Detecting Inconsistency and Using Non-randomized Studies in Research Synthesis

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Hong Zhao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Advised by Bradley P. Carlin, Ph.D

May, 2016

Acknowledgements

It has been a wonderful experience to be a graduate student in the Devision of Biostatistics at University of Minnesota. I would like to extend my gratitude to many people who made this long process of pursuing a Ph.D. much more enjoyable than it otherwise could have been.

First and foremost, I am heartily thankful to my advisor, Dr. Bradley Carlin, who is an enthusiastic and energetic researcher. His guidance is essential not only to my current study, but also to my path from a student to an independent biostatistician in the near future. He always listens to me open-mindedly and encourages me to think on my own, guides my scientific writing through each sentence of my papers, and supports me when I meet difficulties. I have benefited greatly from his outstanding expertise in Bayesian analysis and learned from him to face problems positively.

I also owe my deepest gratitude to Dr. James Hodges, who shared with me incredible intuition on my research. He provided enormous help with his expertise in hierarchical modeling and its application in network meta-analysis through class projects and thesis development. His excellent suggestions were absolutely essential in consolidating my research projects. Thanks also go to the other members of my Ph.D. committee. Dr. Haitao Chu, an expert in network meta-analysis, helped me to understand the main model and its assumptions. Dr. Sue Duval provided insightful comments during the course of my research, especially at committee meetings. All their suggestions helped

my research to be more sound and complete.

I am grateful to Dr. Qi Jiang and Dr. Haijun Ma from Amgen for arranging the projects, and for their insight from an industry perspective. Thanks also go to Dr. Carlin's previous students who helped me with my current study. Hwanhee Hong provided invaluable guidance on many technique details of network meta-analysis methods, and I greatly appreciated her discussion when I first started my research. Brian Hobbs and Thomas Murray gave me a lot of help on the details of commensurate priors, and Brian also introduced me to several casual inference methods for analyzing non-randomized data.

I highly appreciate the support from my dear friends throughout my study. Wherever they are, they deserve my sincere thanks and best wishes.

Finally, no words can express my gratitude for my beloved family. I am grateful to my parents and parents-in-law for their support and endless love. Special thanks to my dearest Ming, for his comfort and encouragement when I was frustrated. I am so lucky that he is always by my side sharing my sorrow and happiness. My strength comes from him and our son Benjamin, who brings me joy every day.

Dedication

To my parents, Encang Zhao and Fangxian Wu, and my brothers, Jun Zhao and Peng Zhao, for their endless love, support and encouragement during my education.

Abstract

In scientific research, multiple studies on the same intervention occur for many reasons, such as using different study populations or designs. Research synthesis attempts to integrate different data on the same topic for the purpose of making generalizations, and provides us with a formal method to systematically combine all available evidence. In this thesis, we focus on the synthesis of evidence from multiple clinical trials using network meta-analysis, and the more challenging problem of combining information from randomized clinical trials and less rigorous observational studies.

Network meta-analysis (NMA) is an extension of standard pairwise meta-analysis to permit combination of results on more than two treatments. This enables both direct and indirect comparisons of treatments, and addresses the comparative effectiveness or safety of the treatments based on all sources of data. Current NMA methods are usually based on a contrast-based (CB) model to estimate the relative treatment effects for each study. While popular and often effective, this model suffers from certain limitations. An alternative is the arm-based (AB) model, which estimates the mean response directly for each treatment. Compared to the CB framework, AB models are more straightforward to interpret, especially when implemented in a missing-data framework, by allowing use of a common baseline treatment across all trials.

In a NMA, when direct and indirect evidence differ, the analysis is said to suffer from inconsistency, and the treatment effect estimates may be biased. Inconsistency detection methods using CB models have already been developed, but no corresponding method based on the newer AB models has yet been proposed. Here, we develop a Bayesian AB approach to detecting inconsistency. After detecting inconsistency, formal diagnostic tests should be performed to check whether this violation of assumption results in the

change of treatment effects. Therefore, we next explore whether the trial-arm combinations that are sources of inconsistency are influential or outlying observations. To do this, we modify the "constraint case" method to produce diagnostics suitable for generalized linear models in NMA using either AB or CB models, where the outcome is binary. Lastly, we develop methods to combine the data from a randomized clinical trial and a propensity score-matched non-randomized study using commensurate priors. The approach determines the proper degree of borrowing from the non-randomized data by the similarity of the estimated treatment effects in the two studies. Performance of all our methods is evaluated via both example datasets and simulation studies.

In summary, this dissertation work enables improved research synthesis in biomedical applications and sheds light on future research directions in the aforementioned areas.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In scientific research, multiple studies on the same phenomenon or hypothesis arise for many reasons. Research syntheses attempt to integrate different data sources on the same topic for the purpose of creating generalizations [1], and provide a formal method to systematically evaluate the scientific evidence. In this thesis, we first investigate inconsistency detection in clinical trial network meta-analysis, followed by diagnostics of influential observations and outliers for the hierarchical models used in NMA. Finally, we propose a two-step method to combine information from both randomized controlled trials (RCTs) and observational studies (OSs) to make causal inferences.

1.1 Inconsistency Detection in NMA

In comparative effectiveness research, *network meta-analysis*, also known as *mixed treatment comparisons* [2, 3], is an extension of the pairwise meta-analysis method [4] to compare the results from two or more studies that have at least one treatment in common. This enables both direct and indirect comparisons and addresses the comparative effectiveness or safety of the interventions based on all sources of data. Because of its

enormous potential, interest in this method has grown substantially, and application of NMA is increasingly common [5]. Moreover, government regulators and national health agency staff have increasingly adopted such methods [6, 7].

One of the key assumptions for NMA is *consistency* [2, 8, 9, 10]. As shown in the NMA network graph in Figure 1.1, each vertex represents a treatment and each edge represents a pairwise comparison. In Figure 1.1(a), the comparison of placebo (P) vs. A and P vs. B is direct evidence, and there is no head-to-head comparison between A and B. To make inference about A vs. B, we can only use the indirect information from studies of P vs. A and P vs. B. In Figure 1.1(b), where we instead have direct evidence comparing A and B, we can now compare A and B using both direct and indirect information.

When direct and indirect evidence differ in a NMA, the model is said to suffer from *inconsistency*. Using Figure 1.1(b) as an example, consistency holds if

$$d_{AB} = d_{PB} - d_{PA}, \tag{1.1}$$

where d_{AB} is the average treatment effect (e.g., log odds ratio) of B vs. A (the treatment effect is "average" due to the possible existence of between-study heterogeneity). Inconsistency can arise due to many causes, including non-comparability of trials, different control groups, or differences in patient characteristics. If inconsistency is present, information from different sources may disagree, and the treatment effect estimates obtained from the NMA may be biased or hard to interpret. Many remedies have been suggested, including adding fixed covariate effects to account for different patient characteristics. As a first step, it is important to detect inconsistency and its likely sources, but as the number of studies increases, network complexity often precludes a quick diagnosis of inconsistency and identification of which studies or treatments are its sources.

In Chapter 2 of this dissertation, we review inconsistency detection methods using various NMA models, and propose another. Bayesian hierarchical models for NMA

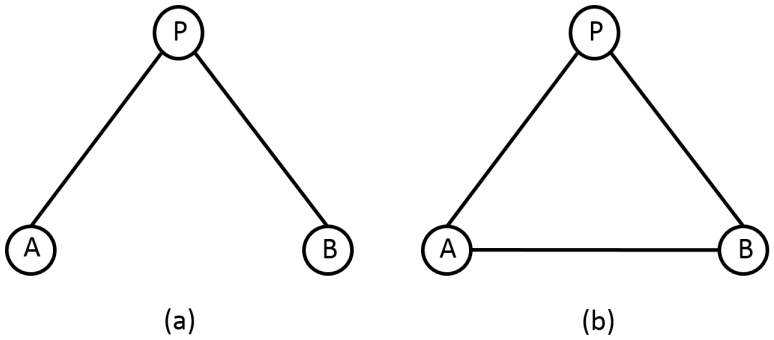

Figure 1.1: Networks of treatment comparisons. Each vertex represents a treatment and each edge represents a pairwise comparison. (a) Indirect comparison only between A and B. (b) Both direct and indirect comparison between A and B.

with binary outcomes have been well studied. The logit model initially proposed by Lu and Ades [3] is a *contrast-based* (CB) model, which uses the log odds ratio to estimate relative effects of two treatments. In the CB framework, a fixed effects model is used when we assume treatment effects do not vary between studies, while a random effects model is implemented when heterogeneity between trials is allowed.

Inconsistency detection is thoroughly discussed by Lu and Ades [8] for CB models using two illustrative datasets [11, 12]. They proposed examining *loop-based inconsistency* by adding one parameter called an *inconsistency factor* (ICF), w, to the consistency relationship shown in (1.1): $d_{AB} = d_{PB} - d_{PA} + w_{ABP}$, where three treatments are connected in a cycle, like the loop ABP in Figure 1.1(b). The posterior distribution of w reflects the extent of inconsistency in a particular evidence loop. Most recently, the back-calculation and node-splitting methods proposed by Dias et al. [9] using CB models have streamlined the process of inconsistency detection by looking in more detail at specific comparisons. Both approaches extend inconsistency detection to any network, not merely triangular node structures. Another type of inconsistency detection method is the design-by-treatment interaction model. Instead of exploring inconsistency arising

from studies that provide direct or indirect information on a particular comparison, this approach introduces a further source of variation to allow for inconsistency in the network. It was introduced by Lumley [2], who defines an inconsistency factor for each different design (i.e., "set of treatments compared within the study" as in [10]), and assumes that all inconsistency factors follow a common random effects distribution. However, the model is constructed only for two-arm trials. Later, Higgins et al. [10] and White et al. [13] conducted inconsistency detection in NMA using multivariate meta-regression (treating inconsistency parameters as fixed effects), extending Lumley's model to include multi-arm trials. Shortly thereafter, Jackson et al. [14] proposed a special case of the design-by-treatment interaction model [10, 13], which is a generalization of Lumley's model, by treating inconsistency parameters as random effects. (In the Appendix, we describe some of these models and compare them to our approach.)

All of the inconsistency detection methods mentioned above are based on CB models. However, these methods have certain limitations due to their focus on relative effects [15, 16]. Several authors (e.g., Hong et al. [15] and Zhang et al. [16]) have suggested an alternative method for NMA often called the *arm-based* (AB) model, which models the absolute (rather than relative) effect of each treatment. The AB model requires an assumption of exchangeability of treatment arms across studies, while the CB model assumes exchangeability of the *relative* treatment effects compared to baseline ("treatment contrasts"), measured on a common scale (e.g., OR) across studies. Specifically, Shuster et al. [17] distinguished two types of assumptions for random effects in meta-analysis: studies at random (SR) and effects at random (ER); AB models assume the former, while CB models assume the latter. See Hong et al. [15], its discussion, and rejoinder for a recent debate regarding the relative merits of CB and AB models.

As mentioned above, inconsistency detection methods using CB models have already been developed, but there remains a need to propose a corresponding method for AB

models. Like earlier authors [9, 10], we review and compare to Lu and Ades' method, since it is the reference standard for this research, and the comparison can be made for both loop- and non-loop-based methods. Under our AB model framework, we look for possible inconsistency in two ways: (1) by using estimates of 4 fixed effects in the AB model to test the discrepancy of the direct and indirect evidence for comparing two treatments, either loop-based or not, and (2) by using estimates of specific AB random effects to detect inconsistency at certain trial-by-arm combinations, once inconsistency has been detected through the AB model fixed effects.

1.2 Diagnostics for Generalized Linear Hierarchical Models in NMA

Although many methods for inconsistency detection have been proposed, it remains unclear how to proceed after inconsistency is detected [10]. Dias et al. [18] suggested any adjustment in response to inconsistency is post hoc and one should always reconsider the entire network if inconsistency is identified. For instance, even if we can reduce inconsistency using an adjustment method, it could result in very different estimates with different interpretations. Later, Jackson et al. [14] proposed a method to incorporate inconsistency using random effects, which estimates treatment effects without assuming consistency. However, if inconsistency is identified in a NMA, drawing inferences with inconsistent evidence is controversial and might only be appropriate in reasonably large networks.

We believe after inconsistency is identified, the key question is whether it affects substantive conclusions drawn in the NMA. Therefore, we aim to examine such discrepancies in a NMA from a diagnostic point of view, which is routinely applied in regression settings. Recently, Zhang et al. [19] proposed a method to detect outlying

trials in NMA, but not at the observation (trial-by-arm) level. Even more recently, Lin et al. [20] examined the influence of treatment exclusion for CB and AB models in a NMA. To our knowledge, there has been no discussion of how to perform diagnostics in NMA at a trial-by-arm level.

In Chapter 3, we present methods to detect influential and outlying observations, which have a large effect on parameter estimates or which deviate markedly from other observations, respectively. It is essential to investigate the relationship between those influential or outlying observations and detected sources of inconsistency, to see whether inconsistency affects parameter estimation substantially. In our work, we extend an approach introduced by Hodges [21] to diagnostics for generalized linear hierarchical AB models used in NMA, then to CB models, and finally perform influential observation and outlier detection on both simulated and real example datasets.

1.3 Combining RCTs with Observational Studies

Another area in research synthesis is development of methods for combining observational data with clinical trial data. In Chapter 4, we discuss the need to incorporate observational studies and introduce a causal inference method for analyzing randomized and non-randomized data.

1.3.1 Randomized Clinical Trials and Observational Studies

RCTs and OSs are the two primary approaches for evaluating the effectiveness of therapeutic interventions. In RCTs, subjects are randomly assigned to treatment groups to eliminate treatment selection bias, which can negatively impact the estimate of association between treatment and outcome. This is based on the assumption that the distribution of observed and unobserved covariates is balanced across the treatment groups on average after randomization. Therefore, RCTs have long been recognized as

the gold standard for testing the efficacy or safety of an intervention, and are considered to yield the highest grade of evidence in the hierarchy of research designs [22, 23]. Although RCTs facilitate valid treatment comparisons, they may be restrictive if the study population, those willing to undergo randomization, represent only a subset of the patient population. By contrast, OSs may enroll patient cohorts that better reflect the broader patient population because they often use less restrictive inclusion criteria. However, OSs can suffer from selection bias, and may yield invalid treatment comparisons even after adjusting for known confounders, which are associated with both outcome and treatment selection.

Combining RCTs and OSs in research synthesis is often criticized due to the limitations of OSs. Table 1.1 shows a "grades of evidence" ranking [23] using internal validity as the primary criterion. The lowest grade includes anecdotal case histories and expert opinion, while OSs such as well-designed cohort or case control studies fall at intermediate levels. Recent advances in epidemiology and statistics, which have enhanced understanding of the implications of study design and analysis for causal inference (the process of analyzing a causal connection based on the conditions of the occurrence of an effect [24]), challenge this evidence hierarchy [25]. MacLehose et al. [26] concluded that discrepancies between RCT and OS estimates of effect size and outcome frequency for different groups were small for high-quality studies, but potentially large for low-quality studies. Concato et al. [27] found that well-designed OSs provided similar estimates for treatment effects compared to those in RCTs across five clinical topics and 99 reports evaluated. More recent literature supports the opinion that systematic reviews of treatment effects should not be restricted to specific study types in all cases. Moreover, because RCTs often admit limited patient populations, conclusions obtained from RCTs may be limited in scope and thereby contradict results obtained from studies of broader patient cohorts [28]. Furthermore, if treatment effects are estimated from a single RCT,

Table 1.1: Grades of evidence: hierarchical rankings for different study designs.

I	Evidence obtained from at least one properly randomized, controlled trial.
II-1	Evidence obtained from well-designed controlled trials without randomization.
II-2	Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.
II-3	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments could also be regarded as this type of evidence.
III	Opinions of respected authorities, based on clinical experience; descriptive studies and case reports; or reports of expert committees.

they might not be more reliable than those obtained after integrating the information acquired from several well-designed OSs. Therefore, examining the extent to which well-designed non-randomized studies can complement RCTs should further advance our understanding of how to effectuate evidence-based medicine through integration of all available sources of information.

1.3.2 Propensity Score Matching for Observational Studies

Since treatment selection is often influenced by clinical factors, the strongest argument against using OSs for causal inferences is the potential for lurking confounders. These lurking covariates may not be equally distributed among treatment cohorts, and thereby influence the outcome variable independently of treatment. For example, in mediation analysis, the confounders may have independent effects, or may mediate the effect of treatment [29].

To adjust for confounders in non-randomized studies, many methods have been developed. Applied researchers have historically relied on regression methods to adjust for differences in baseline characteristics among groups, but this fails to account for selection bias if the covariates are associated with treatment assignment. More recently,

propensity score (PS) analysis has emerged as a widely used method to reduce the effects of confounding, and enable improved treatment effect estimates in OSs. This approach, introduced by Rosenbaum and Rubin [30], uses the estimated conditional probability of being assigned to a certain treatment given a group of observed covariates, under the key assumption of no unmeasured confounders.

To understand how randomized trials and PS analysis using OS compare for causal inference, we first need to consider attributes of the design and model that determine causal effects. Using the notation in Rosenbaum and Rubin [30], suppose we have subjects $i = 1, \cdots, n$ and z_i is the potential treatment assignment for the i^{th} subject, where $z_i = 0$ denotes the control group and $z_i = 1$ the active treatment group. In the two-sample setting, we will have two potential outcomes (patient endpoints) $r(z_i)$ for the i^{th} subject: $r(0)$ under the control regimen and $r(1)$ under active treatment, though usually only one of them is observed. The one not observed is called the *counterfactual alternative* for the observed outcome. For each subject, the treatment effect then is defined as $r(1) - r(0)$ [30, 31], whence the *average treatment effect* (ATE) in the population is defined as $E[r(1) - r(0)]$. A related treatment effect measurement is *average treatment effect for the treated* (ATT) [31], which is defined as $E[r(1) - r(0)|Z = 1]$.

In RCTs, since treatments are assigned at random, we assume that, on average, treated subjects have similar characteristics to those in the competing study arms. As such, these two measures of treatment effects should coincide, and enable valid estimation of the causal relation between treatment and outcome. In OSs, however, usually it is necessary to assume that $E[r(1)] \neq E[r(1)|Z = 1]$ (and similarly for the control group) due to selection bias. In such cases, given covariates X, we can compute the PS as $e(X) = P(Z = 1|X)$. Rosenbaum and Rubin [30] proved that $X \perp Z|e(X)$ (where "$\perp$" denotes "is independent of"), and that $[r(1) - r(0)] \perp Z|e(X)$ given the assumption of "strongly ignorable treatment assignment" (i.e., $[r(1) - r(0)] \perp Z|X$).

This implies that for subjects with roughly the same propensity score, treatments may be viewed as being randomly assigned, yielding unbiased estimates of the ATE [32]. Therefore, the causal relationship can be investigated in OSs with the critical assumption that there are no unknown confounders.

There are many methods for PS estimation, the most common of which regresses treatment assignment on baseline covariates through logistic regression [33, 34, 35]. Adjustment using these estimated propensity scores is then accomplished using one or a combination of four main methods: stratification [36], matching [37], inverse probability weighting (IPW) [38], or regression (covariate) adjustment [32]. Previous research has suggested that matching on PSs achieves a greater balance in measured characteristics between treated and untreated patients compared to stratification [39], and in some settings, PS matching removes modestly more imbalance than IPW [32]. Therefore, PS matching will be used in this thesis, but other approaches could also be used to effectuate PS adjustment among the non-randomized patients. Rosenbaum and Rubin [37] demonstrate three techniques using PSs to construct matched subjects: (i) nearest available matching on the estimated propensity score, (ii) Mahalanobis metric matching including the propensity score, and (iii) nearest available Mahalanobis metric matching within calipers defined by the propensity score. Additional algorithms were compared by [40] via simulation.

