boldsymbol{y}_i$, which are not observed but instead generated from their posterior predictive distributions as part of the MCMC sampling order [119]. D_i^k and AD_i^k computed using the observed data that are extreme relative to this reference distribution indicate poor model fit and merit closer examination in the analysis.

A convenient summary measure of the extremeness of the $D_i^k(y_{ik}^*, \boldsymbol{\theta})$ with respect to the $D_i^k(y_{ik}, \boldsymbol{\theta})$ is the posterior predictive tail area, defined as the *Bayesian p-value* for

discrepancy,

$$\begin{aligned} p_{D_i^k} &\equiv P[D_i^k(y_{ik}^*, \boldsymbol{\theta}) > D_i^k(y_{ik}, \boldsymbol{\theta})|\boldsymbol{y}] \\ &= \int P[D_i^k(y_{ik}^*, \boldsymbol{\theta}) > D_i^k(y_{ik}, \boldsymbol{\theta})|\boldsymbol{\theta}]p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}. \end{aligned} \tag{5.6}$$

Similarly, the Bayesian p-value for average discrepancy is defined as

$$\begin{aligned} p_{AD_i} &\equiv P[AD_i(\boldsymbol{y}_i^*, \boldsymbol{\theta}) > AD_i(\boldsymbol{y}_i, \boldsymbol{\theta})|\boldsymbol{y}] \\ &= \int P[AD_i(\boldsymbol{y}_i^*, \boldsymbol{\theta}) > AD_i(\boldsymbol{y}_i, \boldsymbol{\theta})|\boldsymbol{\theta}]p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}. \end{aligned} \tag{5.7}$$

Note that $p_{D_i^k}$ and p_{AD_i} should not be used to compare models. Rather, they serve only as measures of discrepancy between the proposed model and the observed data, and therefore provide information concerning overall model adequacy and outlier detection. Other summaries focused on other measures of poor fit (say, in the tail of the distribution) can also be defined; see Gelman, Meng, and Stern [120].

5.3.4 Scale mixtures of normals

The conditioning feature of MCMC computational methods enables another approach related to models employing *scale mixtures of normals* (SMN) (see [119], p.184) to investigate outlyingness. Here we expand model (5.1) to

$$y_{ik} \sim N(\Delta_{ik}, \lambda_i \frac{\sigma^2}{n_{ik}}), \tag{5.8}$$

where the λ_i are unknown scale parameters. We then specify prior distributions for λ_i, for example,

$$\begin{aligned} &\text{SMN}_1 : \lambda_i \equiv 1 \Rightarrow \text{Normal errors} \\ &\text{SMN}_2 : \lambda_i \sim IG\left(\frac{\upsilon}{2}, \frac{2}{\upsilon}\right) \Rightarrow \text{ Students' } t_\upsilon \text{ errors} \\ \text{and } &\text{SMN}_3 : \lambda_i \sim Expo(2) \Rightarrow \text{Double exponential errors,} \end{aligned}$$

where the distributions behind the arrow symbols identify the possible departures from normality for the error terms. Since extreme observations will correspond to extreme fitted values of these scale parameters λ_i, potential outliers can be identified by examining the λ_i posterior distributions. Doubt is cast on the commensurability of trial i with the rest if the posterior mean (or median) of λ_i is much bigger than 1, or if $P(\lambda_i \geq 1|\boldsymbol{y})$ is larger than some threshold value, say 0.95.

5.4 Application to diabetes data

5.4.1 Outlier detection results with various measures

Relative distance

We first fit model (5.1) to our diabetes NMA data and record the posterior estimates. Then we fit model (5.1) 12 more times with the 1^{st} through 12^{th} trials omitted, respectively, and record all necessary posterior estimates. Finally, we calculated RD_i^k according to (5.2) with $\hat{\mu}_k$ obtained from the full data and $\hat{\mu}_{k(i)}$ obtained from the data with the i^{th} trial deleted. Note that here we let $\hat{\eta}_k = \hat{\mu}_k$, the mean treatment effect, when we calculate RD_i^k. ARD_i is calculated similarly.

Figure 5.2 shows the relative distances separately for the 5 treatments (PIO, Placebo, MET, SU, and ROSI). The vertical axes show relative distances ranging from 0.0 to 1.0, and the horizontal axes index trials that are deleted in the calculation of the relative distances. Thinking of 0.2 as a significance threshold, Trial 6 is mildly influential for PIO; Trials 2, 4, 6 are influential for Placebo; Trial 8 is influential for MET; Trials 6 and 9 are mildly influential for SU; and Trials 2, 4, 6, and 7 are influential for ROSI. Note that Trials 2, 4, and 6 are influential for ROSI even though they do not directly compare this treatment. In short, with 0.2 as the cutoff for RD_i^k, Trials 2, 4, 6, 7, and 8 seem to be potential outliers.

Figure 5.3 shows the average relative distances. ARD_i for $i = 6$, 7, and 8 are above threshold T = 0.1, while those for the others are below this level (though some only narrowly). Thus Trials 6, 7, and 8 are more influential with ARD_i as the evaluation criteria. This is roughly consistent with the results in Figure 5.2.

We further investigate the possibility that Trials 2, 4, 6, 7, and 8 are outliers. First, it seems fair to call Trials 2 and 4 outliers since the mean change values of HbA1C for patients treated with Placebo are all negative except in these two trials. Second, for Trials 6 and 7, the mean change values of HbA1C for patients treated with PIO from both trials are much smaller than those from the other 10 trials. Thus Trials 6 and 7 are extreme in this sense. Third, the mean HbA1C change responses for patients in Trial 8 do not seem abnormal; however, we observe that MET is only contained in Trial 8, so that deleting Trial 8 will of course have a big impact on the estimates. We thus can

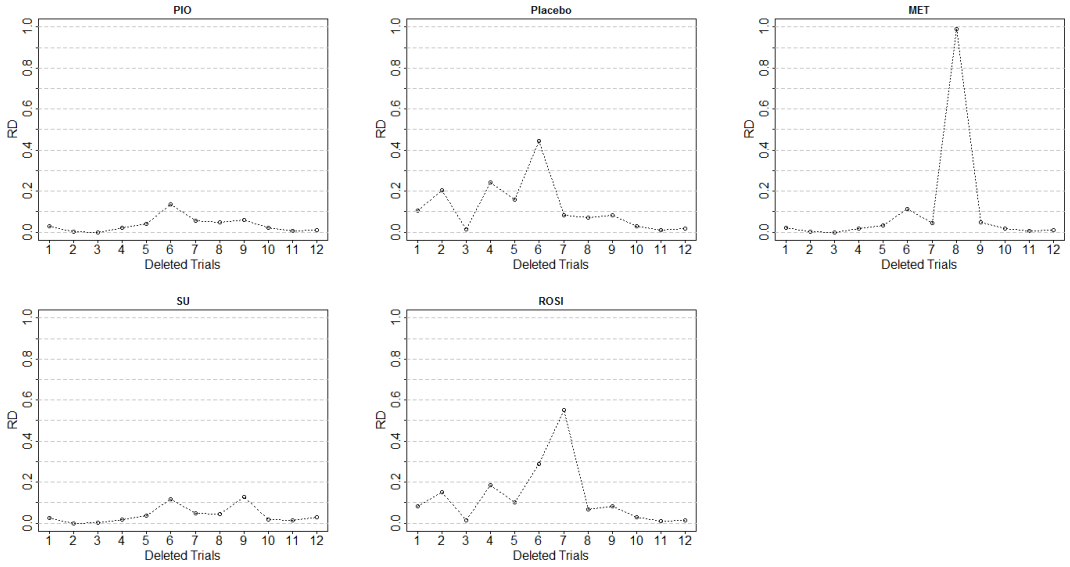

Figure 5.2: Relative distances versus deleted trials for each treatment

infer that RD_8^3 and ARD_8 for Trial 8 are big largely due to lack of information, rather than true "outlyingness".

