Trinity College Dublin
The University of Dublin

The Impact of Performing a Network Meta-Analysis with Imperfect Evidence

A thesis submitted to the University of Dublin, Trinity College in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Statistics)

Joy Leahy

(under the supervision of Dr. Cathal Walsh)

March 2019

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Joy Leahy

Abstract

Network meta-analysis (NMA) is an important aspect of evidence synthesis in a clinical setting, as it allows us to compare treatments which may not have been analysed in the same trial. In an ideal scenario we would have a fully connected network of randomised controlled trials (RCTs) when undertaking an NMA. Ideally, these RCTs would contain the full patient population for a particular disease, and individual patient data (IPD) would be available for all trials. However, in reality we are never going to have all this information. Therefore, this thesis investigates methods for dealing with imperfect evidence. We consider two techniques for adjusting for confounding variables due to differing patient populations in a connected network. Firstly, we assess the benefit of the extra effort involved in obtaining and including IPD in an NMA. Secondly, we evaluate the impact of using IPD to adjust for differing trial populations through the increasingly popular method of matching adjusted indirect comparison. We also propose a method for including single-arm evidence in a disconnected network through aggregate level matching, and analyse the impact of this method. Although our work mainly focuses on the methodological aspects, all methods are illustrated using real world datasets, namely Hepatitis C virus (HCV) infection, melanoma and multiple myeloma.

Acknowledgements

This research was funded by the Health Research Board, Ireland, Grant Number HRB RL2013/4 (PI: Cathal Walsh). Some of the computations in this thesis made use of the statistics computing cluster which was supported by the STATICA project which was funded by the Principal Investigator programme of Science Foundation Ireland, Grant Number 08/IN.1/I1879 (PI: Simon P. Wilson).

Firstly, I would like to thank my supervisor Cathal Walsh for his help over these past years. His knowledge, insight, and motivation have not only guided my research greatly, but have also instilled a deep appreciation and passion for the wider subject area. His advice has been invaluable.

Thank you to Simon Wilson and all the lecturers in the statistics department for their help and support. I am particularly grateful to Arthur White, who selflessly invested his time, wisdom, and expertise at crucial times, and for that I will always be grateful. Thank you to Michael Walsh, and all those who provide technical support in the school of computer science and statistics. I would also like to express my gratitude to Susanne Schmitz, although our paths barely crossed in Trinity, she has given me so much support and advice over the years.

I am grateful to Howard Thom and Jeroen Jansen for inviting me to work with them on our ISPOR workshop. It was a great honour to have such intelligent and knowledgeable collaborators.

I wish to thank all the staff in the NCPE, especially Aisling O'Leary, Emma Gray, and Claire Gorry. They provided me with the opportunity to apply my methods to real world datasets, but more importantly, they have all been such great fun to work with.

Thank you to all the postgrads in the statistics department, especially Angela McCourt, Donnacha Bolger, Shane O'Meachair, and Asmaa Alghamdi. They have all made the last few years so enjoyable. I also wish to thank Owen Cassidy and Laure Ngouanfo for their help in proofreading. Thank you also to Suzy Whorisky and Eilidh Jack, with whom I made such great friends on the APTS courses. Thank you for having me over to Glasgow so many times!

Thank you to all my other collaborators: thank you to Nezam Afdhal, Scott Milligan, and Malte Wehmeyer for making our IPD project such a pleasant experience. Thank you to my colleagues in NUIG: Chris Noone, Eimear Morrissey, John Newell, and Gerry Molloy. I've really enjoyed seeing NMA from a psychology perspective!

Finally, I could not have completed this thesis without the help and support of my family and friends. Thank you to my cousin Siun Tobin for always being available to answer my pharmaceutical questions! Thank you to my parents, for all their help and support over the years. I would like to particularly thank my Dad. I am truly grateful for all the guidance and encouragement he has given me over the years. For his maths puzzles growing up, to helping me with my maths homework when I was stuck, to listening to my talks before conferences! Finally, my biggest thank you goes to my husband Conor. I consider myself truly lucky to have found a partner who is always there for me, who is always willing to listen to my stories at the end of the day, and whose intelligence and unique way of looking at the world have been a great motivation throughout this thesis. Conor, thank for all your love and encouragement over the years.