Given PS-matched non-randomized data as the first step of our approach, in Chapter 4 we discuss how to combine RCT data and PS-matched non-randomized data in more detail.

1.4 Overview

In the following chapters, we apply Bayesian hierarchical modeling in different areas of research syntheses. Chapter 2 introduces our work on inconsistency detection in

NMA using an arm-based approach, followed by the corresponding diagnostics to detect influential observations or outliers in a NMA in Chapter 3. We propose an approach to detecting influential cases that is simpler and less time-consuming than simply refitting the models by rerunning the MCMC algorithm. Chapter 4 then describes our method to combine RCT and non-randomized data using a motivating HIV/AIDS dataset and via simulation. Our statistical approach throughout is primarily Bayesian, but this is mainly for analytic flexibility and convenience, and these methods could be implemented in a frequentist framework if desired. Finally, Chapter 5 closes this dissertation with a summary and gives several possible future research directions.

Chapter 2

Hierarchical Bayesian Approaches for Detecting Inconsistency in Network Meta-Analysis

In this chapter, we propose inconsistency detection methods using the AB model in NMA. In Section 2.1, we describe the motivating dataset used by Lu and Ades [8] and also throughout our analysis. Details of current models for NMA are illustrated and methods to detect inconsistency investigated in Section 2.2, including our novel discrepancy factor approach using AB models. This section also applies our methods to the example dataset, and compares our results to those of Lu and Ades [8]. Section 2.3 then evaluates our methods via simulation. Finally, in Section 2.4 we summarize and discuss possible limitations of our method.

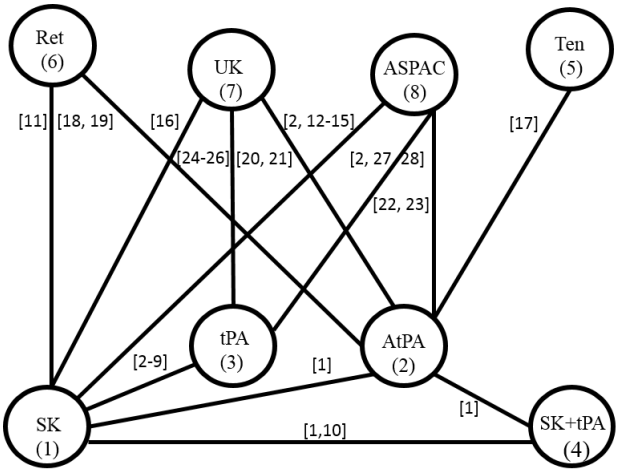

Figure 2.1: Network for thrombolytic drugs dataset. Each vertex represents a treatment and each edge represents a pair of treatments for which at least one direct comparison exists. The indices of the trials having each comparison are shown in square brackets to the left of or above the corresponding edge.

2.1 Motivating Example: Thrombolytic Drugs Dataset

Our illustrative dataset is described in a systematic review [11], and compares eight thrombolytic drugs for use after acute myocardial infarction with the primary outcome being 30-35 day mortality. Twenty-eight trials were conducted to study eight drugs: reteplase (Ret), streptokinase (SK), urokinase (UK), alteplase (tPA), anistreptilase (ASPAC), accelerated alteplase (AtPA), tenecteplase (Ten), and streptokinase plus alteplase (SK + tPA). The dataset is shown in Table 2.1, which displays total number of events over total number of subjects for the treatment groups in each trial. The evidence network is plotted in Figure 2.1, with each vertex representing a treatment and each edge representing a pair of treatments for which at least one direct comparison exists. The indices of the trials having each pairwise comparison are shown in square brackets to the left of or above the corresponding edge.

Table 2.1: Thrombolytic drugs dataset (total number of events/ total number of subjects).

Study number	SK(1)	AtPA(2)	t –PA(3)	SK +tPA(4)	Ten(5)	Ret(6)	UK(7)	ASPAC(8)
1	1472/20163	652/10344		723/10328				
2	1455/13780		1418/13746					1448/13773
3	9/130		6/123					
4	5/63		2/59					
5	3/65		3/64					
6	887/10396		929/10372					
7	7/85		4/86					
8	12/147		7/143					
9	10/135		5/135					
10	4/107			6/109				
11	285/2992					270/2994		
12	3/58							2/52
13	3/86							6/89
14	3/58							2/58
15	13/182							11/188
16	10/203						7/198	
17		522/8488			523/8461			
18		356/4921				757/10138		
19		13/155				7/169		
20		2/26					7/54	
21		12/268					16/350	
22		5/210						17/211
23		3/138						13/147
24			8/132				4/66	
25			10/164				6/166	
26			6/124				5/121	
27			13/164					10/161
28			7/93					5/90

The Appendix offers analysis of another NMA dataset from the literature, the smoking cessation dataset [12] analyzed by Lu and Ades [8] among others. Like Lu and Ades [8], our results also suggest absence of serious inconsistency in this dataset, so we omit these results from this chapter.

2.2 Current Models for Inconsistency Detection in NMA

In this section, we will review various models for NMA, and investigate the associated methods to detect inconsistency in the network. We assume the outcome y_{ik} for each

study follows a binomial distribution as in our sample datasets,

$$y_{ik} \overset{ind}{\sim} Bin(n_{ik}, p_{ik}),\ i = 1, \ldots, I,\ k = 1, \ldots, K, \tag{2.1}$$

where y_{ik} is the total number of subjects with events, n_{ik} is the total number of subjects, and p_{ik} is the probability of the outcome in the k^{th} treatment arm from the i^{th} study.

2.2.1 Contrast-Based Random Effects (CBRE) Models

The NMA logistic model proposed by Lu and Ades [8] is introduced for a single outcome. Using a contrast-based model, we wish to estimate relative effects of treatment pairs. One can use random effects to capture heterogeneity between studies, namely

$$logit(p_{ik}) = \alpha_{iB} + \delta_{iBk}, \tag{2.2}$$

where B refers to study i's baseline treatment. Unless there is treatment common to all studies, the baseline treatments in the studies will be different. Here, α_{iB} is the log odds of the response for the baseline treatment in study i, and δ_{iBk} is the log odds ratio of treatment k versus baseline for the i^{th} study. An independent normal specification for the random effects δ_{iBk} is assumed:

$$\delta_{iBk} \overset{ind}{\sim} N(d_k - d_B, \sigma^2), \tag{2.3}$$

where δ_{iBk} follows a normal distribution with mean equal to the contrast $d_k - d_B$, and δ_{iBk} varies across i, capturing the variability in the log odds ratio of contrast for different studies. This model assumes exchangeability of these differences in d_k on the specific scale that was chosen for the meta-analysis, in this case the log-odds ratio scale which is usually chosen by convention. From the posterior distribution of the d_k, the relative effects of each treatment can be calculated. In this random effects model, the same variance σ^2 is assumed for all treatment groups, so it is called a *homogeneous random*

effects model. If a trial has more than two arms, we need to assume a particular variance-covariance structure for the vector $\vec{\delta_i} = (\delta_{iB2}, ..., \delta_{iBK})'$. The vector $\vec{\delta_i}$ then follows a multivariate normal distribution, with the correlation between any two treatment effects equal to 0.5 under consistency by construction [8]. If we change the variance σ^2 in (2.3) to σ^2_{Bk}, we obtain a *heterogeneous random effects* model, associating different variances with specific pairwise comparisons. As mentioned by Lu and Ades [41], estimating the parameters in the heterogeneity model is a difficult problem due to the implicit constraints on the variances from the NMA structure, which becomes quite complicated when the NMA includes many multi-arm trials. Therefore, we focus on the standard CBRE model with homogeneous random effects.

2.2.2 Arm-Based Random Effects (ABRE) Model

The term "arm-based" is used here to refer to a model proposed by Hong et al. [15] and Zhang et al. [16], although the term has been used somewhat differently elsewhere [42, 43]. In this model, instead of estimating a mean contrast in each trial, we estimate the logit response probability for each treatment for binary data:

$$logit(p_{ik}) = \mu_k + \eta_{ik}, \tag{2.4}$$

where μ_k is the (fixed-effect) mean outcome for treatment k, and η_{ik} is the random effect for treatment k in study i. Then the random effects $\vec{\eta_i}$ for study i are modeled as:

$$\vec{\eta_i} = (\eta_{i1}, \dots, \eta_{iK})' \sim MVN(\mathbf{0}, \Sigma), \tag{2.5}$$

where Σ is an $K \times K$ unstructured covariance matrix to allow correlation between treatment arms in each trial. Compared to the CB model framework, AB models are more straightforward to interpret, especially when implemented in a missing-data framework that imputes values for any treatment arms missing in a given study, thus allowing use

of a common baseline across all trials [15]. However, ABRE models do have slightly more parameters to estimate, since they model the absolute effect of each treatment, rather than relative effects as in CBRE models. As noted, these two approaches make different assumptions of exchangeability: the AB model assumes trials are exchangeable according to the levels of treatment outcomes, while CB methods assume that trials are exchangeable according to treatment contrasts as noted above.

2.2.3 Priors for CBRE and ABRE Models and Model Selection

An important issue in Bayesian modeling is the choice of prior distributions for each of the model parameters. This could be a traditional informative prior, which might come from a literature review or explicitly from an earlier data analysis. But in the situation where there is no previous information about the parameters, we often choose proper, weakly informative prior densities, and let the data drive the posterior distribution [44].

In this chapter, to keep our modeling somewhat generic, we use weakly informative priors for both CBRE and ABRE models. Such priors for CBRE models are described in Lu and Ades [8]. For all our models, the fixed effects (α_{iB}, d_k, and μ_k) assumed to follow a $N(0, 1000)$ distribution, which is very vague but proper. We also assume all CBRE ICFs (the ws) , when present, independently follow a $N(0, \sigma_w^2)$ distribution. In the homogeneous random effects model, we adopt uniform priors $\sigma \sim U(0, 2)$ and $\sigma_w \sim U(0, 2)$ for the standard deviations of the random effects and the ICFs, respectively. For the precision matrix of the random effects in the ABRE model, we choose a Wishart prior $\Sigma_k^{-1} \sim W(V, n)$, where the degrees of freedom $n = K$, the number of treatments, and V is a $K \times K$ matrix with diagonal elements equal to 0.1 and off-diagonal elements equal to 0.005. Using the R function rWishart(), we can calculate that the aforementioned prior corresponds to a 95% prior credible set of $0.14 - 11.52$ for the standard deviation parameters and a 95% credible set of -1.00 to 1.00 for the

correlation parameters, confirming it is weakly informative.

For Bayesian model selection and comparison, we use the Deviance Information Criterion (DIC). DIC is a Bayesian generalization of the Akaike information criterion (AIC), and is calculated as the sum of p_D and $\bar{D}$, where $\bar{D}$ is a measure of goodness of fit and p_D is the effective number of parameters in the model [45]. A model with smaller DIC (say, by at least 5-10 units) is preferred.

2.2.4 Software

`WinBUGS` was used to obtain MCMC samples for all our Bayesian NMA models. Standard diagnostics, including trace plots and sample autocorrelations, were implemented to check MCMC convergence. Lu and Ades [8] provide `WinBUGS` code for CBRE models for the aforementioned two examples. `WinBUGS` code for our ABRE models are available at `http://www.biostat.umn.edu/~brad/software.html`. In all our MCMC runs, we used trace plots and histories of multiple chains to check MCMC convergence, and checked to make sure lag 1 autocorrelations were not excessive. We also used sufficient MC samples to ensure all MCMC errors were small relative to the parameter's standard deviation (maximum around 5% of its standard deviation).

2.2.5 Inconsistency Detection in CB models

In Section 1.1, we introduced loop-based inconsistency method proposed by Lu and Ades [8]. In this chapter, we choose this method as our comparator since it is well-established and has been used as the comparator by many other authors in this field ([9, 10]). In this approach, the number of *inconsistency degrees of freedom* (ICDF) must be determined. ICDF is informally defined as "the number of independent 'loops' of evidence" in the network [18]. When there is no multi-arm trial in the network or when each pair of arms compared in a multi-arm trial is also compared in another trial, ICDF

is calculated as $T - K + 1$, where T is the total number of direct pairwise comparisons and K is the number of treatments. Using the network graph in Figure (2.2) as an example, $T = 6$ and $K = 4$, so there are $ICDF = T - K + 1 = 3$ independent "loops" of evidence for estimating inconsistency in the dataset. To model this inconsistency, (1.1) is modified by adding an inconsistency factor w for each loop:

$$\begin{aligned} d_{BC} &= d_{AC} - d_{AB} + w_{ABC}, \\ d_{BD} &= d_{AD} - d_{AB} + w_{ABD}, \\ \text{and} \quad d_{CD} &= d_{AD} - d_{AC} + w_{ACD}, \end{aligned} \tag{2.6}$$

where the posterior distributions of the 3 added w parameters measure the effect of inconsistency in the 3 respective loops. However, it is not clear how big the inconsistency factor should be for the network to qualify as "inconsistent", instead, these authors recommend use of the posterior probability $\Pr(\sigma_w^2 > \sigma^2 | D)$ as an approximate summary to signal potential evidence inconsistency if this value is high. Given the observation that data tend not provide much information about random-effect variances, one might hypothesize that this test will lack power. For an example discussed in more detail below, using the thrombolytic drugs dataset, although the point estimate for one of the w factors is fairly different from 0 (0.678), the 95% posterior credible interval (CI) for this w factor still covers 0. Moreover, it is not clear which three independent loops from the four that are possible (ABC, ABD, ACD and BCD) should be selected in (2.6). With a different parameterization (i.e., selecting a different 3 set of loops), the estimates of the relative effects using CB models would be different.

A further complication is that when some treatment arms are involved only in multi-arm trials, the aforementioned ICDF formula needs to be reduced by S, which is the number of inconsistency loops where direct comparisons of pairs of arms are only present in multi-arm trials, not in another trial [8]. As shown in Table 2.1, there are $T = 13$ direct comparisons in the thrombolytic drugs dataset: 1-2, 1-3, 1-4, 1-6, 1-7, 1-8, 2-4,

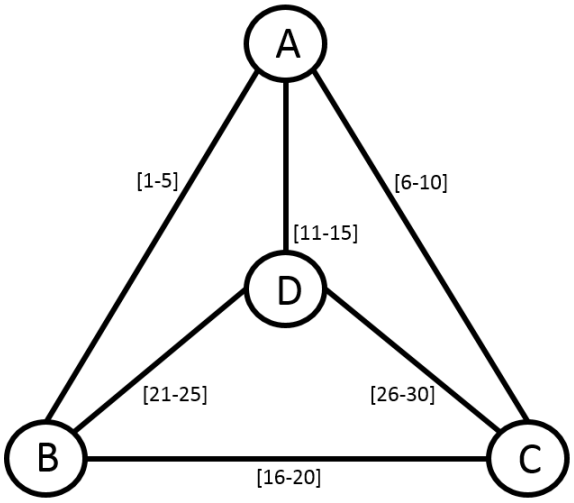

Figure 2.2: Network Graph for Simulation (Treatments A-D are referred to as Treatment 1-4 in Section 2.3 for simplicity). Each vertex represents a treatment and the indices of trials having each comparison are shown in square brackets on the corresponding edge.

2-5, 2-6, 2-7, 2-8, 3-7, and 3-8. Since the comparisons of 1-2 and 2-4 are only estimated in the multi-arm Trial 1, the inconsistency relation for loop 124 cannot be estimated; there is no indirect evidence regarding this loop. Therefore, in this dataset $S = 1$ and $ICDF = T - K + 1 - S = 13 - 8 + 1 - 1 = 5$. A valid set of inconsistency equations modifying model (1.1) is thus

$$\begin{aligned} d_{62} &= d_{61} - d_{21} + w_{126}, \\ d_{72} &= d_{71} - d_{21} + w_{127}, \\ d_{82} &= d_{81} - d_{21} + w_{128}, \\ d_{73} &= d_{71} - d_{31} + w_{137}, \\ \text{and} \quad d_{83} &= d_{81} - d_{31} + w_{138}. \end{aligned} \tag{2.7}$$

Here, the parameterization happens to be unique since there are only five independent loops. However, as we saw above, this will not happen in general. Also, when many multi-arm trials are presented in a NMA, it may become difficult to calculate S, whence

there will be no general formula to determine ICDF [8].

The non-unique parameterization problem of the loop-based inconsistency method has been addressed by Higgins et al. [10]. They find that loop-based inconsistency is a restricted version of their full design-by-treatment interaction model, in which several distinct Lu-Ades models may be nested. However, as Jackson et al. [14] pointed out, if the inconsistency is genuinely due to loop inconsistency, then using the full design-by-treatment interaction model incurs a loss of power, or possibly issues with over-parameterization. More recently, Jackson et al. [46] support that the design-by-treatment interaction model is a unifying framework for modeling loop inconsistency in NMA. Therefore, although our primary goal here is to present an inconsistency detection method using AB model compared to the loop-based method [8], a statistical method for inconsistency detection using AB models compared to the design-by-treatment interaction model is important but awaits further development.

2.2.6 Inconsistency Detection in AB Models

Since AB models implicitly assume consistency, instead of using loop-based w's, we propose using the ABRE fixed and random effects to detect inconsistency. For the thrombolytic drugs dataset, Lu and Ades [8] argue that the posterior distribution of w_{128} indicates evidence for a potential inconsistency problem in loop 128 associated with Trials 22 and 23. Therefore, we use this dataset to show how one might detect inconsistency in AB models.

Inconsistency Detection in AB Models using Fixed Effects

To measure the discrepancy of direct and indirect evidence for comparing two treatments (say, A vs. B) in an arm-based model, we can divide the trials into 4 groups: (i) trials that include both A and B, (ii) trials that include A but not B, (iii) trials that include B

but not A, and (iv) trials that include neither A nor B. Following (2.4), the fixed effects estimating the log odds of an event under treatments A and B in the first group can be denoted as $\mu_A^{(i)}$ and $\mu_B^{(i)}$, respectively. Similarly, the fixed effects estimating the log odds of an event under treatment A in the second group and treatment B in the third group can be denoted as $\mu_A^{(ii)}$ and $\mu_B^{(iii)}$ respectively. The discrepancy between A and B using arm-based models can then be tested by computing the posterior distribution of the *discrepancy factor*

$$\Delta_{AB} = (\mu_A^{(i)} - \mu_B^{(i)}) - (\mu_A^{(ii)} - \mu_B^{(iii)}), \tag{2.8}$$

which is the difference in treatment effects in trials including both arms (the direct evidence) minus the difference in trials including just one arm (the indirect evidence). If zero is in the far tail of this posterior distribution, we conclude that the two sources of evidence for comparing A and B are discrepant, and thus inconsistency exists. This is reminiscent of the CB-based method, node-splitting [9]: we take one pair of treatments at a time and estimate the discrepancy factor for each comparison in separate models. In contrast with the loop-based method, this method allows us to check the inconsistency of the direct and indirect evidence from all sources associated with these two treatments. This is because group (ii) above includes trials with Arm A plus other arms that are never paired with B in any trial, and group (iii) includes trials with Arm B plus other arms that are never paired with A in any trial. Here, we do not use group (iv) in (2.8) since it does not involve either A or B; estimates for comparison AB are primarily shaped by the model and prior from other parts of the dataset. In effect, we effectively assume group (iv) information is broadly consistent with the information from the other three groups in the absence of an excessively strong model or prior. However, if one prefers to include group (iv), the analysis could be redone by combining the last 3 groups as group (II) and modify (2.8) as $\Delta_{AB} = (\mu_A^{(i)} - \mu_B^{(i)}) - (\mu_A^{(II)} - \mu_B^{(II)})$. This alternative approach is discussed briefly in Section 2.4 for the example dataset.