Standardized trial residuals

Table 5.2 shows that the Bayesian standardized trial residuals for Trials 2, 4, and 6 are larger than 1.5 in absolute value (more specifically, $STR_2^2 = -2.24$, $STR_4^2 = -2.29$, and $STR_6^1 = -1.58$). In Trials 2 and 4, the mean changes of HbA1C for patients taking placebo are positive, while those from the other trials that contain placebo ($k = 2$) are negative. Thus it seems reasonable to call Trial 2 and Trial 4 outliers. For STR_6^1, the Trial 6 Bayesian standardized residual for PIO ($k = 1$), we see that the mean change of HbA1C for patients taking PIO in this trial is -0.17, much smaller than that from the other trials. Thus Trial 6 would appear to highly underestimate the efficacy of PIO, i.e., it may also be a legitimate outlier. However, ASTR in Table 5.2 does not identify any significant trial-level outliers, with no values larger than 1.5 and previously unidentified Trial 9 emerging along with Trials 2, 4, and 6 with ASTR > 1.

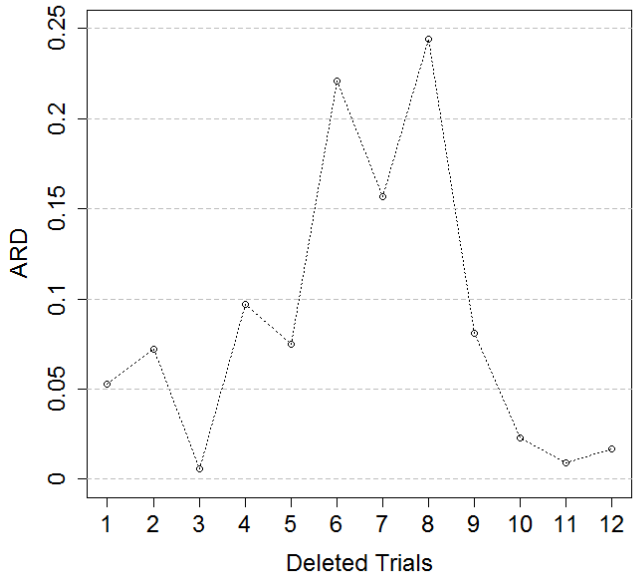

Figure 5.3: Average relative distances versus deleted trials

Bayesian p-values

Table 5.3 shows that Trials 2, 4, 5, 6, and 7 have at least one Bayesian p-value smaller than 0.05. In the case of Trials 2 and 4, this is likely due to the positive responses for placebo in these trials. By contrast, Trials 6 and 7 are likely flagged because the PIO mean changes in HbAIC values in these two trials are much smaller than that from the other trials, i.e., the presence of Trials 6 and 7 in the NMA would underestimate the efficacy of PIO. Oddly, the Bayesian p-values for Trial 5 are also smaller than 0.05, an apparent significance that may be inflated by small variances and merits further investigation. At any rate, it suggests the omnibus goodness of fit measure adopted in Section 4.3 may not be optimal in this particular setting.

Scale mixtures of normals

Figure 5.4 shows that Trials 2, 4, 6, and 7 once again emerge as outliers under both models SMN2 and SMN3, since the posterior estimates for the scale parameters λ_i are

Table 5.2: Results for Bayesian standard trial residuals

Trial	Treatment	STR	ASTR	Trial	Treatment	STR	ASTR
1	1	-0.35	0.64	7	1	-1.20	0.51
	2	-0.92			2	-0.33	
—	—	—	—		5	0.00	
2	1	-0.06	1.15	8	1	1.13	0.60
	2	**-2.24**			3	0.06	
3	1	0.11	0.40	9	1	1.28	1.38
	2	0.69			4	1.48	
4	1	0.09	1.19	10	1	0.55	0.54
	2	**-2.29**			4	0.53	
5	1	0.52	0.56	11	1	0.17	0.16
	2	-0.60			4	-0.16	
6	1	**-1.58**	1.38	12	1	0.23	0.38
	2	1.18			4	-0.53	

Note: Bold cells have $|\text{STR}| > 1.5$.

significantly larger than 1 (with their log-scale 95% CIs significantly higher than 0). The specific estimated values of λ_i are listed in Table 5.4. In addition, Table 5.4 shows that the probabilities that the scale parameters λ_i are larger than 1 in Trials 2, 4, 6, and 7, are all 0.99 or greater, which further suggests the outlyingness of these trials, in broad agreement with the results of the previous detection approaches.

5.4.2 Results with and without outliers

We compare the estimated values for parameters of interest from full data with those computed without our identified trial-level outliers (Trials 2, 4, 6, and 7). A value for μ_5 is not computable without outliers since Trial 7, an outlier, is the only source of information on Treatment 5 (ROSI). As shown in Table 5.5, μ_k estimates are quite different before and after deleting the outliers. For example, in the case of μ_2, the relative difference is $\frac{-0.67-(-0.47)}{-0.47} = 42.6\%$. Relative changes for the other μ's are similarly meaningful, though are less impressive for the variance parameters γ and

Table 5.3: Bayesian p-values for discrepancy

Bayesian p-values									
$p_{D_1^1}$	0.25	$p_{D_3^2}$	0.37	$p_{D_6^1}$	**0.00**	$p_{D_8^1}$	0.50	$p_{D_{10}^4}$	0.57
$p_{D_1^2}$	0.23	$p_{D_4^1}$	**0.00**	$p_{D_6^2}$	**0.00**	$p_{D_8^3}$	0.50	$p_{D_{11}^1}$	0.50
$p_{D_2^1}$	**0.02**	$p_{D_4^2}$	**0.00**	$p_{D_7^1}$	**0.01**	$p_{D_9^1}$	0.50	$p_{D_{11}^4}$	0.50
$p_{D_2^2}$	**0.00**	$p_{D_5^1}$	**0.03**	$p_{D_7^2}$	0.20	$p_{D_9^4}$	0.48	$p_{D_{12}^1}$	0.31
$p_{D_3^1}$	0.51	$p_{D_5^2}$	**0.00**	$p_{D_7^5}$	0.50	$p_{D_{10}^1}$	0.56	$p_{D_{12}^4}$	0.27

Note: Bold cells have p-value < 0.05.