Contents

List of Figures

List of Tables

Table 1: Glossary of commonly used acronyms used throughout this thesis

AgD	Aggregate Data
AIC	Akaike Information Criterion
ASCT	Autologous Stem Cell Transplant
CrI	Credible Interval
CR	Complete Response
DAG	Directed Acyclic Graph
DIC	Deviance Information Criterion
ECOG	Eastern Cooperative Oncology Group
FE	Fixed Effects
GT1	Genotype 1
HCV	Hepatitis C Virus
HR	Hazard Ratio
IPD	Individual Patient Data
ISS	International Staging System
KM	Kaplan Meier
LDH	Lactate Dehydrogenase
LOR	Log Odds Ratio
MA	Meta-Analysis
MAIC	Matching Adjusted Indirect Comparison
MAE	Mean Absolute Error
MC Error	Monte Carlo Error
MCMC	Markov Chain Monte Carlo
NCPE	National Center for Pharmacoeconomics
ndMM	Multiple Myeloma in newly diagnosed patients
NICE	National Institute for Health and Care Excellence
NMA	Network Meta-Analysis
OS	Overall Survival
PFS	Progression-free survival
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomised Controlled Trials
RE	Random Effects
SD	Standard Deviation
STC	Simulated Treatment Comparisons
SUCRA	SUrface under the Cumulative RAnking curve
SVR	Sustained Virological Response
ULN	Upper Limit of Normal
VGPR	Very Good Partial Response
WHO	World Health Organisation

Table 2: Glossary of mathematical notation for Chapter 3

Mathematical Notation	**Explanation**
r_{ij}	Number of events in the j^{th} arm of the i^{th} trial. This is the number of patients who achieved sustained virological response (SVR) in the case of HCV infection.
p_{ij}	Probability of an event in the j^{th} arm of the i^{th} trial.
x_{ij}	Proportion of patients possessing the binary characteristic in the j^{th} arm of the i^{th} trial in the case of one covariate.
$x_{m_{ij}}$	Proportion of patients possessing the binary characteristic m in the j^{th} arm of the i^{th} trial.
β	Coefficient of effect of the covariate when only one covariate is considered.
β_m	Coefficient of effect of covariate m.
μ_i	Baseline risk of having an event in study i.
δ_{ij}	Study-specific treatment effect in the j^{th} arm of the i^{th} trial.
$d_{t_{ij}}$	True effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment.
d_k	True effect of treatment k relative to the reference treatment.
$d_{\text{RCT}[k]}$	True effect of treatment k relative to the reference treatment for RCT studies.
$d_{\text{MATCHED}[k]}$	True effect of treatment k relative to the reference treatment for matched studies.
μ_i	log odds of having an event in the baseline treatment (i.e. the treatment in arm one) for study i. This parameter is considered in the RCT only model, pooled model, and hierarchical model.
ν_i	log odds of having an event in treatment one in trial i. This parameter is considered in the plug-in estimator model.
σ_μ	Standard deviation of the baseline study effect, i.e. variability of the log odds of having an event in baseline treatments (i.e. treatment in arm one) across studies.
σ_δ	Standard deviation of study-specific treatment effects.
σ_{des}	Standard deviation of the between-study design effect.
ω	The variance inflation on the matched evidence.
ξ	Bias in single-arm trials.
T	Number of simulations.
M	Number of identified covariates.
$H_{s,k}$	Cumulative hazard for treatment k in study s in survival model.
d_b	True effect of the baseline treatment in survival model.
$S_{ij}(t)$	Survival function for the j^{th} arm of the i^{th} trial at time t in the survival model.