Table 2.2: Discrepancy factors for thrombolytic drugs dataset using ABRE model. "AB model w/o loop" refers to the inconsistency detection method without defining loops in the NMA and "AB model with loop" refers to the loop-based inconsistency detection method.

method	comparison	discrepancy factor	mean	sd	MC error	2.50%	median	97.50%
AB model w/o loop	2 vs. 8	Δ_{28}	-1.37	0.57	0.02	-2.51	-1.36	-0.26
AB model with loop	loop128: 2 vs. 8	Δ_{28}	-1.31	0.60	0.02	-2.54	-1.29	-0.19

We applied the aforementioned method using (2.8) to calculate the discrepancy factors for 11 different comparisons in the thrombolytic drugs dataset: 1-2, 1-3, 1-6, 1-7, 1-8, 2-4, 2-6, 2-7, 2-8, 3-7, and 3-8. Comparison of 1 vs. 4 was not done since the indirect information on 2-4 is only present in the multi-arm Trial 1 (group (i)); therefore, there is no information for group (iii) (include Treatment 4 but not Treatment 1) here. Similarly, comparison of 2 vs. 5 was not performed since there is no information for group (iii) (include Treatment 5 but not Treatment 2). The first row of Table 2.2 ("AB model w/o loop") summarizes the result for the discrepancy factor; its 95% posterior CI does not include zero (informally, it differs significantly from zero). In this approach, without defining loops, we have successfully detected the discrepancy of sources of evidence for comparing 2 vs. 8, which agrees with the conclusion of Lu and Ades. These authors found a posterior estimate for w_{128} of 0.678 with 95% posterior CI $(-0.03, 1.72)$, indicating the presence of inconsistency in loop 128. A similar result was found by White et al. [13], also suggesting inconsistency around loop 128 from certain *designs* (referring to the set of treatments compared in a trial) using a frequentist method. The "AB model with loop" row will be discussed in the next Section.

Loop-based Inconsistency Detection in AB Models using Fixed Effects

In this section, we show how our methods for detecting inconsistency using AB models are related to the loop-based method. Although we no longer use inconsistency factors to study loops in our AB models due to the implicit consistency assumption, we can investigate inconsistency in a loop-based manner within the AB approach by defining discrepancy factors using a different subsetting method for groups.

To directly compare the AB and CB loop-based methods to detect inconsistency, we can redefine the AB method's 4 subgroups in (2.8), by defining groups corresponding to specific loops. For example, to detect the discrepancy of sources of evidence for comparing A vs. B in loop ABC, we can divide the trials into 4 groups: (i) trials that include both A and B, (ii) trials that include A and C but not B, (iii) trials that include B and C but not A, and (iv) other trials. Then the posterior distribution of $\Delta_{AB} = (\mu_A^{(i)} - \mu_B^{(i)}) - (\mu_A^{(ii)} - \mu_B^{(iii)})$ can be used to detect the discrepancy between the direct and indirect evidence for comparing A vs. B in loop ABC. This is a more direct analogue to Lu and Ades' method in the AB model framework. However, this throws away useful information because groups (ii) and (iii) only include the trials within a specific loop, and exclude other trials involving only one of arms A and B. We applied this method to the thrombolytic drugs dataset, computing discrepancy factors for 15 different pairwise comparisons, for loop 126, 127, 128, 137, and 138, with 3 comparisons in each loop: 1-2, 1-6, 2-6 for loop 126; 1-2, 1-7, 2-7 for loop 127, etc. As shown in the bottom row of Table 2.2 ("AB model with loop"), zero is not contained in the 95% posterior CI for the discrepancy factor for comparison 2-8 in loop 128, indicating that inconsistency exists in loop 128 using loop-based AB model method.

Table 2.3: The most extreme random effects for thrombolytic drugs dataset using ABRE model (i refers to study and k refers to treatment arm).

η[i,k]	η[2,3]	**η[2,8]**	η[2,1]	η[6,3]	**η[22,2]**	η[11,1]	η[20,7]	**η[23,2]**	**η[18,2]**	η[11,6]	η[19,6]
Posterior Mean	0.51	**0.45**	0.39	0.35	**-0.29**	0.26	0.25	**-0.25**	**0.22**	0.21	-0.19

Inconsistency Detection in AB Models using Random Effects

One disadvantage of detecting inconsistency in AB models through fixed effects (Section 2.2.6) is that although it detects inconsistent comparisons, it does not identify the source of the inconsistency as arising from certain trial and treatment combinations. Without defining specific w factors in our AB models, effects of inconsistency, if any, must manifest in either the fixed or the random effects in the ABRE model. Therefore, after detecting inconsistency with fixed effects, we can use the random effects η_{ik} in our ABRE model to investigate the most extreme η_{ik}, say, the top 5% in absolute value. Using the thrombolytic drugs dataset, the posterior means of these random effects are shown in Table 2.3: the random effects for Treatments 3, 8 and 1 in Trial 2 have the estimated values 0.51, 0.45 and 0.39, respectively, followed by Treatment 3 in Trial 6 and Treatment 2 in Trial 22, where Trials 2 and 6 have large sample sizes. The remaining random effects deemed to be large are for Treatments 1 and 6 in Trial 11, Treatment 7 in Trial 20, Treatment 2 in Trials 23 and 18, and Treatment 6 in Trial 19. Here, we recommend that these sources of inconsistency detected by the random effects first be tested using the discrepancy factor method described in Section 2.2.6. In our example, Trials 22 and 23 consider the direct comparison of 2 vs. 8, and Trials 2 and 18 consider the indirect comparison of 2 vs. 8, which is discrepant as shown in Table 2.2. Therefore, Treatment 2 in Trials 22 and 23, Treatment 8 in Trial 2 and Treatment 2 in Trial 18 can be identified as potential sources of the inconsistency. This agrees with the findings in Lu and Ades [8] as they detected Treatment 2 in Trials 22 and 23 as the sources of

Table 2.4: Model comparisons with DIC for thrombolytic drugs dataset.

	CB random effect models			AB random effect models		
	DIC	p_D	$\bar{D}$	DIC	p_D	$\bar{D}$
Full data	93.29	40.83	52.46	79.65	27.03	52.62
Trial 22 and 23 deleted	83.38	36.01	47.37	70.37	26.03	44.34

inconsistency as well.

To see the effects of Trials 22 and 23, we fit the same ABRE model without them. The results for CBRE and ABRE models are compared using DIC, for the datasets with and without Trials 22 and 23. As shown in Table 2.4, the ABRE models have smaller DIC compared to CBRE models whether or not these two trials are included, indicating that the model-based imputation of the unobserved arms yields better DIC performance than ignoring such information for these data. The ABRE and CBRE models have similar $\bar{D}$, indicating similar goodness of fit, so the reduction in DIC for the ABRE model is mainly due to the reduction in p_D, i.e., the ABRE has a smaller effective number of parameters.

2.3 Simulation Studies

In this section, we generate artificial data from inconsistency structures in datasets from CBRE models to evaluate the performance of inconsistency detection using our ABRE models, and compare the results to those of the loop-based method of Lu and Ades [8].

2.3.1 Simulation Settings

We simulate data for a network meta-analysis from the CBRE model for 30 trials (each with only 2 treatment arms) to compare 4 treatments using binary outcomes as in Figure 2.2. In the CBRE models, the true values of $\alpha_{iB}, i = 1, ..., 30$ were assigned as

a sequence of length 30 from -2 to -3. True values of d_{12}, d_{13} and d_{14} were chosen as 0.5, 0.9 and 1.2 respectively, with $w_{123} = w_{124} = 0.01$ and $w_{134} = 2.5$ (what we call the *alternative scenario*, in contrast to a *null scenario* in which all w = 0). Using (2.6), the true values of d_{23}, d_{24} and d_{34} were calculated as 0.41, 0.71, and 2.8, respectively. The standard deviation σ in (2.3) was set equal to 2, and we set $n_{ik} = 100$ for the k^{th} treatment arm in the i^{th} study (Results for simulations using unequal sample sizes and asymmetric network structure are shown in the Appendix). Artificial data y_{ik} can be generated according to (2.1) and (2.2). From this data generation scheme, the evidence in loop 134 is inconsistent, especially in the comparison of 3 vs. 4. This can be confirmed by comparing the randomly generated datasets using the aforementioned settings to consistent datasets generated assuming all $w = 0$ but otherwise using the aforementioned specification (the null scenario). For the experimental design, since there are 6 pairwise comparisons for these 4 treatments (1-2, 1-3, 1-4, 2-3, 2-4, and 3-4), we assigned 5 trials to each comparison, 30 trials in total, each with two arms. We generated 1000 simulated datasets for each scenario using the `Brugs` package in `R`, where we call `OpenBUGS` [47] from `R`, once for each simulated dataset. Each simulated dataset took less than 30 seconds to run on a Dell Latitude E7440 Laptop with 4th gen Intel Core i5-4310U Processor (2.0GHz), with MCMC convergence checking performed as described in Section 2.2.4 for a 1% subsample of simulated datasets.

2.3.2 Simulation Results

In this simulation, our hypothesis is that when analyzing data simulated from the alternative scenario based on CB models, we are still able to detect inconsistency using the fixed effects in AB models for the comparison 3 vs. 4 as stated above. We also perform the analysis using datasets from the null scenario, which should show no evidence of

Table 2.5: Discrepancy factors for simulated datasets using ABRE model. Each entry represents the summary of results from 1000 datasets under each scenario.

w	discrepancy factor	mean	2.50%	97.50%	95% BCI width	Power
w[123]=0.01, w[124]=0.01, w[134]=2.5	Δ_{12}	-1.02	-3.21	1.18	4.39	0.11
	Δ_{13}	-1.25	-3.51	0.99	4.50	0.19
	Δ_{14}	-0.13	-2.42	2.17	4.59	0.05
	Δ_{23}	-0.61	-3.00	1.80	4.80	0.08
	Δ_{24}	0.55	-1.78	2.91	4.69	0.08
	Δ_{34}	**-2.90**	**-5.20**	**-0.61**	**4.59**	**0.66**
all w=0	Δ_{12}	-1.02	-3.19	1.17	4.36	0.11
	Δ_{13}	-1.27	-3.50	0.96	4.46	0.20
	Δ_{14}	-1.33	-3.55	0.89	4.44	0.21
	Δ_{23}	-0.60	-2.97	1.79	4.76	0.07
	Δ_{24}	-0.61	-2.95	1.73	4.68	0.08
	Δ_{34}	-0.50	-2.81	1.84	4.65	0.12

inconsistency using the same method (Section 2.2.6). If the discrepancy factors indicate inconsistency in the dataset, we further investigate the sources of inconsistency by considering the estimated random effects in the AB models. To check the accuracy of the data generation, we first use CB models to fit the datasets from the alternative scenario, and evaluate the coverage probabilities of 95% posterior CIs for these true values. The coverage probabilities of the w factors (w_{123}, w_{124} and w_{134}) and the relative treatment effects (d_{12}, d_{13} and d_{14}) are (0.927, 0.956, 0.935) and (0.936, 0.933 and 0.904), respectively, indicating good data generation and not very influential prior distributions.

Our proposed methods using AB models (as in Section 2.2.6) were then applied. Table 2.5 displays the 2.5% and 97.5% posterior quantiles, and 95% Bayesian CI (BCI) widths for the discrepancy factors for these 6 pairwise comparisons, where each summary is the average of its quantity over the 1000 simulated datasets. The average of the posterior means for Δ_{34} in the simulation for the alternative scenario is -2.90, with average 95% BCI $(-5.20, -0.61)$. The mean estimate for Δ_{34} (comparing Arms 3 and 4)

is very different from 0, so we conclude that the two sources of evidence in that comparison are discrepant, and thus inconsistency exists. For each dataset, we also calculated whether the 95% BCI excluded zero, and counted the fraction for which this was true out of 1000 simulations ("Power"). For comparison 3 vs. 4, this estimates the power under the alternative scenario (reject the null $\Delta_{34} = 0$ when the alternative is true), which emerges as 0.66. Here, since we generated the datasets using CB models with alternative hypothesis $w_{134} = 2.5$, we assume $\Delta_{34} \neq 0$ under this condition. However, it also suggests a small inconsistency in comparisons 1 vs. 4 and 1 vs. 3. On the other hand, all the mean estimates of the discrepancy factors are not that different from zero when we set all $w = 0$, indicating no evidence of inconsistency for null datasets using our AB inconsistency detection approach. When all the w = 0 (the null scenario), the estimate of "Power" for comparison 3 vs. 4 now is 0.12, giving us an estimate of Type I error under the null scenario (reject the null $\Delta_{34} = 0$ when the null is true). All other "Power" estimates for the other comparisons under either alternative or null hypothesis are low, showing that they are essentially Type I errors (null hypothesis that all other Δs are 0 from our simulation).

The sources of inconsistency for the alternative scenario were further investigated with the AB model using random effects. As shown in Figure 2.3, we have examined the most extreme elements (the top 5% in absolute value) at specific trial-arm levels using the posterior mean of η_{ik} averaged over 1000 simulated datasets: Treatment 4 in Trials 25, 26, 27, 29, and Treatment 3 in Trials 29 and 30. All these six extreme random effects are from either direct of indirect comparison 3 vs. 4, suggesting that both inconsistency detection methods using AB models work well on the simulated datasets. We have also checked the other extreme observations in Figure 2.3, and found most of them to be from the 3 vs. 4 comparison as well.

We have also detected inconsistency using CB models as described in Lu and Ades [8]

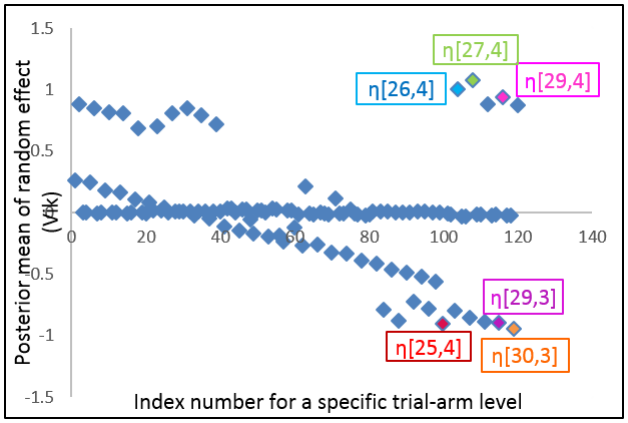

Figure 2.3: Simulated datasets under alternative scenario: the most inconsistent trial by treatment combinations using AB model are shown in rectangle by flagging the most extreme random effects.

for comparison. CBRE models with ICFs were applied to the simulated alternative-scenario datasets, and the large value for $w_{134} = 2.31$ suggests inconsistency in loop 134, though it is not significantly different from 0 according to its 95% BCI $(-0.30, 4.96)$. Furthermore, the sources of the inconsistency can be investigated by comparing the mean residual deviance with and without inconsistency factors using CBRE models. Three of these outliers are from loop 134, suggesting slightly weaker inconsistency detection than using our method.

2.4 Discussion

In this chapter, we have proposed methods for detection of inconsistency using an arm-based random effects model for NMA. Our implementation is fully Bayesian, but our AB model can also be analyzed using classical tools in a generalized linear mixed effect model framework (see Section A.4 in the Appendix). Discrepancy factors have previously appeared in the Bayesian literature in the context of checking model adequacy [48, 49], but to our knowledge, this is their first use in NMA.

Our methods can consider loops but do not need to, and they permit users to address issues previously tackled by CB models. Compared to Lu and Ades [8], our approach can examine specific comparisons in detail, not only in a loop-based manner. This shows how sources of direct and indirect information differ (c.f. Dias et al. [9]). More recent research on the flow of direct evidence to the network estimates can be found in Senn et al. [50] and König et al. [51]. After the Lu and Ades [8] approach detects inconsistency (say, for loop 128 for the thrombolytic drugs dataset), it still needs to examine the source of inconsistency by checking each comparison in the loop. Using the fixed effects in ABRE models to check the discrepancy of the different sources of evidence for comparing two treatments seems to us a simpler and more direct approach for inconsistency detection, and one that can be done using either all sources of information, or only the information in a specific loop.

We identified similar sources of inconsistency using ABRE models as using CB methods in both the example and simulated datasets. The significance levels of the inconsistency factors are only summarized in Lu and Ades [8]; they do not specify how big the ICFs should be to declare network inconsistency. Instead, our methods flag discrepancy factors significantly different from 0 based on their 95% posterior CIs, providing more objective evidence of inconsistency.

We further identified the sources of inconsistency by the most extreme fitted random effects at trial levels. Here, we recommend using both of our proposed methods in concert to identify trial-level inconsistency, which cannot be detected by the discrepancy factor method alone. One limitation of using random effects is that the cutoff for selecting the most extreme elements (say, the top 5% in absolute value) is rather arbitrary. But since we detect inconsistency using discrepancy factors first, the precise criterion we use here is less important. In next chapter, we will examine detection of outlying or influential trial-treatment combinations using diagnostics for hierarchical

models introduced by Hodges [21].

There are of course some limitations or potential concerns associated with our proposed NMA approaches. First, a common concern with AB models is that they "break the randomization" by assigning exchangeable absolute treatment effects across trials. Often this criticism refers to the modeling of pooled treatment arms independently instead of jointly, which we also consider inappropriate. We are using a random effects model with an unstructured covariance matrix on the random effect terms, mitigating this concern somewhat since it gives the desired correlations between treatments in a specific trial. Second, there exist other good methods for inconsistency detection, such as the node-splitting method of Dias et al. [9] and the multivariate meta-regression of Higgins et al. [10]. We found that our discrepancy factor model has much in common with Dias' method, though we perform analysis based on AB model framework while their method is based on the CB model (see further discussion in Appendix). Lastly, the node-splitting method compares "direct evidence" on X vs. Y with what would be predicted from all the remaining evidence (not just the trials including X vs. Y). Although we reported only the discrepancy factor results based on the first 3 of the 4 groups described in Section 2.2.6, we also described a modification using the discrepancy factors arising from all 4 groups ("direct" evidence vs. all remaining evidence), which gave similar results in an analysis of our data (results not shown). We acknowledge that network structure and evidence flow, as well as the location of inconsistency sources in the network, might affect whether this alternative implementation of our approach gives different answers from the original version. Also, our approach is primarily a tool for seeking inconsistency, and (like a node-splitting model) not generally appropriate for making inference about pooled effects.

Chapter 3

Diagnostics for Generalized Linear Hierarchical Models in NMA

Diagnostics for linear hierarchical models with additive errors were proposed by Hodges [21]. The key is to express a hierarchical model in the form of a linear model, which is accomplished by adding artificial "cases" to the dataset (Hodges [52], Section 2.1.1, gives a brief history of this technique). Since the error terms of the new model are not homoscedastic, the model is then transformed to achieve equal variances for the error terms, as in ordinary linear models, after which diagnostics are applied.

This chapter uses the example dataset described in Section 2.1 to show how to diagnose influential observations and outliers, with particular interest in data points (study arms) that have previously been detected as sources of inconsistency. We present the AB model first because this method is simpler, and motivates the development for the less straightforward CB model. Section 3.1 proposes our diagnostic methods in NMA using AB models, followed by Section 3.2 using CB models. Section 3.3 summarizes

results for the example dataset, and then we evaluate our detection of influential observations for AB models via simulation in Section 3.4. Finally, Section 3.5 summarizes and gives conclusions.