σ^2. In a nutshell, when trial-level outliers exist in an NMA, they can wield significant influence on estimates of the parameters of interest.

5.5 Simulations

In this section we evaluate the performance of our proposed detection measures using simulations, and demonstrate the advantages of different measures in different situations.

5.5.1 Simulation settings

We generate a continuous-data network meta-analysis with 12 trials to compare 4 treatments. We set the parameters of interest according to the results of our diabetes data analysis without outliers in Table 5.5 as follows: $\mu_1 = -1.19$, $\mu_2 = -0.47$, $\mu_3 = -1.39$, $\mu_4 = -1.32$, and $\sigma^2 = 1.20$. In order to limit the variability of the random effects, we let $\gamma = 0.10$. For simplicity we assign $n_{ik} = 100$ patients to each arm k in each trial i. Then artificial data y_{ik} can be generated according to model (5.1).

Unbalanced Design Mimicking Motivating Data: In order to make the simulated data as realistic as possible, we let Trials 1-7 compare only Arm 1 and Arm 2, Trial 8 compare only Arm 1 and Arm 3, and Trials 9-12 compare only Arm 1 and Arm 4, again mimicking the motivating diabetes data in Table 5.1.

Balanced Design: In this setting, we still let Arm 1 be compared in all 12 trials, but Arms 2-4 are assumed to exist in the same numbers of trials. Specifically, Trials

Table 5.4: Results for scale mixtures of normals

	λ_i		$P(\lambda_i > 1|\boldsymbol{y})$	
Trial	SMN_2	SMN_3	SMN_2	SMN_3
1	1.57(0.30,20.06)	1.03(0.16,15.68)	0.67	0.51
2	**10.22(1.89,97.33)**	**9.90(1.92,95.03)**	**0.99**	**0.99**
3	3.11(0.67,26.84)	2.71(0.56,24.46)	0.92	0.89
4	**12.71(2.52,114.50)**	**12.39(2.60,110.30)**	**1.00**	**1.00**
5	1.69(0.29,28.37)	1.08(0.15,22.77)	0.68	0.53
6	**100(24.66,788.60)**	**99.38(24.89,760.60)**	**1.00**	**1.00**
7	**24.82(4.27,222.20)**	**24.24(4.36,212.60)**	**1.00**	**0.99**
8	1.44(0.27,36.64)	0.73(0.14,19.00)	0.63	0.40
9	1.35(0.27,19.96)	0.78(0.15,12.65)	0.62	0.41
10	1.03(0.24,11.63)	0.54(0.12,6.37)	0.51	0.28
11	1.16(0.28,10.98)	0.76(0.18,7.54)	0.57	0.39
12	2.08(0.51,18.51)	1.70(0.41,15.74)	0.82	0.74

Note: λ_i denotes the scale parameter; and $P(\lambda_i > 1|\boldsymbol{y})$ denotes the probability that the scale parameter λ_i is larger than 1 given the data. Bold cells represent the outliers.

1-4 compare Arm 1 and Arm 2, Trials 5-8 compare Arm 1 and Arm 3, and Trials 9-12 compare Arm 1 and Arm 4; a simple "star network" in NMA parlance.

Finally, for both designs, we manipulate the data so that Trials 4 and 8 will be outliers: the fixed value 5 is added to Arm 2 in Trial 4 and Arm 3 in Trial 8. We summarize our proposed methods over 1000 simulated data sets.

5.5.2 Simulation results

For each detection measure, we compare its performance under the unbalanced and balanced designs.

Relative distance: We show RD_i^k for $i = 1, \ldots, 12$ and $k = 1, \ldots, 4$ in Table 5.6. For the unbalanced design, most of the relative distances are close to 0 and some are around 0.50. However, $RD_4^2 = 2.99$ and $RD_8^3 = 1.01$, suggesting that Trials 4 and 8

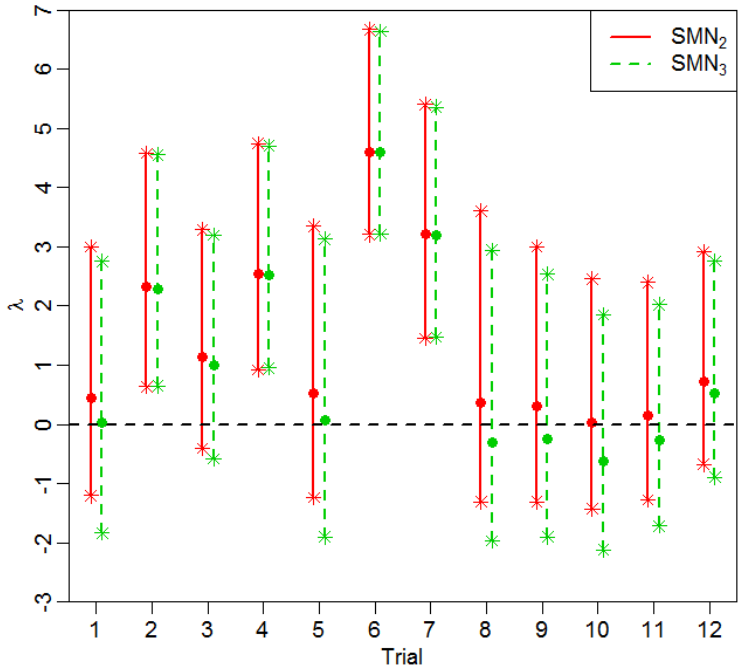

Figure 5.4: Posterior λ_i in log scale for SMN_2 and SMN_3

are influential for Arms 2 and 3, respectively, as we expected. In contrast to these straightforward findings, the results of the balanced design are shown in the right part of Table 5.6. $RD_4^2 = 1.60$ implies that Trial 4 is influential for Arm 2; the remaining RD_i^2 are roughly 1/3 this magnitude (0.53). Also, $RD_5^3 = RD_6^3 = 3.19$, $RD_7^3 = 3.20$, and $RD_8^3 = 9.58$ suggests that all trials containing Arm 3 are influential for it, with outlying Trial 8 as the most influential but the other trials again roughly 1/3 as large. In short, relative distances can successfully detect influential trials, but the relative sizes of these statistics can still vary over a surprising range, making a decision on cutoff value difficult. Here, the "inflation" in values for Arm 3 appears due to the fact that μ_3 is close to 0 (-0.47). Other simulations (not shown) reveal that RD_4^2 and RD_8^3 for the balanced design have roughly the same magnitude if μ_k is constant across k.