Table 3: Glossary of mathematical notation for Chapter 4

Mathematical Notation	**Explanation**
r_{ij}	Number of events in the j^{th} arm of the i^{th} trial. This is the number of patients who achieved SVR in the case of HCV infection.
r_{ijl}	Binary outcome for the l^{th} patient in the j^{th} arm of the i^{th} trial. It is whether or not a patient attained SVR in the case of HCV infection.
p_{ij}	Probability of an event in the j^{th} arm of the i^{th} trial.
p_{ijl}	Probability of an event for the individual patient l in the j^{th} arm of the i^{th} trial.
x_{ij}	Proportion of patients possessing the characteristic in the case of a binary covariate, or mean of the continuous covariate, in the j^{th} arm of the i^{th} trial.
x_{ijl}	Indicator variable for the presence of the characteristic associated with the covariate for patient l in the j^{th} arm of the i^{th} trial.
β	Effect of the covariate.
β_k	Effect of the covariate interacting with treatment k when interactions are considered independent or exchangeable.
$\beta_{t_{ij}}$	Effect of the covariate interacting with the treatment in the j^{th} arm of the i^{th} trial when interactions are considered independent or exchangeable.
μ_β	Mean of covariate effect across treatments when interactions are considered exchangeable.
σ_β	Standard deviation of covariate effect across treatments when interactions are considered exchangeable.
μ_i	Baseline risk in trial i.
δ_{ij}	Study-specific treatment effect in the j^{th} arm of the i^{th} trial.
σ_δ	Standard deviation of study-specific treatment effects.
d_k	True effect of treatment k relative to the reference treatment.
$d_{t_{ij}}$	True effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment.
S	Number of simulations.

Table 4: Glossary of mathematical notation for Chapter 5

Mathematical Notation	**Explanation**
AB-IPD trial	Trial with individual patient data comparing treatment A to treatment B.
BC-AgD trial	Trial with aggregate data only comparing treatment B to treatment C.
$Y_{lk(AB)}$	Outcome for patient l on treatment k in the AB trial.
$N_{k(AB)}$	Number of patients assigned to treatment k in the AB trial.
w_{lk}	Weight assigned to the patient l receiving treatment k.
T_{ijl}	time-to-event for individual l, in study i and arm j.
C_{ijl}	Censored time for individual l, in study i and arm j.
$\beta_{m,t_{ij}}$	Coefficient for the effect of covariate m for the treatment in arm j of study i, relative to treatment 1, when the covariate is an effect modifier.
β_m	Effect of covariate m for all treatments when the covariate is a prognostic variable.
$x_{m,ijl}$	Binary indicator for the presence of the characteristic m for patient l in arm j of study i.
$x_{m_1m_2}$	Proportion of the population possessing the relevant characteristics in the case of two covariates.
μ_i	Baseline risk in trial i.
δ_{ij}	Study-specific treatment effect in the j^{th} arm of the i^{th} trial.
d_k	True effect of treatment k relative to the reference treatment.
$d_{t_{ij}}$	True effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment.
d_b	True effect of the reference treatment.
H_{ik}	Hazard ratio of treatment k versus the baseline treatment in study i.
$re_{i,k}$	Random effect deviation for arm k of study i.
$S_{ij}(t)$	Survival function for the j^{th} arm of the i^{th} trial at time t.
β_{0_i}	Covariate effect in trial i when using an IPD model.
σ_δ	Standard deviation of study-specific treatment effects.
Q	Number of simulations.
M	Number of covariates.

Chapter 1

Introduction

Network meta-analysis (NMA) is a method for comparing multiple treatment interventions simultaneously in a single analysis, by combining direct and indirect evidence within a network (Rouse et al. (2017)). NMA has the potential to be the highest level of evidence. However, it is crucial that there is clear and robust methodological techniques and high reporting standards (Faltinsen et al. (2018)). Pharmaceutical companies stand to make tremendous amounts of money if their treatments are routinely used to treat a particular illness. Therefore, pharmaceutical companies put a lot of time and effort into running randomised controlled trials (RCTs) to show that their treatments are safe and effective. However, as we will discuss in Chapter 2, there are multiple reasons why trials cannot give a perfect estimate of efficacy of a treatment, no matter how much investigators try to minimise bias. On top of this, we must be cognizant of the incentive for pharmaceutical companies to show their treatments in the best light. Therefore, the aim of this thesis is to contribute to the statistical methodology behind NMA, so that all analysis can be carried out in the most objective way possible. This is done through proposing new methods, assessing new and existing methods through simulation studies, and examining the implementation of these methods in a number of different disease areas.