3.1 Diagnostics for the AB Model

3.1.1 Reformulation Generalized Linear Hierarchical Models as Linear Models

In applying Hodges' method to generalized linear hierarchical models in NMA, one difficulty is that we do not have normal errors because our outcome is binary. One solution is to use a normal approximation to the likelihood [44], which approximates the generalized linear model as a linear model in the following way: for each data point y_i, we construct a pseudodatum $\tilde{y}_i$ and corresponding pseudovariance, and assume $\tilde{y}_i$ has approximately a normal likelihood with that pseudovariance. One way to specify the pseudovariance of the error terms is called "the quasiexact method" [53], which uses the variance of the observable outcome conditional on the mean structure. To approximate our AB model as a linear model this way, we propose to construct new pseudodata as $\tilde{y}_{ik} = logit(\frac{y_{ik}}{n_{ik}})$. The binomial distribution of y_{ik} has mean $n_{ik}p_{ik}$ and variance $n_{ik}p_{ik}(1-p_{ik})$, so by the delta method, $Var(\tilde{y}_{ik}) \approx \frac{1}{n_{ik}p_{ik}(1-p_{ik})}$.

We can now approximate the binary-outcome hierarchical model (2.4)-(2.5) as follows:

$$\begin{aligned} \tilde{y}_{ik} &= \xi_{ik} + \epsilon_{ik}, \\ \xi_{ik} &= \mu_k + \eta_{ik}, \\ \text{and} \quad \mu_k &= M_k + \nu_k, \end{aligned} \tag{3.1}$$

where $\epsilon_{ik} \dot{\sim} N(0, \frac{1}{n_{ik}p_{ik}(1-p_{ik})})$, $\boldsymbol{\eta_i}' = (\eta_{i1}, \ldots, \eta_{iK})' \sim MVN(\mathbf{0}, \Sigma)$ as described in (2.5), and the last equation in (3.1) represents the priors for μ_k, with $\nu_k \overset{ind}{\sim} N(0, 1000)$. Here,

ξ_{ik} is the sum of μ_k and η_{ik}, where μ_k is defined in (2.4). Following Hodges [21], we can move the known terms to the left side of these equations and keep the unknown terms on the right, rewriting the above equations as

$$\begin{aligned} \tilde{y}_{ik} &= \xi_{ik} + \epsilon_{ik}, \\ 0 &= -\xi_{ik} + \mu_k + \eta_{ik}, \\ \text{and} \quad -M_k &= -\mu_k + \nu_k. \end{aligned} \tag{3.2}$$

Equations (3.2) have the form of a linear model:

$$\left(\begin{array}{c} \tilde{\mathbf{y}} \\ \hline \mathbf{0}_N \\ \hline -\mathbf{M} \end{array}\right) = \left(\begin{array}{c|c} I_{N\times N} & 0_{N\times K} \\ \hline -I_{N\times N} & H_{N\times K} \\ \hline 0_{K\times N} & I_{K\times K} \end{array}\right) \left(\begin{array}{c} \xi_{1,1} \\ \vdots \\ \xi_{i,k} \\ \vdots \\ \xi_{I,K} \\ \hline \boldsymbol{\mu} \end{array}\right) + \left(\begin{array}{c} \boldsymbol{\epsilon} \\ \hline \boldsymbol{\eta} \\ \hline \boldsymbol{\nu} \end{array}\right). \tag{3.3}$$

Because we often have missing arms in each of the trials, the possible treatments in a trial (k, the second subscript of ξ) is a subset of all treatments, and the total number of observations is usually much smaller than $I \times K$. Let N be the total number of observations, $\tilde{\mathbf{y}}' = (\tilde{y}_1, \cdots, \tilde{y}_N)'$ be the observed data, $I_{m\times m}$ be an identity matrix of dimensions m, $0_{a\times b}$ be a matrix of 0s with dimension $a \times b$, $\boldsymbol{\mu}' = (\mu_1, \cdots, \mu_K)'$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are vectors with length N, and $\boldsymbol{\nu}' = (\nu_1, \cdots, \nu_K)'$. As in Section 2.2.2, $M_k = 0$ for

all k. For our example dataset,

$$\left(\begin{array}{c} \tilde{\mathbf{y}}_{58} \\ \hline \mathbf{0}_{58} \\ \hline \mathbf{0}_{8} \end{array}\right) = \left(\begin{array}{c|c} I_{58\times 58} & 0_{58\times 8} \\ \hline -I_{58\times 58} & \left(\begin{array}{cccccc} 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{array}\right)_{58\times 8} \\ \hline 0_{8\times 58} & I_{8\times 8} \end{array}\right) \left(\begin{array}{c} \xi_{1,1} \\ \vdots \\ \xi_{28,8} \\ \hline \mu_1 \\ \vdots \\ \mu_8 \end{array}\right) + \left(\begin{array}{c} \epsilon_{1,1} \\ \vdots \\ \epsilon_{28,8} \\ \hline \eta_{1,1} \\ \vdots \\ \eta_{28,8} \\ \hline \nu_1 \\ \vdots \\ \nu_8 \end{array}\right), \tag{3.4}$$

where N is 58 (since we have 28 trials, with the first 2 trials being 3-arm trials and the rest 2-arm trials) and K is 8. To set up the design matrix, the only part that needs special attention is the matrix $H_{58\times 8}$ in (3.3), which is not an identity matrix. In any NMA, no trial includes all treatments (arms) being considered, so H will conform to the pattern of trial-by-arm combinations in the NMA design. For example, in our dataset, the first trial has Arms 1, 2, and 4, so the 1^{st}, 2^{nd}, and 4^{th} column entries for the first three rows in $H_{58\times 8}$ are 1s, as shown in (3.4).

Equation (3.3) can be expressed in more compact notation as

$$Y = X\Theta + E. \tag{3.5}$$

Here Y, the pseudodata, and X, the design matrix, are known, Θ is an unknown parameter vector, and E is an error term with mean zero and block diagonal covariance matrix Γ. The blocks of Γ correspond to the covariance matrices of $\boldsymbol{\epsilon}$, $\boldsymbol{\eta}$, and $\mathbf{v}$, where the upper-left 58×58 block of Γ contains the pseudovariances for the ϵ_{ik}, the next 58 rows and columns are composed of 28 block covariance matrices for the random effects of the 28 trials, and the last 8×8 diagonal block contains the variances of the ν_k (the

prior variances), all set to 1000. The parameter Θ is comprised of $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$, and our interest is in the estimates of the μ_k's.

Rows of X, Y, and E in (3.5) corresponding to $\tilde{\mathbf{y}}$ (and thus the actual data y_{ik}) are called *data cases*: the pseudodata $\tilde{\mathbf{y}}$ enter directly into the joint posterior through these observations. Rows of X, Y, and E corresponding to H are *constraint cases*, which effectively impose constraints on the parameters Θ. Finally, rows of X, Y, and E with variances ν_k are denoted as *prior cases*, generated by the hyperprior. We refer to this format as the *constraint case formulation* of a hierarchical model.

3.1.2 Pre-steps before Diagnostics

In the reformulation (3.4), the variances of the error terms are not constant. To obtain equal variances for the error terms, we need to premultiply each term in (3.5) by $\Gamma^{-\frac{1}{2}}$ to give

$$\boldsymbol{Y} = \boldsymbol{X\Theta} + \boldsymbol{E}, \tag{3.6}$$

where the bold font indicates premultiplication by $\Gamma^{-\frac{1}{2}}$, so the covariance matrix of $\mathbf{E}$ is an identity matrix. Further, define the generalized hat matrix $\mathbf{V} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, so one estimate of the residuals $\mathbf{E}$ is

$$\hat{\mathbf{E}} = (\mathbf{I} - \mathbf{V})\mathbf{Y}. \tag{3.7}$$

3.1.3 Diagnostics for Case Influence

Case influence diagnostics show how estimates of parameters change when cases are deleted. In our example dataset, we want to investigate which trial-by-arm combinations are influential. In a Bayes MCMC framework, the most accurate way to detect influential cases would be to rerun the MCMC algorithm for each deleted case, but this could be a prohibitive computing task. Therefore, we use the approach described above with

the single-parameter analogue to Cook's distance proposed by Hodges [21], henceforth called the *linear approximation method*. Fixing Γ at its posterior mean computed using the full dataset, we approximate the change in $\hat{\mathbf{\Theta}}$ arising by deleting the r^{th} case as

$$\hat{\mathbf{\Theta}}_{(-\mathbf{r})} - \hat{\mathbf{\Theta}} \approx -(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{x}_\mathbf{r}'\hat{\mathbf{E}}_\mathbf{r}/(\mathbf{1} - \mathbf{v_{rr}}), \tag{3.8}$$

where $\mathbf{X}$ is computed using the posterior mean of Γ, $\hat{\mathbf{\Theta}}$ is the posterior mean of $\mathbf{\Theta}$ using the full dataset, and $\hat{\mathbf{\Theta}}_{(-\boldsymbol{r})}$ is the posterior mean of $\mathbf{\Theta}$ after deleting the r^{th} case, $\hat{\mathbf{E}}_\mathbf{r}$ is the r^{th} row of

$$\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\mathbf{\Theta}}, \tag{3.9}$$

where $\mathbf{x_r}$ is the r^{th} row of $\mathbf{X}$, and $\mathbf{v_{rr}}$ is the r^{th} diagonal element of $\mathbf{V}$. The change in the estimate of the i^{th} element of $\hat{\mathbf{\Theta}}$ arising from deleting the r^{th} observation is approximately the i^{th} element in (3.8). To facilitate judgments about the size of these changes, let θ be an element of $\mathbf{\Theta}$, and define the *relative change* arising from deleting the r^{th} observation as

$$\mathbf{RC}\{\theta; \mathbf{r}\} = (\hat{\theta}_{(-\mathbf{r})} - \hat{\theta})/\mathbf{psd}\{\theta|\mathbf{Y}\}, \tag{3.10}$$

where *psd* stands for "posterior standard deviation" computed using the full dataset. Analogously to the standard calibration of Cook's distance, $|\mathbf{RC}| \geq 2$ suggests an influential case. For comparison, we calculate $\mathbf{RC}$'s approximately from the linear approximation method, and also exactly for selected trial-by-arm combinations by rerunning the MCMC algorithm.

3.1.4 Outlier Detection Using Residuals

Residuals are often used to detect outlying observations in linear models. After approximating the generalized linear hierarchical model by a linear model as above, we can obtain residuals using either (3.7) or (3.9). Usually, standardized residuals are used to

investigate outliers. Hodges [21] applied the method proposed by Chaloner [54] to the hierarchical reformulation (3.6) to give:

$$cov(\hat{\mathbf{E}}_{\mathbf{d}}) = E\{cov(\hat{\mathbf{E}}_{\mathbf{d}}|\boldsymbol{\Gamma})\} + cov\{E(\hat{\mathbf{E}}_{\mathbf{d}}|\boldsymbol{\Gamma}) = [E(\mathbf{I}-\mathbf{V})]_{\mathbf{d}}, \tag{3.11}$$

with the subscript $\mathbf{d}$ referring to the data cases. Therefore, we can calculate a standardized residual for case r as $\hat{\mathbf{E}}_{\mathbf{r}}/\mathbf{var}(\hat{\mathbf{E}}_{\mathbf{r}})$. Also, although for ordinary linear models the mean of the residuals is zero by construction, this is not true in general for hierarchical models. As shown in equation (4.18) of Hodges [21], for a fixed Γ, the mean of the data case residuals conditional on $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ is

$$E(\hat{\mathbf{E}}_{\mathbf{d}}|\boldsymbol{\eta},\boldsymbol{\nu}) = (\mathbf{I}-\mathbf{V})\boldsymbol{\Gamma}^{-\frac{1}{2}} \begin{pmatrix} 0 \\ \hline \boldsymbol{\eta} \\ \hline \boldsymbol{\nu} \end{pmatrix}, \tag{3.12}$$

which need not be zero; in other words, these residuals are biased. To correct the bias, we can subtract an estimate of the bias from an estimates of $\hat{\mathbf{E}}$, and then standardize them.

3.2 Diagnostics Using CB Model

We can apply similar diagnostic methods to the CB model. To do so, we need to express the CB model in (2.2) and (2.3) in the form of a linear model again using pseudo outcomes $\tilde{y}_{ik} = logit(\frac{y_{ik}}{n_{ik}})$. Then we can formulate the hierarchical CB model as

$$\begin{aligned} \tilde{y}_{ik} &= \xi_{ik} + \epsilon_{ik}, \\ \xi_{ik} &= \alpha_{iB} + (d_k - d_B) + \delta_{ik}, \\ \alpha_{iB} &= M_{iB} + w_{iB}, \\ \text{and} \quad d_k &= z_k + \phi_k, \end{aligned} \tag{3.13}$$

where $\epsilon_{ik} \dot{\sim} N(0, \frac{1}{n_{ik}p_{ik}(1-p_{ik})})$, $\delta_{ik} \overset{ind}{\sim} N(0, \sigma^2)$, and the latter two equations represent the priors for α_{iB} and d_k, where the w_{iB} and ϕ_k are independently assigned $N(0, 1000)$ priors. Here, ξ_{ik} is the sum of α_{iB} and δ_{iBk} in (2.2). Moving the known terms to the left and keeping the unknown terms on the right of the equation as before, we can rewrite (3.13) as

$$\begin{aligned} \tilde{y}_{ik} &= \xi_{ik} + \epsilon_{ik}, \\ 0 &= -\xi_{ik} + \alpha_{iB} + (d_k - d_B) + \delta_{ik}, \\ -M_{iB} &= -\alpha_{iB} + w_{iB}, \\ \text{and} \quad -z_k &= -d_k + \phi_k, \end{aligned} \tag{3.14}$$

which can be written in the form of a linear model

$$\left(\begin{array}{c} \tilde{\mathbf{y}} \\ \hline \mathbf{0}_N \\ \hline -\mathbf{M} \\ \hline -\mathbf{z} \end{array}\right) = \left(\begin{array}{c|c|c} I_{N\times N} & 0_{N\times I} & 0_{N\times(K-1)} \\ \hline -I_{N\times N} & (H_1)_{N\times I} & (H_2)_{N\times(K-1)} \\ \hline 0_{I\times N} & I_{I\times I} & 0_{I\times(K-1)} \\ \hline 0_{(K-1)\times N} & 0_{(K-1)\times I} & I_{(K-1)\times(K-1)} \end{array}\right) \left(\begin{array}{c} \xi_{1,1} \\ \vdots \\ \xi_{i,k} \\ \vdots \\ \xi_{I,K} \\ \hline \boldsymbol{\alpha} \\ \hline \mathbf{d} \end{array}\right) + \left(\begin{array}{c} \boldsymbol{\epsilon} \\ \hline \boldsymbol{\delta} \\ \hline \mathbf{w} \\ \hline \boldsymbol{\phi} \end{array}\right), \tag{3.15}$$

where $I_{m\times m}$, $0_{a\times b}$, $\tilde{\mathbf{y}}$ and $\boldsymbol{\epsilon}$ are as described in Section 3.1.1. The model includes the nuisance parameter $\boldsymbol{\alpha}' = (\alpha_1, \cdots, \alpha_I)'$, treatment effect $\mathbf{d}' = (d_2, \cdots, d_K)'$ ($d_1 = 0$ for the common reference treatment), random effect vector $\boldsymbol{\delta}$ with length N, and priors "errors" $\mathbf{w}' = (w_1, \cdots, w_I)'$ and $\boldsymbol{\phi}' = (\phi_2, \cdots, \phi_K)'$. Using our real dataset as an example, where the prior means M_{iB} and z_k are 0 for all 28 trials, this becomes

$$\left(\begin{array}{c} \tilde{\mathbf{y}}_{58} \\ \hline \mathbf{0}_{58} \\ \hline \mathbf{0}_{28} \\ \hline \mathbf{0}_{7} \end{array}\right) = \left(\begin{array}{c|c|c} I_{58\times 58} & 0_{58\times 28} & 0_{58\times 7} \\ \hline -I_{58\times 58} & \underset{58\times 28}{\begin{pmatrix} \mathbf{1}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{n_I} & \cdots & \mathbf{1}_{n_I} \end{pmatrix}} & \underset{58\times 7}{\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 1 \end{pmatrix}} \\ \hline 0_{28\times 58} & I_{28\times 28} & 0_{28\times 7} \\ \hline 0_{7\times 58} & 0_{7\times 28} & I_{7\times 7} \end{array}\right) \left(\begin{array}{c} \xi_{1,1} \\ \vdots \\ \xi_{28,8} \\ \hline \alpha_1 \\ \vdots \\ \alpha_{28} \\ \hline d_2 \\ \vdots \\ d_8 \end{array}\right) + \left(\epsilon_{1,1} \cdots \epsilon_{28,8} \middle| \delta_{1,1} \cdots \delta_{28,8} \middle| w_1 \cdots w_{28} \middle| \phi_2 \cdots \phi_8 \right)' \tag{3.16}$$

where n_i is the number of treatments in trial i, so that $\mathbf{1}_{n_i}$ is a column 3-vector of 1s for the first two trials, and a column 2-vector of 1s for all the other trials. The parameter vector Θ is comprised of ξ, α, and d; the d's are the parameters of interest. As before, we need to pay special attention to the block matrix H_2 in (3.15), which represents the coefficients of d_k in $d_k - d_B$, in (3.14). First, for the row in H_2 corresponding to the baseline treatment of a trial, the row entries are all 0s since $d_B - d_B = 0$. Second, for a non-baseline arm in a trial, if the baseline in that trial is Treatment 1, then the row has all 0 entries except for a one for the non-baseline treatment's d_k, because Treatment 1 is always chosen as the common reference ($d_1 = 0$). Third, if the baseline in a trial is not Treatment 1, then in the row of H_2 for a non-baseline arm, the entries are -1 for the

baseline treatment, 1 for the non-baseline arm, and 0 otherwise. For our dataset, the 1st trial has Treatments 1, 2, and 4, therefore the entries in the first row in H_2 are all 0s for baseline Treatment 1, and in the second and third rows the entries are 0 except for 1s in the columns for d_2 and d_4, for Treatments 2 and 4 respectively. For Trial 28, because it has the two Treatments 3 and 8, the entries are -1 and 1 for Treatments 3 and 8 in the last row, respectively, since now the baseline is not Treatment 1. Equation (3.16) can also be written in matrix notation as in (3.5), where now the outcome vector Y has length 151 and the unknown parameter vector Θ has length 93. For the error covariance matrix Γ, the first 58 diagonal elements are the pseudovariances for the ϵ_{ik}, the next 58 rows are composed of 28 block variance-covariance matrices for the random effects in these 28 trials, and the last 35 elements are the variances of the priors for the fixed effects, again all set to 1000.

After reformulation as above, and summarizing (3.16) as $Y = X\Theta + E$, as in (3.5), we again premultiply each term in (3.5) by $\Gamma^{-\frac{1}{2}}$ to get (3.6), so that $cov(\mathbf{E})$ is again an identity matrix. Similarly, we define the hat matrix $\mathbf{V} = \mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}$, and estimate $\mathbf{E}$ using (3.7).