Standardized trial residuals: Table 5.7 provides standardized trial residuals for

Table 5.5: Posterior summaries for parameters of interest with and without outliers

	μ_1	μ_2	μ_3	μ_4	μ_5	γ	σ^2
With Outliers	-0.95	-0.67	-1.16	-1.08	-0.70	0.57	1.34
	(-1.28, -0.62)	(-1.01,-0.33)	(-1.51, -0.80)	(-1.42, -0.74)	(-1.09, -0.31)	(0.39, 0.86)	(1.29, 1.39)
Without Outliers	-1.19	-0.47	-1.39	-1.32	—	0.60	1.20
	(-1.65, -0.71)	(-0.95, 0.02)	(-1.87, -0.90)	(-1.78, -0.84)	—	(0.38, 1.01)	(1.15, 1.26)

Note that μ_5 is not available in Row 2 because the only trial (Trial 7) containing Treatment 5 is an outlier.

the 12 trials, displaying 24 values for each design. For the balanced design, STR successfully detects Arm 2 in Trial 4 and Arm 3 in Trial 8 to be outliers. ASTR also works as it should for the balanced design. However, for the unbalanced design, STR and ASTR only successfully detect Trial 4, but not Trial 8. Here with only a single trial to estimate this arm, there is no way for the procedure to identify the arm as "outlying", with the extra 5 units added to every simulated value simply absorbed into the mean structure.

Bayesian p-values: We now turn to the posterior predictive model checks (Bayesian p-values) with the omnibus goodness of fit discrepancy measure $D_i^k(y_{ik}, \boldsymbol{\theta})$ detailed in Section 5.3.3. Unfortunately, as shown in Table 5.8, we find Bayesian p-values perform poorly, with false positives common. After deeper exploration, we found the poor performance was due to the discrepancy measure $D_i^k(y_{ik}, \boldsymbol{\theta}) = \frac{[y_{ik} - E(Y_{ik}|\boldsymbol{\theta})]^2}{Var(Y_{ik}|\boldsymbol{\theta})}$. Hypothetical future values y_{ik}^* generated from their posterior predictive distributions are always very close to $E(Y_{ik}|\boldsymbol{\theta})$, which leads to very small values of $D_i^k(y_{ik}^*, \boldsymbol{\theta})$. Since the $E(Y_{ik}|\boldsymbol{\theta})$ are not always close to the observed y_{ik}, the $D_i^k(y_{ik}, \boldsymbol{\theta})$ are mostly larger than the $D_i^k(y_{ik}^*, \boldsymbol{\theta})$, which leads to small Bayesian p-values (i.e., $p_{D_i^k} < 0.05$). Thus as suggested earlier, more sensible discrepancy measures appear to be needed for this approach to be feasible, e.g., $D_i^k(y_{ik}, \boldsymbol{\theta}) = \frac{|y_{ik} - E(Y_{ik}|\boldsymbol{\theta})|}{\sqrt{Var(Y_{ik}|\boldsymbol{\theta})}}$.

Scale mixture of normals: For each trial i, Table 5.9 shows the average posterior medians of the scale parameters $\widehat{\lambda_i}$, average $P(\lambda_i > 1|\boldsymbol{y})$, and P(Positive), defined as the empirical proportion of the 1000 simulations where $P(\lambda_i > 1|\boldsymbol{y}) > 0.95$ (i.e., a "positive outcome"). Under both balanced and unbalanced designs, $\widehat{\lambda_4}$ is much larger than 1, the average $P(\lambda_4 > 1|\boldsymbol{y}) = 1.00 > 0.95$, and P(Positive) is always equal to 1 for Trial 4. For Trial 8, $\widehat{\lambda_8}$ is much larger than 1 under the balanced design ($\widehat{\lambda_8} = 572.50$),

Table 5.6: Relative distances for the unbalanced and balanced designs in the simulation study

	Unbalanced Design				Balanced Design			
Trial	Arm 1	Arm 2	Arm 3	Arm 4	Arm 1	Arm 2	Arm 3	Arm4
1	0.01	0.50	0	0.01	0.01	0.53	0.05	0
2	0.01	0.50	0	0.01	0.01	0.53	0.06	0.01
3	0.01	0.50	0	0.01	0.01	0.53	0.05	0.01
4	0.01	**2.99**	0	0.01	0.01	**1.60**	0.06	0.01
5	0.01	0.50	0	0.01	0.01	0.01	**3.19**	0.01
6	0.01	0.50	0	0.01	0.01	0.01	**3.19**	0.01
7	0.01	0.50	0	0.01	0.01	0.01	**3.20**	0.01
8	0.01	0.03	**1.01**	0.01	0.01	0.01	**9.58**	0.01
9	0.01	0.03	0	0.01	0.01	0.01	0.06	0.01
10	0.01	0.03	0	0.01	0.01	0.01	0.05	0.01
11	0.01	0.03	0	0.01	0.01	0.01	0.05	0.01
12	0.01	0.03	0	0.01	0.01	0.01	0.06	0.01

Note: RDs bigger than 1 are in bold.

while for the unbalanced design it is only a little larger than 1 ($\widehat{\lambda_8} = 1.33$). The average $P(\lambda_i > 1|\boldsymbol{y})$ is 1.00 for the balanced design, but only 0.61 for the unbalanced design; similarly P(Positive) is 1.00 for balanced Trial 8, whereas it is 0 for the unbalanced case. The reason is that there are no other sources of information about Arm 3 in this design except Trial 8. As in Table 5.7, the algorithm does not have any information that can be treated as reference to detect the outlyingness of Trial 8 for Arm 3. To sum up, our SMN measure can successfully detect trial-level outliers with very high accuracy under a balanced replicated design, but are likely to be less effective for unbalanced designs.

In summary, our investigation with simulated data suggests that our proposed measures RD, STR, and SMN can often uncover trial-level outliers and thus facilitate more accurate evidence synthesis. Bayesian p-values appear more dependent on a sensible choice of discrepancy measure, and are not competitive with the other methods without further exploration.

Table 5.7: Standardized trial residuals for unbalanced and balanced designs in the simulation study

Unbalanced Design								Balanced Design							
Trial	Arm	STR	ASTR	Trial	Arm	STR	ASTR	Trial	Arm	STR	ASTR	Trial	Arm	STR	ASTR
1	1	0.10	0.50	7	1	0.09	0.50	1	1	0.07	0.78	7	1	0.07	0.78
	2	0.90			2	0.91			2	1.48			3	1.49	
2	1	0.10	0.50	8	1	0.09	0.10	2	1	0.08	0.78	8	1	0.09	**2.83**
	2	0.91			3	0.11			3	1.49			3	**5.56**	
3	1	0.09	0.50	9	1	0.09	0.09	3	1	0.07	0.78	9	1	0.08	0.08
	2	0.91			4	0.09			2	1.49			4	0.08	
4	1	0.16	**4.55**	10	1	0.10	0.09	4	1	0.09	**2.84**	10	1	0.07	0.07
	2	**8.94**			4	0.09			2	**5.59**			4	0.07	
5	1	0.09	0.50	11	1	0.09	0.09	5	1	0.07	0.78	11	1	0.07	0.07
	2	0.90			4	0.09			3	1.48			4	0.07	
6	1	0.09	0.49	12	1	0.09	0.09	6	1	0.08	0.78	12	1	0.08	0.08
	2	0.90			4	0.09			3	1.49			4	0.08	

Note: Cells with STR and ASTR larger than 1.5 are in bold.