NMA is particularly useful when attempting to compare the efficacy of different treatment options. For most diseases there are many different drugs that can be used to treat them. Therefore, it is crucial that we are able to compare available treatments. For patients it is important that doctors are able to identify the best treatment for them. For governments it is important to be able to assess new treatments in comparison to other treatments, in order to decide whether they are cost effective. For many illnesses it may be necessary to treat the patient with multiple different treatments over the lifetime of the illness. Patients may have negative reactions to a specific drug or in some illnesses, such as cancer, the disease may build up a resistance to a drug and therefore alternative treatment options are needed. This emphasises the importance of not just identifying the best treatment, but being able to rank all treatment options.

NMA is routinely used in Health Technology Assessments (HTAs), whereby HTA agencies compare the efficacy of one treatment relative to its competitors, in order to assess whether the government in that country should agree to pay the price being offered for that treatment. Ideally, an NMA should be as informative as possible and consist of high quality RCTs. However, HTA agencies are increasingly being asked to assess NMAs comprised of poorer quality evidence, coupled with increasingly complex methodology to compensate. This thesis aims to investigate some of this methodology, and to determine whether or not it is possible to obtain an improved or even sufficiently accurate estimate using these techniques.

1.1 Outline of Chapters

- Chapter 2 - Background: This chapter introduces the concepts contained in this thesis and provides some background. Various types of trials are discussed, along with the motivation for NMA and some possible complications that arise. We then introduce Bayesian methods and their application to NMA.

- Chapter 3 - Aggregate Level Matching of Single-Arm Evidence: Single-armed evidence is increasingly being used to demonstrate the efficacy of treatments. Although it is recognised that RCTs provide a higher standard of evidence, these are not available for many new agents which have been granted licences in recent years. Therefore, it is important to examine whether alternative strategies for assessing this evidence may be used.

  While it is important to assess the potential benefit of including this form of evidence, we must be mindful that this has the potential to increase bias. Nevertheless, given that this information is being used in practice, in this chapter we aim to quantify the potential bias, provide guidance to identify situations where this method may or may not be appropriate, and provide a clear method for including this type of evidence, which is backed up by systematic investigation. In this chapter we examine approaches to incorporate single-armed evidence formally in the evaluation process. We consider matching aggregate level covariates to comparator arms or trials, and including this evidence in an NMA. We consider two methods of matching:

  1. We include the chosen matched arm in the dataset itself as a comparator

for the single-arm trial,

2. We use the baseline odds of an event in a chosen matched trial to use as a plug-in estimator for the single-arm trial.

We illustrate that the syntheses of evidence resulting from such a setup is sensitive to the between-study variability, formulation of the prior for the between-study design effect, the weight given to the single-arm evidence, and the extent of the bias in single-armed evidence. We provide a flow chart for the process involved in such a synthesis, and highlight additional sensitivity analyses that should be carried out. This work was motivated by a Hepatitis C virus (HCV) infection dataset where many agents have only been examined in single-arm studies. We present the results of our methods applied to this dataset and to two melanoma networks.

- Chapter 4 - Individual Patient Data: This chapter assesses the impact of incorporating individual patient data (IPD) into an NMA where possible, and using aggregate data (AgD) for the remaining studies in the network. The use of IPD in NMA is becoming increasingly popular. However, as most studies do not report IPD, most NMAs are carried out using AgD for at least some, if not all, of the studies. We investigate the benefits of including varying proportions of IPD studies in an NMA.