Now that we have reformulated (and approximated) the CB model as a linear model, we can use case-deletion diagnostics to detect influential observations. However, because of the different assumptions made by CB and AB models, the route we take from here must differ from that used for the AB method. We can no longer simply consider deletion of individual data cases in (3.16), as we did for the AB model, because various scenarios can arise when we delete trial-by-arm combinations under the CB model. For example, in a two-arm study, if we delete one arm, only one treatment arm is left. We can keep this single-arm study in the analysis under the AB model, since it still contributes to the likelihood function under the missing data framework [20]. Under the CB model, however, a single-arm trial no longer provides information for estimating relative effects

(treatment contrasts). Therefore, the whole two-arm trial must be deleted if either arm is deleted; we refer to this setting as Scenario 1. In this scenario, we can no longer use (3.7) for single-arm deletion, but must instead resort to case influence determination by deleting multiple cases using the more general equation (8.15) in Section 8.4 of Hodges [52], i.e.

$$\hat{\mathbf{\Theta}}_{(\mathbf{I})} - \hat{\mathbf{\Theta}} = (\mathbf{A}'_{(\mathbf{I})}\mathbf{A}_{(\mathbf{I})})^{-\mathbf{1}}\mathbf{A}'_{(\mathbf{I})}\mathbf{Y}_{(\mathbf{I})}, \tag{3.17}$$

where subscript $(\mathbf{I})$ indicates deleting multiple cases indexed by $\mathbf{I}$, and

$$(\mathbf{A}'_{(\mathbf{I})}\mathbf{A}_{(\mathbf{I})})^{-\mathbf{1}} = (\mathbf{A}'\mathbf{A})^{-\mathbf{1}} + (\mathbf{A}'\mathbf{A})^{-\mathbf{1}}\mathbf{A}'_{\mathbf{I}}(\mathbf{I} - \mathbf{V}_{\mathbf{I}})^{-\mathbf{1}}\mathbf{A}_{\mathbf{I}}(\mathbf{A}'\mathbf{A})^{-\mathbf{1}}, \tag{3.18}$$

where $\mathbf{A_I}$ is the matrix composed of the deleted rows of $\mathbf{A}$, and $\mathbf{V_I} = \mathbf{A_I}(\mathbf{A}'\mathbf{A})^{-\mathbf{1}}\mathbf{A}'_{\mathbf{I}}$. Then the relative change $\mathbf{RC}$ can be calculated by dividing $\hat{\mathbf{\Theta}}_{(\mathbf{I})} - \hat{\mathbf{\Theta}}$ by psd, as in (3.10).

For multi-arm trials, such as Trials 1 and 2 in our example dataset, we investigate 3 scenarios when deleting certain trial-by-arm combinations. Here we can either delete the entire 3-arm trial (Scenario 2), or we can delete one trial-by-arm observation (since two treatment arms will remain to estimate the mean treatment contrast). If we choose to delete only one of the trial-by-arm observations within a multi-arm trial, the method used depends on whether the deleted observation is the baseline arm (Scenario 3) or not (Scenario 4). We can still use (3.17) and (3.18) for Scenario 2 to detect influential studies, where the set I now consists of three trial-by-arm observations. For deleting single trial-by-arm combinations, we can use (3.8) for Scenarios 3 and 4, keeping in mind that under Scenario 3 we also need to change the baseline to the first treatment arm remaining in the dataset after deletion of the original baseline in the full dataset. So the design matrix X changes in Scenario 3, more than simply omitting a row in X.

As before, $|\mathbf{RC}| \geq 2$ suggests an influential case, and for comparison purposes, we calculate $\mathbf{RC}$'s using the linear approximation method proposed here and, for selected

trial-by-arm combinations, by rerunning the MCMC algorithm for comparison.

3.3 Results for the Example Dataset

Previous inconsistency detection methods based on both CB and AB models (Chapter 2) have suggested that Trials 22 and 23 are sources of inconsistency in the dataset [55]. Therefore, we are interested in checking whether these observations are influential cases or outliers.

For case influence, the linear approximation was performed under the AB model and, for these selected study-arm combinations, the MCMC was rerun. One advantage of the linear approximation method for the AB model is that it can calculate **RC**'s for all the observations in one step, avoiding the cumbersome rerunning of the MCMC for each deleted observation. Using the linear approximation method, we found only 1 influential observation: Trial 17 Treatment 5, with **RC** = 10.72 as shown in Table 3.1. Rerunning the MCMC without that observation gives the exact **RC**, 10.40. This is reasonable since observation [17, 5] is the only trial having Arm 5; therefore, deleting that case means the estimate for Arm 5 is determined entirely by the prior for μ_5, which results in a large **RC**. We also examined whether the sources of inconsistency (i.e., the four arms from Trials 22 and 23) are influential. As shown in Table 3.1, the approximate and exact **RC**'s are close, indicating good performance for the linear approximation method, and none of these 4 observations emerges as particularly influential.

We also considered exact and approximate case influence under the CB model (Section 3.2). As noted, here we need to differentiate Scenarios 1-4 to apply the approximation method, so the **RC**'s for all observations can no longer be calculated in one step. As shown in Table 3.2, only deletion of Trial 1 Arm 4 gives a large relative change in d_4 (**RC** = 3.68) using the linear approximation method, compared to the exact **RC** of 3.83 obtained by rerunning the MCMC. Trial 1 Arm 4 is influential using the CB

Table 3.1: **RC**'s for influential case and observations as sources of inconsistency for thrombolytic drugs dataset using AB model.

	RC				
Omitted arm $\longrightarrow$	trial 22 arm 2	trial 22 arm 8	trial 23 arm 2	trial 23 arm 8	trial 17 arm 5
Parameter	μ_2	μ_8	μ_2	μ_8	μ_5
Linear approximation	0.67	-0.13	0.56	-0.18	**10.72**
Rerun MCMC	0.50	-0.12	0.39	-0.23	**10.40**

Table 3.2: **RC**'s for influential case and observations as sources of inconsistency for thrombolytic drugs dataset using CB model.

	RC				
Omitted arm $\longrightarrow$	trial 22 (arm 2)	trial 22 (arm 8)	trial 23 (arm 2)	trial 23 (arm 8)	trial 1 arm 4
Parameter	d_2	d_8	d_2	d_8	d_4
Scenario	1	1	1	1	4
Linear approximation	0.18	-0.23	0.12	-0.15	**3.68**
Rerun MCMC	0.31	-0.25	0.22	-0.23	**3.83**

model, perhaps due to its very large sample size (N=10,328). As in the AB analysis, the **RC**'s from deleting Trial 22 and Trial 23 are all much smaller than 2, indicating modest influence on estimated effects from these sources of inconsistency. **RC**'s from the linear approximation and exact methods are close, showing satisfactory performance of our diagnostic method using the CB model. Although the results from the AB and CB models identified different influential cases, this can be explained by the different parameterizations: in the AB model the estimated μ's are absolute effects, while in the CB model the d's are contrast effects.

We show outlier detection using standardized residuals for the AB linear model approximation. Table 3.3 presents the trial-by-arm combinations having the most extreme standardized residuals and standardized bias-corrected residuals. The absolute standardized residuals for Arm 2 in Trials 22 and 23 are both greater than 2, though they are somewhat attenuated after the bias correction. This indicates that the sources

Table 3.3: Outlier detection for thrombolytic drugs dataset.

trial	arm	standardized residuals	trial	arm	standardized bias-corrected residual
19	6	-1.5110	19	6	-0.8936
10	1	-1.5308	10	1	-1.1195
13	1	-1.5859	13	1	-1.2525
20	7	1.8580	**22**	**2**	-1.2976
23	**2**	**-2.0150**	20	7	1.3899
22	**2**	**-2.0805**	**23**	**2**	-1.4219

of inconsistency are outliers in the example dataset to some extent, but that outliers are not necessarily influential, as is well-known from linear model theory.

3.4 Simulation Studies

This section further investigates case influence under the AB model framework using simulations. In these simulations, we generate datasets from CB models in Zhao et al. [55] with fairly extreme inconsistency; our purpose is to investigate whether these sources of inconsistency are influential.

3.4.1 Simulation Settings

We simulated data from a network meta-analysis with 4 treatments and binary outcomes as shown in Figure 3.1. There are 6 pairwise comparisons (1-2, 1-3, 1-4, 2-3, 2-4, and 3-4) among these 4 treatments. We simulated data using the CB model, with inconsistency modeled using an inconsistency factor w for each loop (123, 124 and 134) as follows (see Section 2.2.5):

$$\begin{aligned} d_{23} &= d_{13} - d_{12} + w_{123}, \\ d_{24} &= d_{14} - d_{12} + w_{124}, \\ \text{and} \quad d_{34} &= d_{14} - d_{13} + w_{134}. \end{aligned} \tag{3.19}$$

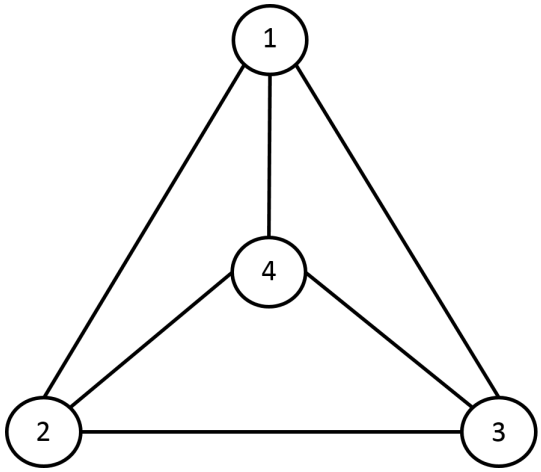

Figure 3.1: Network Graph for Simulation (Each vertex represents a treatment).

Similar as in Section 2.3.1, we assigned the true values of the α_{iB} from -2 to -3. True values of d_{12}, d_{13}, and d_{14} were still 0.5, 0.9 and 1.2 respectively, with true inconsistency factors $w_{123} = w_{124} = 0.01$ and $w_{134} = 2.5$. We set the standard deviation σ in (2.3) as 2, and $n_{ik} = 100$ for all treatments k and studies i. From this data generation scheme, the evidence in loop 134 is strongly inconsistent, especially for comparing 3 vs. 4. To create different degrees of inconsistency with this set-up, we considered several settings, as follows: (a) assign 5 trials to each pairwise comparison, 30 trials in total (I=30); (b) assign 1 trial to each comparison, 6 trials in total (I=6); and (c) repeat setting (b), but specify the simulated data as 90 events out of 100 person (i.e. $y_{ik} = 90$ and $n_{ik} = 100$) for both Arms 3 and 4 in trial 6 (for comparison 3 vs. 4). We simulated and analyzed 1000 datasets for each setting using the `Brugs` package in `R`, where we call `OpenBUGS` [47] from `R`, once for each simulated dataset. In each setting, the simulated proportions of events meeting the standard criterion $|\mathbf{RC}| \geq 2$ were calculated. To avoid using too strict a criterion to label cases as influential, proportions of events with $|\mathbf{RC}| \geq 1$ were also calculated.

Table 3.4: **RC**'s for simulated datasets under simulation setting 2. The entry of each cell is fraction of events with $|\mathbf{RC}| \geq 2$ out of 1000 simulations.

	Fractions with RC≥ 2 (out of 1000 simulations) for simulation setting (b)											
	[1,1]	[1,2]	[2,1]	[2,3]	[3,1]	[3,4]	[4,2]	[4,3]	[5,2]	[5,4]	[6,3]	[6,4]
μ_1	0	0	0	0	0	0	0	0	0	0	0	0
μ_2	0	0.7%	0	0	0	0	0	0	0	0	0	0
μ_3	0	0	0	1.4%	0	0	0	0.1%	0	0	0	0
μ_4	0	0	0	0	0	0.4%	0	0	0	0.4%	0	1.8%

3.4.2 Simulation Results

In simulation setting (a), none of the observations were influential cases using the criteria $|\mathbf{RC}| \geq 2$ or $|\mathbf{RC}| \geq 1$, even though inconsistency was truly present (results are omitted here). This might be because we have many other trials providing informations about Arms 3 and 4, besides the inconsistent trials, so that the effects of the discrepant evidence are diluted and thus less influential. Setting (b) used only one trial for each direct comparison. Using the criterion $|\mathbf{RC}| \geq 2$, the highest fraction of influential cases in the direct comparison of 3 vs. 4 is just 1.8% in Table 3.4. Overall, these low fractions show that the discrepant cases are not influential, by the $|\mathbf{RC}| \geq 2$ criterion.

Setting (c) further enhances the strength of inconsistency by artificially forcing the data to have a large number of events for the already discrepant treatments, which we expected to deliver bigger **RC**'s. Table 3.5 shows the influence of Trial 6 for Settings (b) and (c), for two different **RC** criteria. With the criterion $|\mathbf{RC}| \geq 2$, the discrepant trial-arms are found to be influential in few datasets. But more influential cases emerge when we increase the evidence for inconsistency in setting (c). This trend is more obvious when we use the criterion $|\mathbf{RC}| \geq 1$, with the proportion of influential cases being 41.2% in Setting (c), versus just 16.9% in setting (b) for Arm 3 in Trial 6. This indicates that the proportion of cases labeled as influential increases in a setting with

Table 3.5: Discrepancy factors for simulated datasets using ABRE model. The entry of each column is the summary of results from 1000 datasets under each scenario.

	RC≥2		RC≥1	
Omitted arm ⟶	trial 6 arm 3	trial 6 arm 4	trial 6 arm 3	trial 6 arm 4
Simulation setting (b)	0	1.8%	16.9%	35.0%
Simulation setting (c)	1.8%	0.9%	41.2%	35.4%

increased the evidence for inconsistency, but the big discrepancy factor in the data-generation mechanism does not imply that Trial 6's arms are influential even half the time; we also need to "cook" the NMA design (Settings (b) and (c)) to make it happen even that often.

3.5 Discussion

In this chapter, we proposed novel diagnostic methods to detect influential observations and outliers at the trial-by-arm level in NMA. Results using our example dataset show that the linear approximation method to detect influential cases performs well compared to an exact computation, while being more practical and computationally less expensive. Our method can be applied to NMA using either the AB or CB model, although the procedure is more complex for the CB model. Moreover, we found that the sources of inconsistency in NMA tend not to be influential under the standard criterion $|\mathbf{RC}| \geq 2$. Simulation studies confirmed this overall conclusion for case influence, though the chance of having cases declared influential increases when more severe discrepancy exists in NMA; this was more obvious with the less stringent criterion $|\mathbf{RC}| \geq 1$. Therefore, as a matter of data-analytic strategy, if it is easy to run discrepancy checks, it may be efficient to use them to identify trial-arm combinations that are potentially influential. The ultimate question is not whether direct and indirect evidence are discrepant according

to a significance test (Bayesian or frequentist), but whether that discrepancy affects the substantive conclusions we draw from a NMA, i.e., whether this discrepant information is influential.

To detect outlying observations, we used both standardized residuals and bias-corrected standardized residuals. In theory, bias-corrected standardized residuals should be preferred in diagnostics for hierarchical models, but because the phenomenon of biased residuals in hierarchical models is not widely known and bias-correction is not widely used, we have presented both kinds of residuals. They performed similarly in our example, identifying inconsistency sources as outlying observations, though differing in the apparent severity of outlyingness. Though outlier detection was not investigated via simulation, it is likely that outlying points tend to be sources of inconsistency: consistency holds when direct and indirect evidence agree, so if there are outlying observations, treatment effects estimated from them will be more likely to be discrepant. This is also mentioned in Zhang et al. [19], who suggest that outlying trials can be the primary source of inconsistency. These authors also recommend caution in deleting outlying observations, and prefer examination of inclusion criteria for trials in the network, where more stringent criteria may lower the numbers of outlying observations in NMA.

Chapter 4

Combining RCTs with Observational Studies

In this chapter, we propose methods to combine randomized study and PS matched non-randomized (NR) study using commensurate priors [56]. Section 4.1 introduces an example that exemplifies the need for such combining due to the small sample size of a RCT. Section 4.2 describes a commensurate prior approach to adaptively borrowing information from the matched NR cohort to complement the parameter estimates from the randomized study (RS) data. In Section 4.3, we apply our approach to the FIRST dataset. Naive modeling based on either the RS or NR data is described; the results are surprisingly discrepant, motivating use of our new method to combine the matched NR data with the RS data. Section 4.4 then evaluates our approach via simulation, showing different degrees of borrowing between the information contributed by NR and RS data, depending to some extent on the model and prior distribution we choose. Finally, in Section 4.5 we summarize and discuss some limitations of our method.

4.1 A Clinical Trial Example

In some cases, owing to the inherent difficulty of implementing RCTs, which require that the participating clinicians agree to forgo using their clinical expertise and randomly select therapies for their patients (as well as additional institutional oversight), statistical methods are needed to facilitate integrative analysis based on both types of data. The Flexible Initial Retrovirus Suppressive Therapies (FIRST) trial, conducted by the Community Programs for Clinical Research on AIDS (CPCRA), offers an example. As described in MacArthur et al. [57], highly active antiretroviral therapy-naive, HIV-infected subjects were randomized to three strategies (nucleoside reverse transcriptase inhibitor (NRTI) was used in all three strategies): a two-class protease inhibitor (PI+NRTI), a two-class non-nucleoside reverse transcriptase inhibitor (NNRTI+NRTI), and a three-class strategy (PI+NNRTI+NRTI). Participants within the two strategies involving NNRTIs could further specify whether they wanted to be randomly assigned to a NNRTI drug (nevirapine, NVP, or efavirenz, EFV) before the randomization to strategy arms, or permit a study clinician to prescribe one of the two drugs. The three strategies were compared for long-term virological and immunological durability, drug resistance, and disease progression.

Figure 4.1 offers a pictorial representation of the study design. Here, we consider only the data from the two-class NNRTI strategy, including randomized EFV ($n = 45$) or NVP ($n = 53$), as well as patients whose clinician chose EFV ($n = 211$) or NVP ($n = 100$). Our data set excludes patients missing an 8-month plasma HIV RNA measurement. Our goal is to compare the probability of virological suppression (HIV RNA < 50 copies/ml) under EFV and NVP at 8 months, adjusting for several baseline covariates. Analysis of the small randomized substudy data alone yields insufficient power. The larger non-randomized cohort is likely more representative of the HIV/AIDS population at large, but is subject to selection bias due to patient preference, local medical

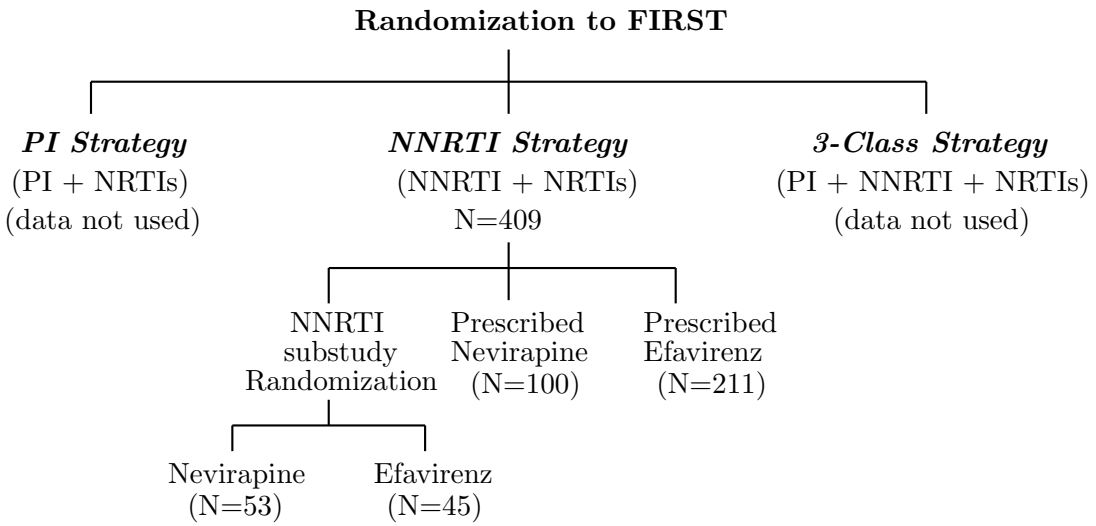

Figure 4.1: Outline of FIRST design and randomization for eligible subjects.

practice patterns, and other factors. Thus, we need to develop a method to cautiously combine the randomized and non-randomized data, after attempting to correct the latter for treatment selection bias.

4.2 Statistical Methods

After accounting for measurable imbalances in attributes among the NR patients using the propensity-score methods described in Chapter 1, we can use Bayesian hierarchical modeling to combine the NR and RS data for integrative analysis. Because randomization precludes systematic selection bias, the matching step only pertains to the NR cohort. This section describes models for integrating the NR and RS data.