5.6 Discussion and future work

Though methods for network meta-analysis have been extensively discussed and explored in the current literature, few previous papers appear to have mentioned trial-level outliers. In this paper, we proposed four detection measures for trial-level outliers, and applied them to a diabetes network meta-analysis, as well as simulated their performance in balanced and unbalanced designs. Our results suggest RD (ARD), STR (ASTR), and SMN perform well and are promising tools for detecting trial-level outliers.

Our detection measures can be easily extended to binary data. For example, instead of (5.1) for Gaussian data, Zhang et al. [8] proposed an arm-based method for binary data wherein $\Phi^{-1}(p_{ik}) = \mu_k + \gamma\nu_{ik}$, where p_{ik} is the event rate for the k^{th} treatment in the i^{th} trial, and μ_k, γ, and ν_{ik} have the same representations as (5.1). In this binary data setting, RD (ARD) and STR (ASTR) can be defined similarly as in the continuous

Table 5.8: Mean Bayesian p-values of 1000 replicates of simulations for unbalanced and balanced designs

Unbalanced Design						Balanced Design					
$p_{D_1^1}$	0.03	$p_{D_9^1}$	0.58	$p_{D_5^2}$	0.02	$p_{D_1^1}$	0	$p_{D_9^1}$	0.58	$p_{D_5^2}$	0
$p_{D_2^1}$	0.03	$p_{D_{10}^1}$	0.58	$p_{D_6^2}$	0.02	$p_{D_2^1}$	0	$p_{D_{10}^1}$	0.58	$p_{D_6^2}$	0
$p_{D_3^1}$	0.03	$p_{D_{11}^1}$	0.58	$p_{D_7^2}$	0.02	$p_{D_3^1}$	0	$p_{D_{11}^1}$	0.58	$p_{D_7^2}$	0
$p_{D_4^1}$	0	$p_{D_{12}^1}$	0.58	$p_{D_8^3}$	0.50	$p_{D_4^1}$	0	$p_{D_{12}^1}$	0.58	$p_{D_8^3}$	0
$p_{D_5^1}$	0.03	$p_{D_1^2}$	0.02	$p_{D_9^4}$	0.57	$p_{D_5^1}$	0	$p_{D_1^2}$	0	$p_{D_9^4}$	0.57
$p_{D_6^1}$	0.03	$p_{D_2^2}$	0.02	$p_{D_{10}^4}$	0.57	$p_{D_6^1}$	0	$p_{D_2^2}$	0	$p_{D_{10}^4}$	0.57
$p_{D_7^1}$	0.03	$p_{D_3^2}$	0.02	$p_{D_{11}^4}$	0.58	$p_{D_7^1}$	0	$p_{D_3^2}$	0	$p_{D_{11}^4}$	0.57
$p_{D_8^1}$	0.50	$p_{D_4^2}$	0	$p_{D_{12}^4}$	0.57	$p_{D_8^1}$	0	$p_{D_4^2}$	0	$p_{D_{12}^4}$	0.57

data setting. The implementation of SMN for binary data relies on rewriting (5.1) as

$$\begin{aligned} & Y_{ik} \sim Bin(n_{ik}, p_{ik}), k \in S_i, i = 1, \ldots, I \\ \text{where } & p_{ik} = P(Y_{ik}^* > 0) \\ \text{and } & Y_{ik}^* = \mu_k + \gamma \nu_{ik} + \epsilon_{ik}, \end{aligned}$$

where the Y_{ik}^* are latent variables and $\epsilon_{ik} \overset{iid}{\sim} N(0,1)$; see Albert and Chib [121]. Using this formulation, we can adapt the SMN method in Section 5.3.4 accordingly.

We acknowledge that cutoff values for detecting the practical significance of a potential outlier appear hard to select, as they can be context-, model-, and data-specific. We have done simulations (not shown in this paper) to investigate different cutoff values for the SMN method and found that the $P(\lambda_i > c|\boldsymbol{y})$ varies with different predetermined values for the cutoff c. For example, in the unbalanced design, with $c = 1$ as the cutoff criterion, $P(\lambda_4 > 1|\boldsymbol{y})$ is 1.00 for Trial 4 but around 0.45 for the other trials, as is shown in Table 5.9. When the cutoff value is set to be 5, however, $P(\lambda_4 > 5|\boldsymbol{y})$ keeps being 1 for Trial 4, but $P(\lambda_i > 5|\boldsymbol{y})$ is around 0.06 for the other trials. If we let the cutoff value be 10, the $P(\lambda_4 > 10|\boldsymbol{y})$ is again 1 for Trial 4 but around 0.02 for the other trials. The cutoff selection issue also plagues the other measures. However, simulations like those in Section 6 can still help guide this selection, based on the design of interest and information content of each trial (as measured by sample sizes and variance estimates).

Table 5.9: Scale mixtures of normals for unbalanced and balanced designs in the simulation study

	Unbalanced Design			Balanced Design		
Trial i	$\widehat{\lambda_i}$	$P(\lambda_i > 1 \mid \boldsymbol{y})$	P(Positive)	$\widehat{\lambda_i}$	$P(\lambda_i > 1 \mid \boldsymbol{y})$	P(Positive)
1	0.89	0.45	0	0.95	0.48	0
2	0.89	0.45	0	0.95	0.48	0
3	0.89	0.45	0	0.95	0.48	0
4	**583.69**	**1.00**	**1.00**	**575.53**	**1.00**	**1.00**
5	0.89	0.45	0	0.95	0.48	0
6	0.89	0.45	0	0.95	0.48	0
7	0.89	0.45	0	0.95	0.48	0
8	1.33	0.61	0	**572.50**	**1.00**	**1.00**
9	0.92	0.46	0	0.92	0.46	0
10	0.92	0.46	0	0.92	0.46	0
11	0.92	0.46	0	0.92	0.46	0
12	0.92	0.46	0	0.92	0.46	0

Note: $\widehat{\lambda_i}$: the mean value of 1000 exceedance posterior medians; $P(\lambda_i > 1 \mid \boldsymbol{y})$: mean values of 1000 probablities that λ_i is larger than 1; P(Positive): indicates probability of positive results (exceedance probability greater than 0.95) over the 1000 simulations. Bold values mark detected outlying trials.

Turning to future work, we are interested in developing methods for automatic downweighting of outlying trials. Borrowing the idea of Ibrahim and Chen [122], power priors offer a simple and intuitive approach, by raising the outlying likelihood to a power $\alpha_0 \in [0, 1]$, and re-standardizing the result to a proper distribution. Hobbs et al. [123] proposed an extension called hierarchical commensurate and power priors for adaptive incorporation of information, which could also be applied here. Future work also looks toward extension to outlier detection for models incorporating baseline covariates and individual-level patient data. Note that when baseline covariates are present, the definition of outliers ought to be modified, since a trial could then be outlying simply by having an unusual population (e.g., more older enrollees).