  Several models have previously been developed for including both AgD and IPD in the same NMA. We carried out a simulation study based on these models to examine the impact of additional IPD studies on the accuracy and precision of the estimates of both the treatment effect and the covariate effect. We also compared the Deviance Information Criterion (DIC) between models to assess model fit. An increased proportion of IPD resulted in more accurate and precise estimates for most models and datasets. However, in certain scenarios coverage probability decreased, particularly when the model was misspecified. The use of IPD leads to greater differences in DIC, which allows us choose the correct model more often.

  We analysed a HCV infection network consisting of three IPD observational studies. The ranking of treatments remained the same for all models and datasets. We observed similar results to the simulation study: the use of IPD leads to differences in DIC and more precise estimates for the covariate effect. However, IPD sometimes increased the posterior SD of the treatment effect estimate, which may indicate between-study variability. We recommend that IPD should be used where feasible, especially for assessing model fit. If a researcher has access to IPD, it should be included in an NMA. If a researcher has no IPD or limited IPD then the benefit and cost of obtaining additional IPD needs to be assessed in the context of the extra time and effort required. We provide a framework for assessing the benefit of collecting this additional IPD.

- Chapter 5 - Matching Adjusted Indirect Comparison (MAIC): Although IPD is the highest form of evidence to consider in an NMA, the situation often arises where a researcher has IPD for trials concerning a particular treatment (for example from a sponsor), but none for other trials. Therefore, one could reweight the IPD so that the covariate characteristics in the IPD trials match that of the AgD trials, using an MAIC.

  We assess the impact of using the reweighted aggregated data, obtained by the MAIC, in a Bayesian NMA for a connected treatment network. We compare the performance of this method to the standard NMA model and a mixed AgD/IPD NMA model, similar to the models outlined in Chapter 4. We apply this method to a network of multiple myeloma treatments in newly diagnosed patients (ndMM), where the outcome is progression-free survival. We investigate the reliability of our methods and results through a simulation study which mirrors the ndMM network. The ndMM network consists of three IPD studies comparing lenalidomide to placebo (Len-Placebo), one AgD study comparing Len-Placebo, and one AgD study comparing thalidomide to placebo (Thal-Placebo). We therefore investigate two options of weighting the covariates:

  1. The IPD within each of the three studies are reweighted such that the AgD of each reweighted study matches the AgD of the Thal-Placebo trial.
  2. All three IPD studies are pooled together for the reweighting element, such that the AgD of the three studies combined matches the AgD of the Thal-Placebo trial, but they are put into the NMA as three separate trials. Note that this is a less stringent requirement as it generally involves less reweighting.

  We observe limited benefit to MAIC in the full network population. While MAIC can be beneficial as a sensitivity analysis to confirm results across patient populations, we advise that MAIC is used and interpreted with caution. We recommend the either a standard NMA model or the mixed AgD/IPD model over an MAIC for the base case analysis.

- Chapter 6 gives a summary of the work in this thesis and gives some suggestions for future work. We discuss the importance of running RCTs, and the potential bias that can arise from statistical methods which deal with imperfect data. We also discuss types of evidence for which we are most likely to be able to obtain IPD.

1.2 Research Contributions

The novel contributions of this work are as follows:

- In Chaper 3:
  - Novel methods for including single-arm studies in an NMA are proposed and analysed through simulation studies in Section 3.2.1.
  - We suggest a number of sensitivity analyses that can be undertaken when including single arm evidence.
  - We provide an algorithm for deciding whether or not to include single arm evidence in Figure 3.1.
- In Chaper 4:
  - We quantify the impact of IPD in an NMA by performing a simulation study.
  - We assess how much IPD can help to assess how a covariate affects different treatments equally (ie treatment-covariate interaction).
  - We present the consequences of making an incorrect assumption about the nature of the treatment-covariate interaction.
- In Chaper 5:
  - We undertake a simulation study to compare MAIC to a standard NMA or a standard NMA using a covariate term.
  - We assess how to best reweight three IPD trials by using either an MAIC pooled trials method or MAIC separate trials method.
  - We compare how MAIC affects hazard ratio model and median models.
- Throughout the thesis we provide examples of how to quantify unknown and important real-world issues using synthesised evidence.