Several authors have considered methods for incorporating historical data using Bayesian hierarchical modeling. Pocock et al. [58] advocated incorporating historical control data into clinical trial analysis under certain "acceptability" conditions. Ibrahim and Chen [59] introduced *power prior* (PP) methods to down-weight the historical data

relative to the current data. The PP approach assumes identical model parameters in the historical and current data and uses a weight (likelihood exponent) between 0 and 1 to control the extent to which the historical information influences the posterior. However, the weight is often difficult to estimate from the data, especially for non-Gaussian models.

Hobbs et al. [60] proposed *commensurate priors* for linear and generalized linear mixed models to facilitate dynamic partial pooling of between-source information, in which the extent of borrowing is estimated flexibly. This Bayesian hierarchical model assumes that the parameter vector θ for the current data follows a normal distribution with mean θ_0 estimated from historical data, and a precision or *commensurability* parameter τ, that is like a precision (reciprocal of variance). When evidence for commensurability is very strong between the two sources of data, τ will be sufficiently large that θ is constrained to be very close to θ_0. On the other hand, when τ is close to 0, the variance of the conditional prior for θ given θ_0 is large, indicating weak commensurability and thus less borrowing between θ_0 and θ. In previous work, the goal was to develop hierarchical priors that facilitate a model where the degree of borrowing was driven by the similarity of the historical and current data, and thus the authors considered posterior inference using various types of priors for the commensurability parameters.

Similarly, we can link the modeling of matched NR data and unmatched RS data using a commensurate prior framework. First, we can fit a Bayesian hierarchical model to the matched NR data as described by Agresti et al. [61]. Our clinical outcome can be a member of the exponential family [62], and here we assume it is binary and follows a Bernoulli distribution (logit link). Let the observations for pair m be (y_{m1}, y_{m2}), where m refers to the matched pair index, and the second index $k = 1$ or 2 refers to the control and treatment groups, respectively. Then we can impose a standard fixed treatment

effects model for pair (y_{m1}, y_{m2}) as follows:

$$\log\left(\frac{\pi_{m1}}{1-\pi_{m1}}\right) = \alpha_m \text{ and } \log\left(\frac{\pi_{m2}}{1-\pi_{m2}}\right) = \alpha_m + \lambda_0 \,, \tag{4.1}$$

where π_{mk} denotes the probability that $y_{mk} = 1$ and α_m the baseline effect for the control arm in the m^{th} pair. Here by allowing for pairwise specific intercepts, we implicitly assume that the matches themselves characterize varying magnitudes of prognostic effects, while the extent to which one treatment should be preferred with respect to another, λ_0, is global. Therefore, we expect that the log-odds of obtaining a response is heterogeneous among the pairs and we can account for this using a hierarchical model to estimate the extent of inter-pair heterogeneity. We do so by letting the α_m follow a normal distribution with mean μ_0 and precision parameter τ_1, and assume λ_0 is a common treatment effect.

If we additionally let the treatment effects vary between pairs, we can switch to a random treatment effects model,

$$\log\left(\frac{\pi_{m1}}{1-\pi_{m1}}\right) = \alpha_m \text{ and } \log\left(\frac{\pi_{m2}}{1-\pi_{m2}}\right) = \alpha_m + \lambda_m \,, \tag{4.2}$$

where λ_m denotes the now pair-specific treatment effect for m^{th} pair relative to the baseline. This model acknowledges that the extent to which one treatment might be favored versus another varies by pair, suggesting an interaction between treatment and the covariates used for PS matching. For estimation, we can assume that the random effects λ_m are exchangeable with a common mean λ_0 and precision parameter τ_2 to characterize inter-pair heterogeneity in treatment effectiveness as a variance component. Without loss of generality, weakly informative priors can be used for these parameters. Specifically, the fixed effects (μ_0 and λ_0) are assumed to follow independent $N(0, 1000)$ distributions since these parameters are well informed by the data in this model, and we adopt conventional Inverse Gamma$(0.1, 0.1)$ priors for the precisions τ_1 and τ_2.

For the unmatched RS data, we can incorporate the supplemental information from the matched NR study by fitting another general linear model with logistic link function

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mu + \lambda z_i \ . \tag{4.3}$$

Here i indexes subjects, μ is the baseline effect, λ is the coefficient for the treatment effect, and z is the $0-1$ indicator for treatment arms. To incorporate the matched NR data, a commensurate prior can be structured for λ so that $\lambda \sim N(\lambda_0, \frac{1}{\tau_3})$, where the commensurability parameter τ_3 (a precision) controls the degree of borrowing from the NR data. Here, we specify the prior for μ as a $N(0, 1000)$ distribution. However, as discussed in Hobbs et al. [56], estimation of τ_3 is inherently difficult. Therefore, a "spike and slab" prior [63] for τ_3 is more appropriate here, since it induces sparsity for estimating hierarchical variance components that are difficult to estimate from the data, and yields a dynamic borrowing procedures that has desirable bias-variance tradeoffs when integrating information that is potentially biased. In essence, this distribution is a mixture of a uniform distribution and a probability mass concentrated at a point $a > S_u$:

$$\begin{aligned} P(\tau_3 < S_l) &= 0, \\ P(\tau_3 < v) &= (1-p_0)\frac{v - S_l}{S_u - S_l} \quad (S_l \le v \le S_u), \\ \text{and} \quad P(\tau_3 > S_u) &= P(\tau_3 = a) = p_0, \end{aligned} \tag{4.4}$$

where S_u and S_l are the upper and lower bounds for a uniform distribution, and p_0 denotes the prior probability that τ_3 attains the value of the spike. Usually, we calibrate the values for a (the spike), the boundaries of the uniform distribution (the slab), and p_0 to represent different degrees of informative context (weakly informative to very informative). We can reduce our hyperparameter specification by setting $S_l = 0$ and $S_u = a$ ("two-parameter spike and slab" prior) when we have a smaller sample size. A

very small p_0 will produce less commensurability when a is small, but more commensurability when a is large.

Sensitivity of the spike and slab prior to hyperparameter selection has been investigated in previous research. Murray et al. [64] found that posterior inference for λ was not sensitive to modest shifts in these hyperparameters except for p_0. These authors suggest further simplifying this prior, which we do by modifying it to a "two-spike" prior by introducing a second spike at a small value r. This implies $\lambda \sim p_0 N(\lambda_0, 1/R) + (1 - p_0)N(\mu_0, 1/r)$, a two-part scale mixture. With this specification, one can either choose a value for p_0 or specify a hyperprior for it.

4.3 Results from the FIRST Trial

4.3.1 Baseline Characteristics and Naive Modeling

In this section, we demonstrate our method with an application using the FIRST dataset from Section 4.1. Baseline covariates were compared between treatment groups using the standard t test for continuous variables and the chi-squared test for categorical variables, respectively. Table 4.1 gives summaries of available covariates, including age, race (white vs. others), progression of disease before randomization or not ("podbl", 1: yes), baseline average cd4 count ("cd4bl"), log baseline HIV RNA level ("lrnabb"), gender (1: male), whether a male subject has homosexual activity with other men ("malesex", 1: yes), and injection drug use ("idu", 1: yes) for NVP-treated or EFV-treated participants in both the RS and NR groups. Results show that the baseline covariates were acceptably balanced between treatment groups in the randomized cohort, while one covariate (baseline average CD4 count) differed significantly between the NVP or EFV participants in the NR cohort (unadjusted p-value = 0.03). This could be the result of an early study [65] that suggested EFV was more efficacious for low CD4 patients. NVP

Table 4.1: Comparison of baseline characteristics between FIRST treatment groups in the randomized and non-randomized cohorts prior to matching. Continuous variables are reported as mean ± standard deviation. Dichotomous variables are reported as N (percent)

.

	Randomized cohort			Non-randomized cohort		
	EFV (N=45)	NVP (N=53)	p-value	EFV (N=211)	NVP (N=100)	p-value
age	39.0±7.6	36.7±8.5	0.17	38.6±9.9	38.9±8.4	0.78
race	12 (26.7)	14 (26.4)	0.98	61 (28.9)	26 (26.0)	0.59
podbl	16 (35.6)	22 (41.5)	0.55	83 (39.3)	38 (38.0)	0.82
cd4bl	227.7±207.3	209.5±193.0	0.66	190.2±189.0	242.9±227.3	**0.03**
lrnabb	5.1±0.9	5.2±0.8	0.56	5.1±0.8	4.9±0.8	0.08
gender	35 (77.8)	40 (75.5)	0.79	167 (79.2)	77 (77.0)	0.67
malesex	21 (46.7)	22 (41.5)	0.61	100 (47.4)	41 (41.0)	0.29
idu	5 (11.1)	11 (21.2)	0.18	25 (11.9)	17 (17.0)	0.22

was also suspected of causing liver problems [66], whereas EFV's side effects (dreams, nightmares) were thought to be less severe, possibly further explaining the physicians' overall preference for EFV (211 vs 100).

Let $Y_i = 1$ if patient i experiences virological suppression (VS) at the 8-month visit. We assume $Y_i \sim Bernoulli(\pi_i)$, where π_i denotes the probability of VS for $i = 1, \ldots, N_c,\ c = RS, NR$, where N_c represents sample size for either the RS or NR cohort. Now let $x_1, \ldots, x_p$ denote p baseline covariates, and z an indicator variable for the intervention group (such that $z_i = 1$ is NVP; $z_i = 0$ is EFV). Then we can compare the odds of VS between treatments using a generalized version of model (4.3),

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = (\mathbf{Xb})_i + \lambda z_i\ , \tag{4.5}$$

where $\mathbf{b} = (\beta_1, \ldots, \beta_8)'$, $\mathbf{X}$ is a $N_c \times 8$ design matrix, and λ is the log-odds ratio of VS for NVP versus EFV. If we perform a naive frequentist analysis for the randomized and NR data separately without any PS adjustment in the NR cohort, we find that the

difference between NVP and EFV groups was significant in the NR cohort (p-value = 0.02), but not in the randomized cohort (p-value = 0.24). More surprisingly, the results differed in the direction of the effect. Specifically, the log-odds ratio λ using EFV as the reference group was -0.68 with a 95% confidence interval (CI) $[-1.25, -0.11]$ in the NR cohort, indicating an increase in the odds of VS for patients receiving EFV relative to those receiving NVP, but was 0.79 in the RS cohort with a 95% CI $[-0.54, 2.13]$, indicating a relative *decrease* in odds of VS for patients receiving EFV. The smaller sample size in the RS cohort leads to the decrease in precision (wider interval estimate).

Motivated by this discrepancy, we seek to combine the RS and NR data using the methods described in Section 4.2. Specifically, we will use PS matching within the NR cohort, and then combine the information with the RS data using Bayesian hierarchical commensurate prior models, which gives privilege to the RS data.

4.3.2 Combining RS and PS-matched NR Study Data

As mentioned in Section 1.3.2, we first estimate PS in the NR cohort using logistic regression, i.e., we regress the probability of being assigned to the NVP group on the 8 covariates described in Table 4.1. For the FIRST dataset, after we calculate the PSs using the NR cohort, we must match participants between its NVP and EFV arms. We match one treatment observation with a control observation, subject to a maximum allowable difference between their propensity scores. This maximum allowable difference is called the *caliper width*, and is usually designated by simply assigning a reasonable but subjective value that determines the extent of allowable dissimilarity among matched pairs; for example, 0.1. After each pair is matched, it is removed from the pool. This procedure is repeated until all NVP patients are matched to the EFV patients, or until no further EFV observations fulfill the matching criteria (89 matched pairs). After matching, balance diagnostics, e.g. ,those proposed by Austin et al. [67],

can be performed to assess whether the propensity score model has been adequately specified. At this point, all the covariates are balanced between NVP and EFV groups. In our data, e.g., we obtained a p-value of 0.81 for the baseline average CD4 count in the NR study using a paired t test, suggesting that the procedure achieved reasonable balance among the matched pairs (all p-values for other matched covariates are large, so results are omitted here). The log-odds ratio λ was re-estimated using the PS-matched NR data, resulting in a point estimate of -0.72 with a 95% CI $[-1.43, -0.01]$ (p-value $= 0.04$).

Results obtained from combining the NR and RS data using the Bayesian hierarchical model described in Section 4.2 are summarized here for the FIRST dataset. For computation, we used `OpenBUGS` [47] to draw two Markov chain Monte Carlo (MCMC) chains from the posterior for 30,000 iterations each, after 20,000 iterations of burn-in. The final estimate of the log-odds ratio λ was -0.30 with a 95% Bayesian credible interval (BCI) of $[-0.88, 0.32]$ for the fixed effects model using the two parameter spike and slab prior. Here, we chose $a = 40$ and $p_0 = 0.3$, which represents a moderately informative prior because of the relatively high quality of the NR data (all subjects, both NR and RS, met the FIRST entry criteria). The random effects model (4.2) was also fitted, but not reported here since its results were similar. The pooled treatment effect of -0.30 lies in between those obtained using either the RS or matched NR data alone, indicating an increase in the odds of VS for EFV patients compared to NVP subjects, though the increase is not statistically significant based on its BCI. The BCI width (1.20) for the log-odds ratio λ is smaller than those obtained using either the RS data alone (2.67) or the matched NR data alone (1.42). The posterior mean of τ_3 was estimated as 25.85 using the fixed effects model, showing a moderate degree of borrowing from the NR data.

Our results provide formal support for the findings in Van den Berg-Wolf et al. [68],

showing a lower rate of virological failure in persons taking EFV as compared to those taking NVP, although our combined data result is not statistically significant. This might be because we cautiously borrowed from the matched subset of NR data, while they combined all the NR and RS data from both the NNRTI and 3-class strategies, making the strong assumption that these data sources are exchangeable and equally free of bias. They also found an *increased* risk of disease progression or death for those randomized to EFV when compared to those randomized to NVP. These findings indicate pros and cons of using either medication in terms of different clinical endpoints. Our results suggest physicians may rely on their clinical judgment when prescribing these drugs, keeping in mind the expected risk-benefit trade-off for their patients.

4.4 Simulation Studies

We use a series of simulations to examine the performance of our models for combining NR data with randomized data using commensurate priors after propensity score matching. As in our FIRST example, here we consider only binary outcomes.

4.4.1 Simulation Settings

We simulated data with 4 baseline covariates (x_1 to x_4) for both the NR and RS data. These covariates were simulated from independent $N(0, 1)$ distributions. Among these 4 covariates, we designated two of them as affecting treatment selection (x_1 and x_2) for the NR data, while two other covariates (x_2 and x_3) affect the binary outcome for both datasets. Furthermore, these covariates were allowed to have a weak, moderate, strong, or very strong effect on treatment selection or outcome corresponding to β values of log(1.25), log(1.5), log(1.75) and log(2), respectively. For each subject, the true probability of treatment selection (propensity score) was determined from the following

logistic model:

$$\log\left(\frac{PS_i}{1-PS_i}\right) = \beta_{00} + \beta_{01}x_{1i} + \beta_{02}x_{2i}, \tag{4.6}$$

the treatment selection model. Therefore, the true treatment status for each patient i was generated from a Bernoulli distribution with a subject-specific parameter PS_i denoting the probability of being assigned to the treatment group: $Z_i \sim Bernoulli(PS_i)$. We then generated a binary outcome using the formula:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \lambda Z_i, \tag{4.7}$$

the outcome model. Similarly, the outcome Y_i was then generated from a Bernoulli distribution using $Y_i \sim Bernoulli(\pi_i)$.

Several scenarios were simulated to test the degree of borrowing when we used different spike and slab commensurate priors for τ_3. In Scenario 1, we assumed that the same model applied to the outcomes for both the NR and RS data, where β_2 and β_3 were set to log(1.25), log(1.5) respectively. The true treatment effect λ was set to -0.7 when the total sample size was 200 for two cohorts (100 patients in each cohort), indicating an odds ratio of 0.5 comparing treatment to placebo. We also set the true treatment effect to -0.4 for a sample size 850 (425 patients in each cohort), giving an odds ratio of 0.67. In Scenario 2, we kept the same true treatment effects λ for the NR and RS data; however, different sets of β coefficients (see (4.7)) for the covariates were used in the outcome models. Specifically, for the NR data, we used log(1.75) and log(2) for β_2 and β_3, while for the RS data, log(1.25) was used for both β_2 and β_3. In Scenario 3, we assumed that the treatment effects were very different in the NR and RS data (0.7 for the RS cohort but -0.7 for the NR cohort), while β_2 and β_3 in the outcome model stayed the same for both cohorts as in Scenario 1. In Scenario 4, we assumed not only that the treatment effects λ were very different in the NR and RS data (0.7 for the RS cohort but -0.7 for the NR cohort), but also used the same β_2 and β_3 in the

outcome models as in Scenario 2. Besides the above scenarios, we also tested scenarios in which we predicted the PSs from the treatment selection model (see (4.6)) using only x_1 and x_2, only x_3 and x_4, or all four covariates (x_1 to x_4). The results showed that these models perform similarly as long as we predict the PSs using covariates in the true model (x_1 and x_2 in our case); therefore, we will use x_1 to x_4 to predict PS in all settings.

We checked three different commensurate priors under each scenario as follows: (1) a two parameter spike and slab prior with $a = 40$ and $p_0 = 0.3$ ("CP1"), (2) a two parameter spike and slab prior with $a = 1000$ and $p_0 = 0.01$ ("CP2"), and (3) a two-spike prior with $R = 2000$, $r = 0.01$ and $p_0 = 0.1$ ("CP3"). The results using all three commensurate priors were also compared to the model where we used only the RS data ("No CP"). In each of these settings, we generated 1000 simulated datasets (each of sample size 200 or 850) using `R`, where we used the `Brugs` package to call `OpenBUGS`, once for each simulated dataset. Using each of the 1000 simulated datasets, we performed the propensity score matching based on *all* the baseline covariates x_1 to x_4 in the generated NR data. We then used the fixed effects model for paired data as shown in (4.1) on the propensity-score-matched sample. Finally, model (4.3) was used to estimate the treatment effect in the RS data using each commensurate prior for each simulated sample. The average posterior bias, 95% BCI width, and mean squared errors (MSE) for all the parameters (β_2, β_3 and λ) were calculated across the 1000 simulated datasets. Empirical coverages of the 95% BCIs for parameter λ were calculated, and the power for finding a significant treatment effect was estimated as the empirical proportion of these BCIs that did not contain zero out of 1000 simulations under the alternative hypothesis (true treatment effect was different from 0).

Table 4.2: Posterior estimates and MSEs for key parameters, 95% empirical coverages for λ, and power using simulated datasets under Scenarios 1 and 2 (borrowing warranted). Each cell represents an average over 1000 simulations.

sample size	true λ		priors for τ_3	posterior λ estimates		τ_3/p_0 mean	MSE			95% coverage	power
	NR	RS		bias	BCI width		β_2	β_3	λ		
Scenario 1											
200	-0.7	-0.7	No CP	-0.07	1.84	-	0.13	0.15	0.48	0.93	0.40
			CP 1	-0.08	1.49	26.07	0.12	0.14	0.31	0.94	0.54
			CP 2	-0.08	1.42	504.68	0.13	0.14	0.30	0.93	0.59
			CP 3	-0.07	1.73	0.47	0.13	0.15	0.41	0.94	0.44
850	-0.4	-0.4	No CP	-0.01	0.82	-	0.02	0.03	0.09	0.95	0.50
			CP 1	0.00	0.71	26.89	0.02	0.03	0.06	0.96	0.62
			CP 2	0.00	0.66	505.31	0.02	0.03	0.06	0.95	0.69
			CP 3	0.00	0.75	0.65	0.02	0.03	0.07	0.96	0.59
Scenario 2											
200	-0.7	-0.7	No CP	-0.08	1.81	-	0.13	0.14	0.48	0.93	0.41
			CP 1	-0.04	1.48	26.04	0.12	0.13	0.31	0.94	0.51
			CP 2	-0.03	1.41	504.70	0.12	0.13	0.29	0.93	0.54
			CP 3	-0.07	1.71	0.46	0.13	0.13	0.41	0.94	0.44
850	-0.4	-0.4	No CP	-0.01	0.81	-	0.02	0.02	0.09	0.96	0.50
			CP 1	0.01	0.70	26.95	0.02	0.02	0.06	0.96	0.61
			CP 2	0.01	0.65	505.81	0.02	0.02	0.06	0.95	0.67
			CP 3	0.00	0.74	0.66	0.02	0.02	0.07	0.96	0.58

4.4.2 Simulation Results

In the construction of the true models above, we are imposing increasing heterogeneity between the NR and RS studies from Scenario 1 to Scenario 4. Therefore, we expect our models to capture this trend using commensurate priors for λ, borrowing less from the NR data when heterogeneity is high (i.e., we expect τ_3 to be small in models with CP1 and CP2, or p_0 to be small in models with CP3).