Of course, criticisms of our methods can be made. In this paper, we have only considered the trials that were already collected in an NMA. However, trials that were candidates for inclusion but were omitted is another issue worthy of attention. This can be broadly related to the issue of publication bias, the concern that studies with significant results are more likely to be published, and published studies (especially those in the meta-analyst's own language) are more likely to be included in an NMA [28]. Another limitation of this paper is that we have not considered *arm-level* outliers, even though it may be that some treatment arms may not be suitable to be synthesized with the others in an NMA. Approaches for these issues and their evaluation await further exploration.

Chapter 6

Conclusions

6.1 Summary of major findings

This thesis set out to explore novel Bayesian hierarchical methods in the context of network meta-analysis in order to provide better evidence synthesis. Our contribution lies in providing more proper and comprehensive reporting for network meta-analysis, which ultimately enables better decision making for patients, health carers, and policy makers. This thesis has also sought to solve the thorny issues related to nonignorable missingness in the highly sparse network meta-analysis data set, and when some trials in the network are outlying. The proposed selection models method and outlier detection measures successfully incorporates nonignorable missingness and detect outlying trials respectively, facilitating more precise evidence synthesis.

We now summarize our findings and contributions specifically for each chapter. In Chapter 2, we proposed a novel arm-based Bayesian hierarchical method from the missing data perspective, which focuses on estimating direct summary, i.e., event rates, instead of relative summary, i.e., ORs. We compared its performance with the current contrast-based method and the traditional pairwise comparison method, and illustrated how to provide proper summaries with two published network meta-analyses. Our method clears away obstacles for patients in order to comprehensively trade-off efficacy and safety measures. We then built upon this method to incorporate nonignorable missingness in Chapter 3 with the proposed selection models method. It comprises two parts: a model of interest, and a model of missingness. Joint modeling avoided the

bias created by ignoring missingness, or by considering all missing data to be missing at random, verified by various simulations. Chapter 4 investigated whether removal of trials had influence on the estimates of treatment effects, and gathered empirical evidence by reanalyzing 14 published network meta-analyses after excluding specific trials. We found that some trial exclusions can substantially affect the estimated treatment effects. After exploring influence in exclusions of trials that should have been included, in Chapter 5 we studied outlying trials that should not be included but were included. Four Bayesian outlier detection measures were proposed and their performance was studied via application to a diabetes data set and simulation studies.

Turning to practice, this thesis provides a practical guide on how to carry out indirect comparisons and multiple treatment comparisons. Through a series of illustrative examples, it also showed in down-to-earth terms how to synthesize available evidence and make better decisions using appropriate Bayesian statistical techniques. The intended audience for this thesis is therefore mixed, including statisticians, decision makers, and health economists. We believe people with an interest in the production of systematic reviews will also find much that is both new and highly relevant to their work.

Of course, limitations remain. Heterogeneity and inconsistency are the two major issues in network meta-analysis. We incorporated heterogeneity with random effects but did not quantitatively compute it. Inconsistency was out of scope for this thesis, although there are already a number of papers investigating it. Another limitation lies in the approach we used for handling nonignorable missingness. After all, the selection models method serves only as sensitivity analysis, methods searching for the particular reason of nonignorable missingness and methods remedying it await further exploration. Selection of cutoff values for detecting the practical significance of a potential outlier is the third issue worthy of future efforts as these cutoff values can be context-, model-, and data-specific.

6.2 Extensions and future work

In addition to providing immediate important findings, the work in this thesis has also motivated a variety of future projects.

6.2.1 NMA involving multiple type of outcomes

Future work looks to extending our method to include mixed types of outcomes, for example, a time-to-event safety outcome paired with a binary efficacy outcome, or a binary safety outcome paired with a continuous efficacy outcome. For example, we will look at a network meta-analysis of HIV viral load data. Here the primary efficacy endpoint is the number of patients with HIV-RNA below the limit of assay detection (50 copies) at 48 weeks (i.e., binomial outcome), and the safety endpoint is time-to-event disease progression or death (time-to-event outcome). Another example is a network meta-analysis of cancer clinical trials where the tumor size serves as the efficacy endpoint (continuous outcome) while the adverse event death is the safety endpoint (binary outcome).

6.2.2 Evidence synthesis of observational studies

There is no doubt that the RCT produces the most reliable evidence on the comparative effectiveness of interventions; however, good, conclusive randomized evidence may not exist for every treatment decision [124]. Thus there are occasions when it will be necessary to consider observational data on which to estimate effectiveness. In other cases, though RCTs are available, concerns regarding their quality may require supplementary observational evidence.

When observational evidence is considered either instead of, or in combination with, randomized evidence, the question arises as to whether we need to apply new methodology in order to sensibly synthesize it. Observational data potentially produce more biased estimates than data from randomized trials. Current approaches for combining randomized and observational evidence include deriving a prior distribution from observational evidence, bias allowance models for the observational data, and hierarchical models with an extra level of variation to allow for variability in effect sizes between different sources of evidence. However, these methods are all based on largely untestable assumptions. Future investigation here are needed.

6.2.3 Computing and software development

In order for the proposed methods to be fully embraced by researchers in non-statistical fields, we need to provide the means which to implement them conveniently. The most obvious way for achieving this would be by creating a freely available R package with user-friendly functions to conduct the network meta-analysis.

For modest or slightly large sample sizes, usually at the aggregate level, it sometimes takes a couple of hours to analyze the data. We aim to call BUGS software from R in this situation. However, individual patient level data can be a bit computationally burdensome. Implementing the methods to these data in BUGS and R requires much more effort, thus we may need to develop our own package using C++ subroutines embedded with an R wrapper.

References

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

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

[3] A. Whitehead. *Meta-Analysis of Controlled Clinical Trials*. John Wiley, New York, 2003.

[4] S.J. Tanenbaum. Comparative effectiveness research: evidence-based medicine meets health care reform in the usa. *Journal of evaluation in clinical practice*, 15(6):976–984, 2009.

[5] V.S. Conn, T.M. Ruppar, L.J. Phillips, and J.A.D. Chase. Using meta-analyses for comparative effectiveness research. *Nursing outlook*, 60(4):182–190, 2012.

[6] D.L. Sackett, W. Rosenberg, J.A. Gray, R.B. Haynes, and W.S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[7] S. Timmermans and A. Mauck. The promises and pitfalls of evidence-based medicine. *Health Affairs*, 24(1):18–28, 2005.

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

[9] T. Li, M.A. Puhan, S.S. Vedula, S. Singh, K. Dickersin, et al. Network meta-analysis-highly attractive but more methodological research is needed. *BMC medicine*, 9(1):79, 2011.

[10] Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the pharmaceutical benefits advisory committee (version 4.3), 2008.

[11] G.A. Wells, S.A. Sultan, L. Chen, M. Khan, and D. Coyle. Indirect evidence: indirect treatment comparisons in meta-analysis, 2009.

[12] National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal, 2008. Available from `http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf` [Accessed 2008 Jul 30].

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

[14] F. Dominici, G. Parmigiani, R.L. Wolpert, and V. Hasselblad. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *Journal of the American Statistical Association*, 94(445):16–28, 1999.

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

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

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

[18] R.M. Nixon, N. Bansback, and A. Brennan. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Statistics in Medicine*, 26(6):1237–1254, 2007.