Chapter 2

Background

2.1 Meta-Analysis (MA)

Meta-analysis is a statistical technique for combining the results of trials which compare the same two interventions. When there are multiple trials assessing two interventions there can be differences in the results of these trials (Krauss (2018), Welton et al. (2012)). For example, this could be because each trial was carried out only on a subset of the population, who each react to the intervention in slightly different ways (Rothwell (2005)). For instance, let's assume we are comparing an active treatment to a placebo. If a trial consists of relatively healthy patients then many may get better, regardless of which treatment they are given. However, if the trial consists of sicker patients, they may respond only if they receive the active treatment.

A second reason could be because of differences in the drug administration in each trial. For example, some antibiotics can be given both orally or through intravenous therapy (IV), which can impact on how effective the treatment is relative to the competing treatment. Furthermore, some trials may give differing dosages of the same treatment, which can also affect the outcome.

A third complication that may affect the results of trials is how the outcome is measured. For example, blood pressure can be affected by the time of day that the reading is taken at, or whether the patient is sitting or active when the reading is taken. As the method of measuring blood pressure can vary from trial to trial, this can have an impact on the efficacy of one treatment relative to another.

These variations between studies are referred to as the *heterogeneity* between studies. When combining all relevant clinical trials, most of these trials will give us some information about the efficacy of the treatments involved, but no single trial will give us perfect information. The idea is that if we analyse the results of all these trials in one meta-analysis we get the best possible estimate of the true treatment effect (Haidich (2010)).

Another reason why meta-analysis is so important is because smaller trials on their own may not have the power to detect differences between treatments. A meta-analysis of multiple trials has larger power to detect differences and ensure that the best possible treatment is given to all patients.

2.2 Network Meta-Analysis (NMA)

NMA is a method for assessing the entire evidence base for a particular disease when three or more treatment options are available (Caldwell et al. (2005)). It is an extension of traditional pair-wise Meta-Analysis (MA), as described in Section 2.1 which is used when directly evaluating two treatment options. NMA combines direct and indirect evidence to obtain effect estimates by comparing all treatments against all other treatments in the network. Much research has been carried out on its benefits and potential drawbacks (Lumley (2002), Lu & Ades (2004), Cooper et al. (2011), Senn et al. (2013), Salanti et al. (2008), Song et al. (2012), Jansen & Naci (2013), Dias et al. (2018)).

As there are a large number of treatments available for some diseases it is not practical to assess all these treatments in one trial. If we have two competing treatments they may often be compared to a placebo. However, decision makers may be mainly concerned with evaluating those two treatments relative to one another. Therefore, we want to use indirect comparisons to compare these two treatments. Treatments are often tested in placebo controlled trials. However, when a treatment is known to be effective and is routinely used in clinical practice, it could be considered unethical to test a new potential comparator against a placebo as opposed to the known effective treatment. These newer treatments are often assessed versus the best available or the standard care. Therefore we can have a network of treatments with many connections between treatments. Even when direct evidence is available as a comparison for certain treatments it can strengthen the evidence to use indirect evidence as well (Welton et al. (2012)).

Before an NMA can be performed the relevant studies need to be identified. This is done by a *systematic review*. This is defined as “a review of the evidence on a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyse data from the studies that are included in the review” (Khan et al. (2001)). The systematic reviews in this thesis are carried out in accordance with the criteria of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses group (PRISMA) (Moher et al. (2009)).