Tables 4.2 and 4.3 display the posterior estimates for the main parameters λ, τ_3 (or p_0 under CP3), MSEs for λ, β_2 and β_3, as well as 95% empirical coverages for λ and power averaged over the 1000 simulated datasets under each scenario described in

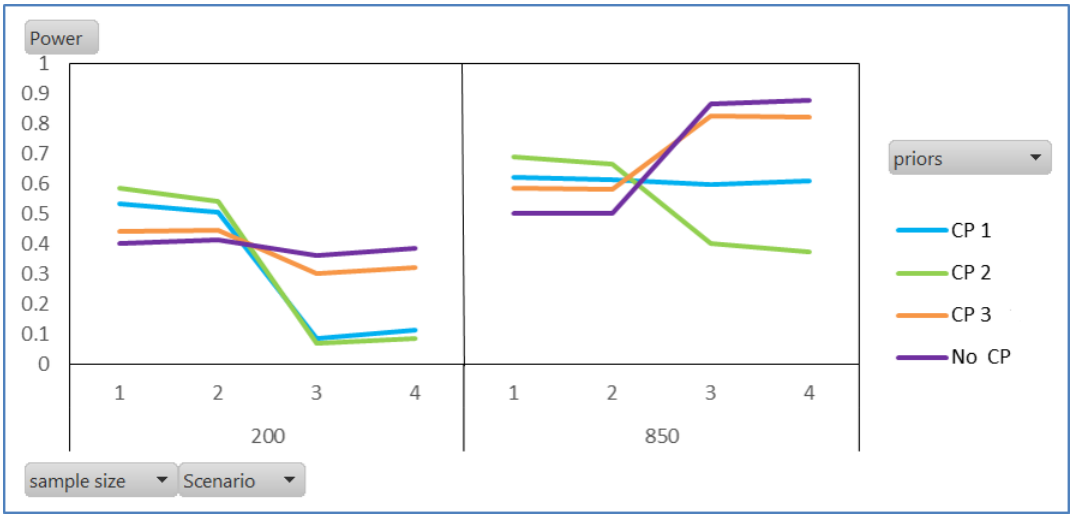

Figure 4.2: Power using simulated datasets under different scenarios and sample sizes.

Section 4.4.1. In Table 4.2, under Scenarios 1 and 2 where it is beneficial to use the NR studies (since the true treatment effects λ were the same across different sources of data), models with CP1 and CP2 (informative to very informative priors) performed better since they gave narrower 95% BCIs for λ that did not cover 0 (high power), indicating a statistically significant treatment effect. Moreover, the posterior estimates for τ_3 were large using models with CP1 and CP2, showing a substantial degree of borrowing from the NR data. These models yielded better (10% to 20% higher) power compared to models with No CP, which can be better viewed in Figure 4.2 under Scenarios 1 and 2 for both sample sizes. We also obtained about 35% reduction in MSE for the CP1-2 approach and about 20% for CP3 compared to models with No CP, where less MSE reduction was observed in the absence of bias as the trade-off. In the situation when we only used the RS data, we see the widest 95% BCI, covering 0 even when we used a larger sample size, indicating a non-significant treatment effect. The model with CP3 showed moderate borrowing from the NR data under first two scenarios, since this prior is weakly informative compared to CP1 and CP2.

Table 4.3: Posterior estimates and MSEs for key parameters, 95% empirical coverages for λ, and power using simulated datasets under Scenarios 3 and 4 (borrowing not warranted). Each cell represents an average over 1000 simulations.

sample size	true λ		priors for τ_3	posterior λ estimates		τ_3/p_0 mean	MSE			95% coverage	power
	NR	RS		bias	BCI width		β_2	β_3	λ		
Scenario 3											
200	-0.7	0.7	No CP	0.16	1.93	-	0.16	0.18	0.65	0.93	0.36
			CP 1	-0.40	1.80	21.41	0.15	0.17	0.60	0.83	0.09
			CP 2	-0.55	1.59	479.21	0.15	0.17	0.66	0.69	0.07
			CP 3	0.08	2.12	0.20	0.16	0.18	0.67	0.93	0.30
850	-0.7	0.7	No CP	0.02	0.91	-	0.03	0.03	0.11	0.95	0.87
			CP 1	-0.16	0.95	8.31	0.03	0.03	0.16	0.86	0.60
			CP 2	-0.26	0.97	201.45	0.03	0.03	0.23	0.74	0.40
			CP 3	0.01	0.93	0.03	0.03	0.03	0.12	0.94	0.83
Scenario 4											
200	-0.7	0.7	No CP	0.16	2.08	-	0.16	0.16	0.64	0.93	0.38
			CP 1	-0.37	1.75	22.25	0.15	0.16	0.56	0.84	0.11
			CP 2	-0.49	1.57	484.00	0.15	0.15	0.59	0.74	0.09
			CP 3	0.07	2.09	0.22	0.16	0.16	0.65	0.92	0.32
850	-0.7	0.7	No CP	0.02	0.91	-	0.03	0.03	0.11	0.95	0.88
			CP 1	-0.16	0.94	9.45	0.03	0.03	0.16	0.87	0.61
			CP 2	-0.28	0.95	240.64	0.03	0.03	0.24	0.71	0.37
			CP 3	0.01	0.93	0.04	0.03	0.03	0.12	0.95	0.82

In Table 4.3 under Scenarios 3 and 4 where we gave opposite signs for the treatment effects for the NR and RS data, the above results are essentially reversed, since now the true model assumes high heterogeneity between different sources of data. The posterior biases for λ are much larger using models with CP1 and CP2 compared to those from models with CP3 or No CP, especially when the sample size is small. Furthermore, although the posterior mean estimates for τ_3 decreased compared to those under Scenario 1 for models with CP1 and CP2, these priors are informative in the sense that they represent a strong preference for borrowing strength across cohorts. This is revealed by the power estimates in Table 4.3, where models with CP1 and CP2 resulted in very

low power (0.09 and 0.07 for Scenario 3; 0.11 and 0.09 for Scenario 4) for studies with sample size 200, and similar decreases relative to No CP and CP3 for sample size 850 (Figure 4.2 under Scenarios 3 and 4 for both sample sizes). By contrast, the model with CP3 resulted in more adaptive borrowing from the NR data, e.g., the posterior mean estimate for p_0 decreased from 0.65 in Scenario 1 to 0.03 in Scenario 3 for sample size 850. Moreover, the power from model with CP3 under Scenario 3 is 0.3 and 0.83 for sample sizes 200 and 850, respectively (similar results as No CP), indicating good performance under the no-borrowing scenario using this weakly informative prior. Results from Scenario 4 were similar to those from Scenario 3, showing similar degree of heterogeneity between different sources of data under these scenarios.

Moreover, results shows that Type-I errors for different models under various settings were well controlled under 0.05, showing no inflation problems (results omitted here).

4.5 Discussion

In this chapter, we developed a practical Bayesian statistical tool to combine OSs and RCTs. This was achieved using hierarchical models with priors that facilitate adaptive borrowing from matched NR data when justified by its commensurability with the corresponding RS data. Although our FIRST dataset is somewhat unique given that the same trial produced both the NR and RS cohorts, our model can be applied more broadly to integrate more and different kinds of OS data. Our approach also has the benefit of increasing the external validity of our results, because in practice only a select subpopulation is willing to be randomized. Borrowing strength is most appropriate when the two data sources are homogeneous, as in our FIRST dataset. However, in other settings, where we have more broadly collected data sources which include lower-quality OSs, models with weakly informative two-spike priors like CP3 would appear to be more appropriate, given their flexibility for adjusting to evidence of heterogeneity

across studies. This is also true when we do not have any prior information regarding the NR studies. As such, our method is very much in the spirit of the 21st Century Cures Act to streamline drug approvals, currently working its way through the U.S. House of Representatives [69, 70].

Our proposed method better uses all available data, but it has certain limitations. Its improvement in precision may not outweigh its potential for bias arising from "low-quality" PS matches. This is because "highest-quality" matches would arise only from identifying pairs of patients who are likely targets for the same treatment, but receive opposite treatments. As a somewhat controversial alternative to our approach in Section 1.3.2, where we first perform PS estimation in thhe NR cohort alone, we could instead use all the data (both RS and NR) in our PS matching. Specifically, for the randomized data, we no longer assume their PSs are fixed at 0.5, but rather estimate the probability of treatment assignment using data from *all* subjects. That is, we "predict" the PS for randomized subjects from the fitted partial regression coefficients obtained from the NR data. These PSs then have the interpretation of a measure of the probability that the patient would have been assigned to the treatment arm using the intrinsic "treatment selection". After PSs have been calculated for all patients, matching *between* the RS and NR data (e.g., matching one randomized control patient with high PS to one non-randomized treated patient with similarly high PS) enables estimation of the treatment selection counterfactuals. Similarly, these counterfactuals can then be used to estimate the average treatment effect using Bayesian hierarchical models for paired binary data as described in Section 4.2. The advantage of this method is that it is more likely to use high-quality matches (counterfactuals), but at the cost of disassembling the "gold standard" RS data, and possibly also discarding many observations in the (much larger) NR study due to the matching across cohorts.

Chapter 5

Conclusions and Future Work

In this chapter, major conclusions drawn from this thesis are summarized, and the aforementioned future directions of the work are discussed.

5.1 Major conclusions

In this dissertation, we have shown that Bayesian hierarchical modeling offers greater flexibility for research syntheses. In Chapter 2, we have proposed an inconsistency detection method for arm-based NMA models, illustrated via an example dataset and simulation, and obtained results comparable to those from its CB model counterpart, with more objective evidence for inconsistency. After inconsistency has been detected, we suggested that the key question was to check whether an inconsistent trial-arm combination would change the substantive conclusions of the NMA. So in Chapter 3, we developed a diagnostic method to explore whether the trial-arm combinations that are sources of inconsistency are influential or outlying observations. We did this by approximating the generalized linear model in NMA as a normal-error linear model, and then applying standard diagnostic tests. Our work results in fast one-step diagnostics

for NMA using AB models, and can also be applied to CB models, although it is more complex for the latter framework.

Besides the above application of Bayesian hierarchical modeling in NMA, we also applied it to combining the data from RCT and NR studies in Chapter 4. After matching on PS, causal inference can be made after mitigation of selection bias using only the NR data. Then we proposed using commensurate priors to model the relationship between the RCT and the matched NR data by controlling the degree of borrowing from the NR data, determined internally by the similarity of the treatment effects in the two studies.

In summary, the proposed Bayesian modeling framework in this dissertation has potential significance in practice. With recent advances in methods and software, Bayesian approaches to research synthesis have become quite popular, as they can allow models of far greater complexity (say, accommodating multiple endpoints), incorporate external information (i.e., historical data), and produce results having a direct interpretation. Therefore, the Bayesian approaches we have proposed should provide an easy-to-understand measure of evidence, which is often attractive to patients and their caregivers.

5.2 Future Perspectives

To further advance inconsistency detection as shown in Chapter 2, more work needs to be done to improve our AB models. Future work looks to extending our methods to continuous, count, or time-to-event outcomes, though the latter will likely require individual-level patient data except for the simplest of models. Rather than noninformative priors, weakly or even informative priors can be used if we have suitable information from historical or observational studies. Individual-level data can also be incorporated with the aggregated data summaries used above, allowing borrowing of strength from patient characteristics to better investigate treatment effects and assess

inconsistency.

For the diagnostic method in Chapter 3, though the outcome in our examples was binary, it can be directly applied to NMA with outcome measures on continuous scales, analyzed using a normal-errors hierarchical model without the need for pseudodata or pseudovariances. It can also be used within a frequentist framework, although our focus here was fully Bayesian. We also hope to extend our method to models including baseline covariates and individual patient-level data. Here again the goal would be to discover sources of influential observations and outliers, possibly downweighting their effects using subgroup analysis or bias adjustment methods [71].

In Chapter 4, we have implemented PS matching methods to balance covariates in NR data, which assumes that there are no unknown confounders. However, this is a strong assumption and is hard to check. So we might use other possible methods to mitigate treatment selection bias, including the simple substitution estimator (i.e. parametric G-computation) [72] or instrumental variable (IV) [73] methods. IV methods are widely used in economics because of the difficulty of doing controlled trials. The main idea of the IV method is to define variables called *instruments* that have two properties. First, they should highly correlate with the treatment choice, and second, they must not be directly related to the outcome measure. Given such variables, one can estimate how much the instrument induces variation in the treatment assignment, which later affects the outcome. A key assumption of the IV method is that there is no direct association between the IV and the outcome except through the treatment variable, since one can then think of the IV as a device to achieve a pseudo-randomization, akin to coin flipping. The difficulty in implementing the IV method is how to find good instrumental variables and validate their selection. This would be especially challenging in our FIRST dataset, since we would need two IVs: one for the patient's choice to accept or decline the substudy randomization, and a second for the clinician's choice of drug (NVP or EFV)

for those declining. Still, we hope to compare results from IV methods to those from PS matching methods using other example datasets and through simulation. Future methodological research in this area looks to extending our Chapter 4 approach to permit more types of auxiliary data to be incorporated into the network meta-analysis. With enough data, we can also let both the α_m in (4.1) and λ_m in (4.2) depend on covariates, with random effect terms soaking up unexplained variability in the model.

References

[1] H. Cooper, L. Hedges, and J. Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2009.

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

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

[4] R. Dersimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188, 1986.

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

[6] Pharmaceutical Benefits Advisory Committee. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (version 4.3)*. Canberra: Australian Government Department of Health and Ageing, 2008.

[7] G. Wells, S. Sultan, L. Chen, M. Khan, and D. Coyle. *Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis*. Ottawa: Canadian Agency for Drugs and Technologies in Health, 2009.

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

[9] S. Dias, N. Welton, D. Caldwell, and A. Ades. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, 29:932–944, 2010.

[10] J. Higgins, D. Jackson, J. Barrett, G Lu, A. Ades, and I. White. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods*, 3:98–110, 2012.

[11] A. Boland, Y. Dundar, A. Bagust, A. Haycox, R. Hill, R. Mujica Mota, T. Walley, and A. Dickson. Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation. *Health Technology Assessment*, 7:1–136, 2003.

[12] M. Fiore, W. Bailey, S. Cohen, and et al. *Smoking Cessation, Clinical Practice Guideline No. 18*. Agency for Health Care Policy and Research, U.S. Department of Health and Human Services, Rockville, MD, 1996.

[13] I. White, J. Barrett, D. Jackson, and J. Higgins. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods*, 3:111–125, 2012.

[14] D. Jackson, J. Barrett, S. Rice, I. White, and J. Higgins. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Statistics in medicine*, 33(21):3639–3654, 2014.

[15] H. Hong, H. Chu, J. Zhang, and B.P. Carlin. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods*, 2015. Accessed Nov. 4, 2015 (with discussion and rejoinder). DOI: 10.1002/jrsm.1153.

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

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

[18] S. Dias, N. Welton, A. Sutton, D. Caldwell, G. Lu, and A. Ades. *NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials*, 2011. URL.

[19] J. Zhang, H. Fu, and B. Carlin. Detecting outlying trials in network meta-analysis. *Statistics in medicine*, 2015.

[20] L. Lin, H. Chu, and J. Hodges. Sensitivity to excluding treatments in network meta-analysis. *Epidemiology*, 2016. to appear.

[21] J. Hodges. Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:497–536, 1998.

[22] Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA: the Journal of the American Medical Association*, 268:2420–2425, 1992.

[23] U.S. Preventive Services Task Force. *Guide to Clinical Preventive Services: Report of the U.S. Preventive Services Task Force*. Williams & Wilkins, Baltimore, 1996.

[24] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[25] R. Ligthelm, V. Borzi, J. Gumprecht, R. Kawamori, Y. Wenying, and P. Valensi. Importance of observational studies in clinical practice. *Clinical Therapeutics*, 29:1284–1292, 2007.

[26] R. MacLehose, B. Reeves, I. Harvey, T. Sheldon, I. Russell, and A. Black. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment*, 4:1–154, 2000.

[27] J. Concato, N. Shah, and R. Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342:1887–1892, 2000.

[28] J. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *JAMA: the Journal of the American Medical Association*, 294:218–228, 2005.

[29] R. Baron and D. Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51:1173–1182, 1986.

[30] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[31] G. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86:4–29, 2004.

[32] P. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399–424, 2011.

[33] R. D'Agostino. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17:2265–2281, 1998.

[34] T. Kurth, A. Walker, R. Glynn, K. Chan, J. Gaziano, K. Berger, and J. Robins. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163:262–270, 2006.

[35] D. Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2:808–840, 2008.

[36] P. Rosenbaum and D. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

[37] P. Rosenbaum and D. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38, 1985.

[38] J. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23:2937–2960, 2004.

[39] P. Austin and M. Mamdani. A comparison of propensity score methods: A case study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*, 25:2084–2106, 2006.

[40] P. Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33:1057–1069, 2014.

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

[42] G. Salanti, J. Higgins, A. Ades, and J. Ioannidis. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17:279–301, 2008.

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

[44] B. Carlin and T. Louis. *Bayesian methods for data analysis*. Chapman and Hall/CRC Press, 2011.

[45] D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde. The deviance information criterion: 12 years on (with discussion). *J. Roy. Statist. Soc., Ser. B*, 76:485–493, 2014.

[46] D. Jackson, P. Boddington, and I. White. The design-by-treatment interaction model: a unifying framework for modelling loop inconsistency in network meta-analysis. *Research Synthesis Methods*, 2015. accessed Nov. 20, 2015. DOI: 10.1002/jrsm.1188.

[47] D. Lunn, D. J. Spiegelhalter, A. Thomas, , and N. Best. The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, 28:3049–3067, 2009.

[48] P. Laud and J. Ibrahim. Predictive model selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):247–262, 1995.

[49] A. Gelfand and S. Ghosh. Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11, 1998.

[50] S. Senn, F. Gavini, D. Magrez, and A. Scheen. Issues in performing a network meta-analysis. *Statistical Methods in Medical Research*, 22(2):169–189, 2013.

[51] J. König, U. Krahn, and H. Binder. Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. *Statistics in Medicine*, 32(30):5414–5429, 2013.

[52] J. Hodges. *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. CRC Press, 2013.

[53] H. Lu, J. Hodges, and B. Carlin. Measuring the complexity of generalized linear hierarchical models. *Canadian Journal of Statistics*, 35:69–87, 2007.

[54] K. Chaloner. Residual analysis and outliers in Bayesian hierarchical models. In A. Smith and P. Freeman, editors, *Aspects of uncertainty: a tribute to DV Lindley*, pages 153–161. Wiley, 1994.

[55] H. Zhao, J. Hodges, H. Ma, Q. Jiang, and B. Carlin. Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis. *Statistics in Medicine*, to appear, 2016.

[56] B. Hobbs, D. Sargent, and B. Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7:639–674, 2012.