[19] S. Dias, A.J. Sutton, A.E. Ades, and N.J. Welton. A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

[20] J.P. Jansen, B. Crawford, G. Bergman, and W. Stam. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value in Health*, 11(5):956–964, 2008.

[21] N.J. Welton, N.J. Cooper, A.E. Ades, G. Lu, and A.J. Sutton. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. *Statistics in Medicine*, 27(27):5620–5639, 2008.

[22] S. Dias, N.J. Welton, A.J. Sutton, and A.E. Ades. NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials, 2011a. Last updated April 2012. Available from `http://www.nicedsu.org.uk`.

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

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

[25] S. Dias, N.J. Welton, A.J. Sutton, D.M. Caldwell, G. Lu, and A.E. Ades. NICE DSU technical support document 4: inconsistency in networks of evidence based on randomised controlled trials, 2011b. Last updated April 2012. Available from `http://www.nicedsu.org.uk`.

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

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

[28] M. Borenstein, L.V. Hedges, J.P.T. Higgins, and H.R. Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2011.

[29] N.J. Cooper, A.J. Sutton, D. Morris, A.E. Ades, and N.J. Welton. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine*, 28(14):18611881, 2009.

[30] G. Lu, N.J. Welton, J.P.T. Higgins, I.R. White, and A.E. Ades. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. *Research Synthesis Methods*, 2(1):43–60, 2011.

[31] H.P. Piepho, E.R. Williams, and L.V. Madden. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics*, 68(4):1269–1277, 2012.

[32] H. Zhao, J.S. Hodges, H. Ma, Q. Jiang, and B.P. Carlin. Arm-based approaches for detecting inconsistency in network meta-analysis. *Research Report 2014–006, Division of Biostatistics, University of Minnesota. Submitted to* Biometrics, 2014.

[33] D.M. Caldwell, A.E. Ades, and J.P.T. Higgins. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*, 331(7521):897–900, 2005.

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

[35] A. Cipriani, C. Barbui, G. Salanti, J. Rendell, R. Brown, S. Stockton, M. Purgato, L.M. Spineli, G.M. Goodwin, and J.R. Geddes. Comparative efficacy and acceptability of antimanic drugs in acute mania: a multiple-treatments meta-analysis. *The Lancet*, 378(9799):1306–1315, 2011.

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

[37] M. Pahor, B.M. Psaty, M.H. Alderman, W.B. Applegate, J.D. Williamson, C. Cavazzini, and C.D. Furberg. Health outcomes associated with calcium antagonists compared with other first-line antihypertensive therapies: a meta-analysis of randomised controlled trials. *The Lancet*, 356(9246):1949–1954, 2000.

[38] F. Song, Y.K. Loke, T. Walsh, A. M. Glenny, A.J. Eastwood, and D.G. Altman. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ*, 338:b1147, 2009. doi: http://dx.doi.org/10.1136/bmj.b1147.

[39] T. Palmerini, G. Biondi-Zoccai, D.D. Riva, C. Stettler, D. Sangiorgi, F. D'Ascenzo, T. Kimura, C. Briguori, M. Sabatè, H. Kim, et al. Stent thrombosis with drug-eluting and bare-metal stents: evidence from a comprehensive network meta-analysis. *The Lancet*, 379(9824):1393–1402, 2012.

[40] J.P. Daniels, L.J. Middleton, R. Champaneria, K.S. Khan, K. Cooper, B.W.J. Mol, and S. Bhattacharya. Second generation endometrial ablation techniques for heavy menstrual bleeding: network meta-analysis. *BMJ*, 344:e2564, 2012. doi: http://dx.doi.org/10.1136/bmj.e2564.

[41] S.Y. Wang, H. Chu, T. Shamliyan, H. Jalal, K.M. Kuntz, R.L. Kane, and B.A. Network meta-analysis of margin threshold for women with ductal carcinoma in situ. *Journal of the National Cancer Institute*, 104(7), 2012. doi: 10.1093/jnci/djs142.

[42] D.G. Altman, J.J. Deeks, and D.L. Sackett. Odds ratios should be avoided when events are common. *BMJ*, 317(7168):1318, 1998. doi: http://dx.doi.org/10.1136/bmj.317.7168.1318.

[43] J. Deeks. When can odds ratios mislead?: Odds ratios should be used only in case-control studies and logistic regression analyses. *BMJ*, 317(7166):1155, 1998.

[44] H.T.O. Davies, I.K. Crombie, and M. Tavakoli. When can odds ratios mislead? *BMJ*, 316(7136):989–991, 1998.

[45] B.M. Psaty, T. Lumley, C.D. Furberg, G. Schellenbaum, M. Pahor, M.H. Alderman, and N.S. Weiss. Health outcomes associated with various antihypertensive therapies used as first-line agents: a network meta-analysis. *JAMA*, 289(19):2534–2544, 2003.

[46] T.A. Trikalinos, A.A. Alsheikh-Ali, A. Tatsioni, B. Nallamothu, and D.M. Kent. 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.

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

[48] G. Salanti, A.E. Ades, and J.P.A. Ioannidis. 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.

[49] H. Chung and T. Lumley. Graphical exploration of network meta-analysis data: the use of multidimensional scaling. *Clinical Trials*, 5(4):301–307, 2008.

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

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

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

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

[54] B. Jones, J. Roger, P.W. Lane, A. Lawton, C. Fletcher, J.C. Cappelleri, H. Tate, and P. Moneuse. Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical statistics*, 10(6):523–531, 2011.

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

[56] E.H. Turner, A.M. Matthews, E. Linardatos, R.A. Tell, and R. Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358(3):252–260, 2008.

[57] C.I. Coleman, O.J. Phung, J.C. Cappelleri, W.L. Baker, J. Kluger, C.W. White, and D.M. Sobieraj. Use of mixed treatment comparisons in systematic reviews. *http://www.ncbi.nlm.nih.gov/books/NBK107330/*, 2012.

[58] E.J. Mills, S. Kanters, K. Thorlund, A. Chaimani, V. Areti-Angeliki, and J.P.A. Ioannidis. The effects of excluding treatments from network meta-analyses: survey. *BMJ*, 347:f5195, 2013. doi: 10.1136/bmj.f5195.

[59] R.J. Little and D.B. Rubin, editors. *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, New York, 2002.

[60] S. Dias, N.J. Welton, A.J. Sutton, and A.E. Ades. NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. *National Institute for Health and Clinical Excellence, London, UK*, 2011. Available from `http://www.nicedsu.org.uk`.

[61] S. Dias, N.J. Welton, A.J. Sutton, and A.E. Ades. NICE DSU technical support document 5: evidence synthesis in the baseline natural history model, 2011. Available from `http://www.nicedsu.org.uk`.

[62] A.J. Sutton and N.J. Welton. NICE DSU technical support document 6: embedding evidence synthesis in probabilistic cost-effectiveness analysis: software choices, 2011. Available from `http://www.nicedsu.org.uk`.