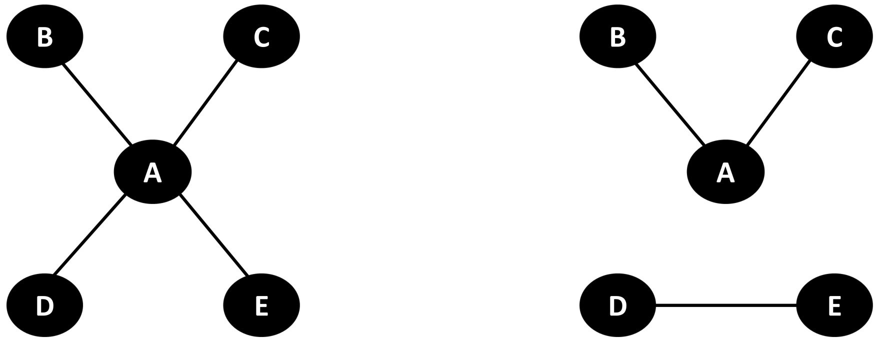
According to O'Rourke (2007), the first meta-analysis looked at typhoid inoculation and was published in 1904 by Simpson and Pearson (Simpson & Pearson (1904)). Higgins & Whitehead (1996) and Bucher et al. (1997) were two of the earliest meta analyses to also include an indirect treatment comparison. In 2004 Lu and Ades used Bayesian methods for NMA (Lu & Ades (2004)). One area where NMA is used is HTA. In a HTA a treatment is assessed to see if it is cost effective. Many countries use HTA in deciding whether or not to fund new treatments, for example the National Centre for Pharmacoeconomics (NCPE) in Ireland and the National Institute for Health and Care Excellence (NICE) in England and Wales. Therefore, it is crucial to be able to quantify the efficacy of any treatment relative to its competitors. The World Health Organisation (WHO) have also recently started using NMA (Kanters et al. (2016)).

The term network meta-analysis was first used in 2002 and is inspired by a graph illustrating which treatments have been compared directly with each other within primary studies (Higgins & Welton (2015)). In an NMA we want to compare all possible treatment regimens so it is necessary to form a connected network. Figure 2.1 shows two examples of networks. The nodes represent the treatments and the edges indicate whether these treatments have been compared in the same study. For example, the edge between treatment A and treatment B indicates that there is a study in our network which has an arm with treatment A and an arm with treatment B, i.e., some patients were assigned treatment A and others were assigned treatment B. Figure 2.1a shows a connected network. We can, therefore, compare any two treatments even if they are not in the same study. For example, although treatment B and treatment D have never been compared directly in a study we can still assess the relative efficacy of these treatments as they have a common comparator in treatment A. However, Figure 2.1b is not a connected network. There is no way of comparing treatment B and treatment D as they have no common comparator.

2.3 Types of Evidence

2.3.1 Observational Evidence

There are many different ways that evidence can be recorded to enable us to compare treatments. A simple way to do this is to keep a database of any treatment that people are taking. This is a form of *observational evidence*. However, this evidence is potentially biased, and we need to be cautious about being over-reliant on this form of evidence (Grieve et al. (2016)). We can look at treatments for



(a) Connected Network Example *(b) Disconnected Network Example*

Figure 2.1: Examples of Networks

cancers, as an example. Let's say there is a promising new treatment available. Doctors may be more likely to prescribe this treatment to particularly sick patients, for example. This may be because either they want to give the best treatments to the sickest patients, or they prefer to have exhausted all other tried and tested treatment options before giving more experimental newer treatments. In this case the new treatment may end up looking less effective than it actually is as it has been given to patients who were less likely to respond to treatments. For this and for other reasons, it is difficult to obtain an unbiased estimate of the treatment effect through this type of evidence.

2.3.2 Randomised Control Trials (RCTs)

For the reasons described in Section 2.3.1, it is important to run *Randomised Control Trials (RCTs).* In an RCT a patient is randomly assigned to a particular treatment. The idea is that both treatments will be given to a similar cohort of patients overall, which allows us to compare like with like. Without randomisation, bias could seep into a trial, for example, through unconscious decisions by the investigator, such as assigning the treatment under investigation to healthier patients.