[57] R. MacArthur, L. Chen, D. Mayers, C. Besch, R. Novak, M. van den Berg-Wolf, T. Yurik, G. Peng, B. Schmetter, B. Brizz, and D. Abrams. The rationale and design of the CPCRA (Terry Beirn Community Programs for Clinical Research on AIDS) 058 FIRST (Flexible Initial Retrovirus Suppressive Therapies) trial. *Controlled Clinical Trials*, 22:176–190, 2001.

[58] S. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29:175–188, 1976.

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

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

[61] A. Agresti and Y. Min. Effects and noneffects of paired identical observations in comparing proportions with binary matched pairs data. *Statistics in Medicine*, 23:65–75, 2004.

[62] C. McCulloch and S. Searle. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, 2001.

[63] T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

[64] T. Murray, B. Hobbs, and B. Carlin. Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. *Annals of Applied Statistics*, 0:0, 2015.

[65] F. Van Leth, S. Andrews, B. Grinsztejn, E. Wilkins, M. Lazanas, J. Lange, and J. Montaner. The effect of baseline CD4 cell count and HIV-1 viral load on the efficacy and safety of nevirapine or efavirenz-based first-line HAART. *AIDS*, 19:463–471, 2005.

[66] J. Stern, P. Robinson, J. Love, S. Lanes, M. Imperiale, and D. Mayers. A comprehensive hepatic safety analysis of nevirapine in different populations of hiv infected

patients. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 34:S21–S33, 2003.

[67] P. Austin. A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27:2037–2049, 2008.

[68] M. van den Berg-Wolf, K. Hullsiek, G. Peng, M. Kozal, R. Novak, L. Chen, L. Crane, and R. MacArthur. Virologic, immunologic, clinical, safety, and resistance outcomes from a long-term comparison of Efavirenz-based versus Nevirapine-based antiretroviral regimens as initial therapy in HIV-1-infected persons. *HIV Clinical Trials*, 9:324–336, 2008.

[69] F. Upton, D. DeGette, J. Pitts, F. Pallone, and G. Green. 21st Century Cures Act, 2015.

[70] J. Avorn and S. Aaron. The 21st Century Cures Act - will it take us back in time? *New England Journal of Medicine*, 372:2473–2475, 2015.

[71] R. Turner, D. Spiegelhalter, G. Smith, and S. Thompson. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47, 2009.

[72] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[73] J. Newhouse and M. McClellan. Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, 19:17–34, 1998.

Appendix A

Hierarchical Bayesian Approaches for Detecting Inconsistency in Network Meta-Analysis

A.1 Additional Motivating Example (Smoking Cessation Dataset)

Our second illustrative NMA dataset compares smoking cessation strategies reported by the Agency for Health Care Policy and Research (AHCPR) Smoking Cessation Guidelines Panel, which has been analyzed by Lu and Ades (2006) among others. It consists of 24 studies to compare the relative effect of three treatments (B: self-help; C: individual counseling; and D: group counseling) vs. the baseline treatment (A: no contact). While we do not show these data explicitly, Figure A.1 gives the evidence network showing the connection between all treatment groups, where again the indices of trials having each comparison are shown in square brackets on the corresponding edge.

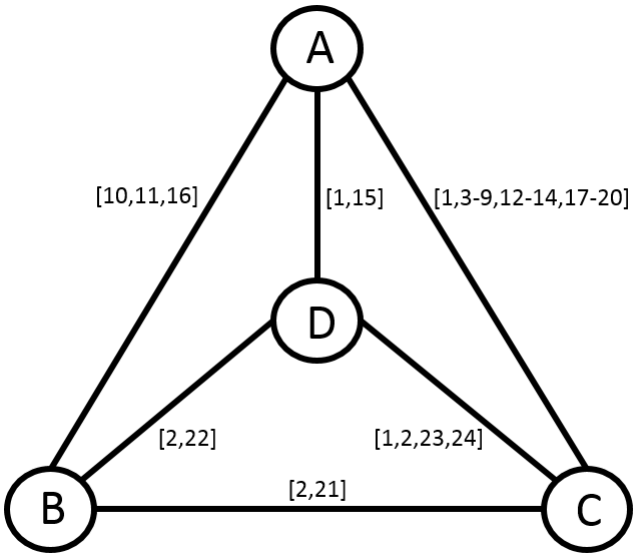

Figure A.1: Network for the smoking cessation dataset. Each vertex represents a treatment, and the indices of trials having each comparison are shown in square brackets on the corresponding edge.

We have applied the method described in Section 3.6.1 to calculate the discrepancy factors (w/o loop) for 6 different comparisons in the smoking cessation dataset: 1-2, 1-3, 1-4, 2-3, 2-4, and 3-4. As shown in Table A.1, none of these discrepancy factors is significantly different from 0 according to its 95% posterior CI. This result agrees with the conclusion by Lu and Ades (2006), finding no presence of serious inconsistency using the same dataset.

A.2 Simulation Studies

In Chapter 2, we have generated artificial data from inconsistency structures in datasets to evaluate the performance of inconsistency detection using our ABRE models, and compared the results to those of the loop-based method of Lu and Ades (2006). In that simulation, we used equal sample sizes $n_{ik} = 100$ for the k^{th} treatment arm in the

Table A.1: Discrepancy factors for the smoking cessation dataset using ABRE model (method described in Section 3.6.1: inconsistency detection in AB Models using fixed effects (w/o loop).

discrepancy factor	mean	sd	MC error	2.50%	median	97.50%
Δ_{12}	0.71	0.88	0.05	-1.29	0.76	2.30
Δ_{13}	0.67	0.70	0.04	-0.72	0.68	2.03
Δ_{14}	-0.81	1.96	0.13	-5.92	-0.58	2.49
Δ_{23}	0.70	1.01	0.05	-1.28	0.70	2.67
Δ_{24}	-0.20	0.90	0.05	-2.02	-0.20	1.63
Δ_{34}	0.92	0.80	0.04	-0.61	0.91	2.55

i^{th} study. Here, we want to consider two more complex settings, one using unbalanced sample sizes for each comparison, and another using asymmetric network structure. Under the same scenarios as in Chapter 2 (alternative and null), the new modified settings are as follows. First, for the unbalanced sample size setting, we assign the n_{ik} as 10, 160, 20, 300, 90, 20, 40, 60, 40, 10, 150, 10, 100, 25, 150, 20, 30, 40, 20, 30, 10, 200, 15, 120, 20, 300, 110, 50, 75, and 10 for the 6 pairwise comparisons (1-2, 1-3, 1-4, 2-3, 2-4, and 3-4), each having 5 simulated studies; see Figure 2.2 in Chapter 2. For the asymmetric network structure setting, we used the same simulation setting ($n_{ik} = 100$) as in Chapter 2, but deleting Trials 21 to 30; see Figure A.2. Since there is now only one loop in the network, we assign $w_{123} = 2.5$ or $w_{123} = 0$ for the alternative or null scenarios, respectively.

Our proposed methods using AB models (as in Section 3.6.1) were then applied to the two new settings. Tables A.2 and A.3 display the 2.5% and 97.5% posterior quantiles, and 95% Bayesian CI (BCI) widths for the discrepancy factors for each pairwise comparison, where each summary is the average of its quantity over the 1000 simulated datasets. In the first setting, the posterior mean of Δ_{34} for the alternative scenario is -3.04, with average 95% BCI $(-5.52, -0.58)$, indicating that these two sources of evidence for comparing Arms 3 and 4 are discrepant, and thus inconsistency exists.

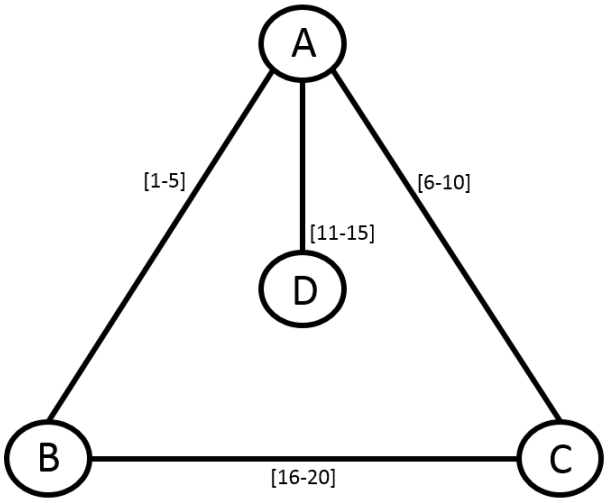

Figure A.2: Asymmetric network graph for additional simulation (Treatments A-D are referred to as Treatment 1-4 in Section A.2 for simplicity). Each vertex represents a treatment, and the indices of trials having each comparison are shown in square brackets on the corresponding edge.

A similar result is obtained in the second setting: the posterior mean of Δ_{23} for the alternative scenario is -2.94, with 95% BCI $(-5.55, -0.34)$, showing evidence of inconsistency in comparison 2 vs. 3. On the other hand, all other discrepancy factors are not that different from 0 when we set all $w = 0$, indicating no evidence of inconsistency for null datasets using our AB inconsistency detection approach. These results supplement our equal sample size findings in Chapter 2.

A.3 Comparison of Our Discrepancy Factor Method With Other Approaches

As mentioned in Chapter 2, there are many other approaches to detect inconsistency in NMA besides the Lu-Ades loop-based method. Here, we aim to present two main alternative approaches, and examine their relationship to our discrepancy factor approach.

Table A.2: Discrepancy factors for simulated datasets (unequal sample size) using ABRE model. Each entry represents the summary of results from 1000 datasets under each scenario.

w	discrepancy factor	mean	2.50%	97.50%	95% BCI width	Power
w[123]=0.01, w[124]=0.01, w[134]=2.5	Δ_{12}	-1.19	-3.60	1.22	4.82	0.13
	Δ_{13}	-1.39	-3.90	1.10	5.00	0.19
	Δ_{14}	-0.01	-2.48	2.46	4.94	0.04
	Δ_{23}	-0.97	-3.65	1.74	5.39	0.10
	Δ_{24}	0.43	-2.18	3.05	5.23	0.06
	Δ_{34}	**-3.04**	**-5.52**	**-0.58**	**4.94**	**0.66**
all w=0	Δ_{12}	-1.19	-3.58	1.20	4.78	0.14
	Δ_{13}	-1.38	-3.87	1.09	4.96	0.18
	Δ_{14}	-1.29	-3.70	1.10	4.80	0.18
	Δ_{23}	-0.95	-3.62	1.74	5.36	0.11
	Δ_{24}	-0.79	-3.40	1.82	5.22	0.09
	Δ_{34}	-0.52	-3.00	2.00	5.00	0.10

A.3.1 Node-splitting Method by Dias et al. (2010)

Dias et al. (2010) propose to detect inconsistency in NMA for certain treatment comparisons, not merely within a triangular loop. Their underlying model is a conventional CB model as described in Section 3.1, equation (3) of their main paper. Using their notation (changing their j to our i to keep the notation in this thesis consistent), it assumes:

$$logit(p_{ik}) = \begin{cases} \mu_{ib}, & k = b;\ b = A, B, C, ... \\ \mu_{ib} + \delta_{ibk}, & k > b;\ b = A, B, C, ... \end{cases} \tag{A.1}$$

where p_{ik} is the probability of an event in trial i and arm k, μ_{ib} is the log odds of success on trial i's baseline treatment b (which is not necessarily the same treatment for all trials), and δ_{ibk} is the log odds ratio of treatment k relative to b in trial i ($k > b$ indicates that k is after b in the alphabet). Study level effects μ_{ib} are treated as nuisance parameters unrelated to other model parameters. In detail, the trial-specific treatment effects δ_{ibk} follows a normal distribution $N(d_{bk}, \sigma^2_{bk})$, where one typically

Table A.3: Discrepancy factors for simulated datasets (asymmetric network structure) using ABRE model. Each entry represents the summary of results from 1000 datasets under each scenario.

w	Discrepancy factor	mean	95% BCI 2.50%	95% BCI 97.50%	95% BCI width	Power
w[123]=2.5	Δ_{12}	-1.09	-3.49	1.34	4.83	0.09
	Δ_{13}	1.42	-0.96	3.84	4.8	0.22
	Δ_{23}	**-2.94**	**-5.55**	**-0.34**	**5.21**	**0.6**
all w=0	Δ_{12}	-1.09	-3.47	1.32	4.79	0.09
	Δ_{13}	-1	-3.43	1.41	4.84	0.16
	Δ_{23}	**-0.51**	**-3.11**	**2.13**	**5.24**	**0.12**

assumes homogeneous variance $\sigma_{bk}^2 = \sigma^2$ for random effects.

To detect inconsistency using the node-splitting method, Dias et al. defines "node" as a particular treatment comparison, X vs. Y, and splits the information about that comparison (node) into direct and indirect evidence depending on whether the trials directly compare X and Y. Two posterior estimates are then obtained for the mean treatment effect d_{XY}: d_{XY}^{dir} using the direct information and d_{XY}^{ind} using all remaining evidence. Formally, the underlying CB model in (A.1) remains the same, and $\delta_{iXY} \sim N(d_{XY}^{dir}, \sigma^2)$ when node (X, Y) is being split. In the same model, the indirect evidence on comparison X vs. Y using the remaining studies is used to estimate d_{XY}^{ind}. Finally, the difference between the mean direct and indirect evidence is used to identify possible inconsistency in the given comparison.

Similar to this node-splitting approach, our discrepancy factor method also detects inconsistency in NMA for specific treatment comparisons, showing how these two sources of information differ. The idea of calculating the difference between the mean direct and indirect effects in our discrepancy factor approach has much in common with the node-splitting method, but the two methods are based on different model assumptions. Node-splitting is based on the CB model, which uses a nuisance parameter μ_{ib} (where

b refers to baseline) when estimating the mean treatment difference d_{bk} between the k^{th} trial and the baseline. If one wants to compare two treatments (e.g., Treatments E and F; neither a baseline), node-splitting uses $d_F - d_E$, the difference between their two relative effects, relative to a common reference group. Assuming the reference group is treatment A, then writing in terms of our AB model terminology, $d_F - d_E = [(\mu_F - \mu_A) - (\mu_E - \mu_A)] = \mu_F - \mu_E$, which is the same fixed-effect parameter comparison we use in our AB model (under the condition that $d_F = \mu_F - \mu_A$ is accepted). Another difference between our method and Dias' method is the grouping, since they divide the information into those including both X and Y and the rest, and we do not use the information from trials that include neither X nor Y (though we have applied the other grouping method using an example dataset and discussed it in Chapter 2).

A.3.2 Design-by-treatment Interaction Model of Jackson et al. (2014)

Jackson et al. (2014) introduced a model to detect inconsistency, which is a special case of Higgins' design-by-treatment interaction model, and a generalization of Lumley's model. The formal definition of "design" is given in Higgins et al. (2012) as "set of treatments compared within the study", and they group studies according to the design, in this sense. Using the notation in Jackson et al. (2014), if we want to estimate the treatment comparison between Arms A and J, the model assumes

$$\mu_{di}^{AJ} = \delta^{AJ} + \beta_{di}^{AJ} + \omega_d^{AJ},$$

where δ^{AJ} is a fixed effect of treatment J relative to A, β_{di}^{AJ} is a study-by-treatment interaction term to reflect what the authors call standard heterogeneity (i.e., unstructured variation between studies indexed by d for design and by i for study within design), and ω_d^{AJ} is a design-by-treatment interaction term to reflect inconsistency (variability of the treatment effect between designs subscripted by d). They further model between-study

heterogeneity using $\beta_{di} \sim N(0, \Sigma_\beta)$, and assume $\omega_d \sim N(0, \Sigma_\omega)$, whereas Higgins et al. assume the ω to be fixed effects.

In the above model, estimates of ω_d^{AJ} can be used to explore inconsistency in the network. As shown in Higgins (2012), the design-by-treatment interaction model can detect two different types of inconsistency: loop-based inconsistency and design inconsistency. Loop inconsistency is a restricted version of their full model (since the analysis for a given loop only considers some of the designs in the full network of trials, instead of all of them). Design inconsistency includes all variation in the design-by-treatment interaction that is not included in loop-based inconsistency. Therefore, the design-by-treatment interaction model has more degrees of freedom than any loop-based inconsistency model, and can detect inconsistency not only arising from discrepancy between direct and indirect information, but also from other sources of inconsistency. The extra complexity of the model results in loss in power in statistical tests, as described in Higgins et al. (2012).

Our discrepancy factor model, on the other hand, is constructed specifically to detect the discrepancy between the direct and indirect information for a particular comparison. By dividing all trials into groups based on whether they supply direct or only indirect evidence (or no apparent evidence, our group (iv)), we in effect introduce an interaction into our model that previously had only main effects. For each particular treatment comparison, we divide the studies into four groups of studies, which (in ANOVA terms) defines a new factor, and then compute the posterior of a particular contrast in this treatment-by-group interaction. The full interaction consists of that contrast plus enough other orthogonal contrasts to account for the degrees of freedom in the interaction. Our model amounts to setting those other contrasts to zero, so that in effect our model is a simplification of the full treatment-by-group interaction. Therefore, our approach (like node-splitting) is primarily a tool for seeking inconsistency, not

generally appropriate for making inference about pooled effects. By contrast, the approach of Jackson et al. (2014) offers a modeling framework to facilitate the ranking of treatments under inconsistency.

A.4 Application of Our Model in a Classical Analyses

In Chapter 2, our ideas are primarily applied in a Bayes-MCMC setting, to facilitate implementation. However, they can also be applied in a classical analyses, since equation (2.4) in Chapter 2

$$logit(p_{ik}) = \mu_k + \eta_{ik}$$

can be seen as a generalized linear mixed model $\eta = X\beta + Z\gamma$ (η here is the linear predictor g(E(y)); the link function is logit for binary outcomes), which we can write in its matrix form as follows:

$$logit(p_{ik}) = \begin{pmatrix} 1 & 0 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & \cdots & & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ & & & \vdots & & \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & & \cdots & & 0 & 1 \end{pmatrix}_{58\times 8} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_8 \end{pmatrix} + I_{58\times 58} \begin{pmatrix} \eta_{1,1} \\ \eta_{1,2} \\ \eta_{1,4} \\ \eta_{2,1} \\ \eta_{2,3} \\ \eta_{2,8} \\ \vdots \\ \eta_{28,3} \\ \eta_{28,8} \end{pmatrix},$$

where $X = \begin{pmatrix} 1 & 0 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & \cdots & & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ & & & \vdots & & \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & & \cdots & & 0 & 1 \end{pmatrix}_{58\times 8}$, $\beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_8 \end{pmatrix}$, $Z = I_{58\times 58}$ and $\gamma = \begin{pmatrix} \eta_{1,1} \\ \eta_{1,2} \\ \eta_{1,4} \\ \eta_{2,1} \\ \eta_{2,3} \\ \eta_{2,8} \\ \vdots \\ \eta_{28,3} \\ \eta_{28,8} \end{pmatrix}$,

and where we assume $\gamma \sim N(0, G)$, where $G = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{14} & 0 & & & \cdots & 0 \\ \sigma_{21} & \sigma_{22} & \sigma_{24} & 0 & & & \cdots & 0 \\ \sigma_{41} & \sigma_{42} & \sigma_{44} & 0 & & & \cdots & 0 \\ 0 & 0 & 0 & \sigma_{11} & \sigma_{13} & \sigma_{18} & \cdots & 0 \\ 0 & 0 & 0 & \sigma_{31} & \sigma_{33} & \sigma_{38} & \cdots & 0 \\ 0 & 0 & 0 & \sigma_{81} & \sigma_{83} & \sigma_{88} & \cdots & 0 \\ & & & \vdots & & & & \\ 0 & 0 & 0 & \cdots & & 0 & \sigma_{33} & \sigma_{38} \\ 0 & 0 & 0 & \cdots & & 0 & \sigma_{83} & \sigma_{88} \end{pmatrix}_{58\times 58}$.

Then restricted likelihood follows immediately from this specification.