[63] J.G. Ibrahim, H. Chu, and M. Chen. Missing data in clinical studies: issues and methods. *Journal of Clinical Oncology*, 30(26):3297–3303, 2012.

[64] D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[65] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken, 1987.

[66] J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall/CRC, Boca Raton, 1997.

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

[68] H. Chu, L. Nie, Y. Chen, Y. Huang, and W. Sun. Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: Methods for the absolute risk difference and relative risk. *Statistical Methods in Medical Research*, 21(6):621–633, 2010.

[69] A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[70] P. Gustafson. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Statistics in medicine*, 24(8):1203–1217, 2005.

[71] D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. A bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

[72] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

[73] A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *JASA*, 85(410):398–409, 1990.

[74] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

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

[76] S.N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of internal medicine*, 130:995–1004, 1999.

[77] S. Yusuf, R. Peto, J. Lewis, R. Collins, P. Sleight, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases*, 27(5):335–371, 1985.

[78] H. Hong, H. Chu, J. Zhang, and B.P. Carlin. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Report 2012-018, Division of Biostatistics, University of Minnesota.*

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

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

[81] J. Zhang and F.Y. Kai. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*, 280(19):1690–1691, 1998.

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

[83] D.M. Caldwell, A.E. Ades, and J.P.T. Higgins. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *British Medical Journal*, 331(7521):897–900, 2005.

[84] F.A. Achana, N.J. Cooper, S. Dias, G. Lu, S.R.C. Rice, D. Kendrick, and A.J. Sutton. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Statistics in medicine*, 32(5):752–771, 2013.

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

[86] G. van Valkenhoef, T. Tervonen, B. de Brock, and H. Hillege. Algorithmic parameterization of mixed treatment comparisons. *Statistics and Computing*, 22(5):1099–1111, 2012.

[87] J.P. Jansen. Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods*, 3(2):177–190, 2012.

[88] P. Saramago, A.J. Sutton, N.J. Cooper, and A. Manca. Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in Medicine*, 31(28):3516–3536, 2012.

[89] J.J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492, 1976.

[90] A. Mason, S. Richardson, and N. Best. Two-pronged strategy for using *DIC* to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, 7(1):109–146, 2012.

[91] P. Gustafson, S. Hossain, and Y.C. MacNab. Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, 34(3):377–390, 2006.

[92] R. Natarajan and C.E. McCulloch. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics*, 7(3):267–277, 1998.

[93] J. Wakefield. Ecological inference for 2$\times$ 2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–445, 2004.

[94] C. Jackson, N. Best, and S. Richardson. Improving ecological inference using individual-level data. *Statistics in Medicine*, 25(12):2136–2159, 2006.

[95] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[96] M.C. Fiore, W.C. Bailey, S.J. Cohen, S.F. Dorfman, M.G. Goldstein, E.R. Gritz, R.B. Heyman, J. Holbrook, C.R. Jaen, T.E. Kottke, et al. Smoking cessation. clinical practice guideline no. 18 (ahcpr publication no. 96-0692). rockville, md: Us department of health and human services. *Public Health Service, Agency for Health Care Policy and Research*, 1996.

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

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

[99] G. Salanti. 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.

[100] R. Ara, A. Pandor, J. Stevens, A. Rees, and R. Rafia. Early high-dose lipid-lowering therapy to avoid cardiac events: a systematic review and economic evaluation. *Health Technol Assess*, 13:1–74,75–118.

[101] J. Ballesteros. 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.

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

[103] M.J. Eisenberg, K.B. Filion, D. Yavin, P. Bélisle, S. Mottillo, L. Joseph, A. Gervais, J. O'Loughlin, G. Paradis, S. Rinfret, et al. Pharmacotherapies for smoking

cessation: a meta-analysis of randomized controlled trials. *Canadian Medical Association Journal*, 179(2):135–144, 2008.

[104] L.J. Middleton, R. Champaneria, J.P. Daniels, S. Bhattacharya, K.G. Cooper, N.H. Hilken, P. ODonovan, M. Gannon, R. Gray, and K.S. Khan. 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. doi: http://dx.doi.org/10.1136/bmj.c3929.

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

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

[107] M.A. Puhan, L.M. Bachmann, J. Kleijnen, G. ter Riet, and A.G. Kessels. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. *BMC medicine*, 7(1), 2009. 10.1186/1741-7015-7-2.

[108] V. Thijs, R. Lemmens, and S. Fieuws. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. *European heart journal*, 2008. doi: 10.1093/eurheartj/ehn106.

[109] D.G. Altman and J.M. Bland. Measurement in medicine: the analysis of method comparison studies. *The statistician*, 32:307–317, 1983.

[110] J.M. Bland and D.G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

[111] J.M. Bland and D.G. Altman. Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2):135–160, 1999.

[112] C.L. Gillies, K.R. Abrams, P.C. Lambert, N.J. Cooper, A.J. Sutton, R.T. Hsu, and K. Khunti. Pharmacological and lifestyle interventions to prevent or delay

type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. *BMJ*, 334(7588):299, 2007.

[113] X. Wang, X. Qin, H. Demirtas, J. Li, G. Mao, Y. Huo, N. Sun, L. Liu, and X. Xu. Efficacy of folic acid supplementation in stroke prevention: a meta-analysis. *The Lancet*, 369(9576):1876–1882, 2007.

[114] D.J. Spiegelhalter, K.R. Abrams, and J.P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, 2004.

[115] Y. Ding and H. Fu. Bayesian indirect and mixed treatment comparisons across longitudinal time points. *Statistics in Medicine*, 32(15):26132628, 2012.

[116] J. Zhang, H. Chu, H. Hong, J.D. Neaton, B.A. Virnig, and B.P. Carlin. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Research Report 2013–018, Division of Biostatistics, University of Minnesota*.

[117] D.B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

[118] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3rd edition, 2013.

[119] B.P. Carlin and T.A. Louis. *Bayesian Methods for Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3rd edition, 2009.

[120] A. Gelman, X. Meng, and H.S. Stern. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6(4):733–807, 1996.

[121] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

[122] J.G. Ibrahim and M. Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

[123] B.P. Hobbs, B.P. Carlin, S.J. Mandrekar, and D.J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

[124] A.J. Sutton, N.J. Cooper, K.R. Abrams, and A.E. Ades. *Evidence synthesis for decision making in healthcare*. John Wiley & Sons, West Sussex, 2012.

Appendix A

Glossary for abbreviations

CER: comparative effectiveness research.

EBM: evidence based medicine.

RCT: randomized controlled trial.

NMA: network meta-analysis.

MCAR: missing completely at random.

MAR: missing at random.

MNAR: missing not at random.

RD: risk difference.

RR: risk ratio.

OR: odds ratio.

CB: contrast-based.

AB: arm-based.

MBHMM: multivariate Bayesian hierarchical mixed model.

NICE: National Institute for Health and Clinical Excellence.

RD_i^k: relative distance after deleting trial i for a particular treatment k.

ARD_i: average relative distance after deleting trial i.

STR: standardized trial residual.

SMN: scale mixtures of normal.

ARC: absolute relative change.