The highest quality RCTs will be doubly blinded RCTs, which means that neither the investigator of the trial nor the patient knows to which treatment arm the patient has been assigned. Without blinding there are a number of ways in which bias can be introduced in a trial. Patients who have been assigned to the seemingly inferior treatment may be less hopeful about the trial and this may affect their adherence to the medication or their interest in their general well-being. On the other hand, if investigators know which treatments the patients are on, they may be expecting more positive results from the treatment that they consider to be better, and therefore they may be more likely to record a positive outcome for patients on these treatments. For further discussion on RCTs see Akobeng (2005).

2.3.3 RCTs versus Observational Evidence

Although there are clear reasons why RCTs are the gold standard of evidence when estimating relative treatment effect, it may be beneficial to consider other forms of evidence as well, especially when it comes to external validity. In fact, Faraoni & Schaefer (2016) suggest that "meta-analyses using both RCT and observational studies should be used to highlight some questions that neither a RCT, nor an observational study would have the ability to solve by themselves." A number of authors have reviewed the benefits and weaknesses of observational evidence versus RCTs (Shrier et al. (2007), Kunz & Oxman (1998), Ioannidis et al. (2001)). Firstly, RCTs may lack generalisability. The specific selection criteria for RCTs may limit how well results extend to the patient population as a whole. For example, RCTs rarely allow children or pregnant women to be included, and therefore we cannot know for certain if the results of the RCT extends to these groups of patients. Secondly, RCTs can lack real world applicability. For example, patients often tend to take particularly good care of themselves when participating in an RCT, so benefits may be over-estimated.

Even if we consider RCTs to be the highest standard of evidence there may be other reasons for including lesser quality evidence. In some cases, if we lack sufficient data, the limited RCT estimates available can mean that we are not able to detect all true differences between treatments. While we may get estimates for all relative treatment effects, the credible intervals associated with these estimates may be very wide, which can lead us to be unable to make recommendations for clinical practice based on the results. In this case including observational evidence such as registry data can allow us to produce more precise estimates, even if we have to be a bit more cautious in relying on the results, due to potential bias.

A further reason for using non-randomised evidence is that we may not be able to compare all treatments using only the RCT evidence available. Take, for example, the disconnected network in Figure 2.1b. When faced with this type of network it may not be possible to carry out a trial which connects the network. Sometimes it may be unethical, for example, comparing a blood transfusion with no transfusion in the case of critical haemorrhage (Trentino et al. (2016)). If there was extra observational evidence available comparing, for example, treatment D and treatment A, then we could use this evidence to connect the network, thus allowing us to draw a comparison between any two treatments in the network, although once again we may need to be more cautious in interpreting the results.

2.3.4 Single-Arm Evidence

A final type of evidence to be considered is that of *single-arm evidence*. This is a type of observational evidence that has no comparator at all. There are a number of reasons why this type of evidence is available (Griffiths et al. (2017)). It can come from Phase I or Phase II studies, when the pharmaceutical company was first gathering evidence on whether the treatment worked, before the additional expense of a more involved RCT. It can also come in the form of Phase IV studies for ongoing monitoring, where the treatment is being used in practice. Sometimes no other treatment may be available for a particular disease and it may be unethical to run a comparative trial where some patients are given a placebo. Another possible reason is because a disease may be so rare that it is not feasible to run a trial where patients are randomised to different arms.

This type of evidence is even more prone to the biases of comparative observational evidence. When examining observational evidence with multiple treatments we cannot know whether there was bias in treatment assignment, and therefore we are unsure how appropriate it is to compare treatment arms. Single-arm evidence has become very common in cancer trials. However, with this type of evidence, we do not even have another arm with which to compare the evidence, so it is very difficult to infer how patients would have responded to a different treatment (Twombly (2006)). Nonetheless, we can sometimes find ourselves in the situation of not having any other evidence for evaluating a specific treatment. In this instance we need to make a judgment call as to whether to incorporate this type of evidence or not. The work in Chapter 3 is designed to aid in this decision making process by identifying circumstances when it may or may be not appropriate to consider this form of evidence. We also examine sensitivity analyses that can be undertaken to check the effect and the appropriateness of including single-arm evidence.

2.4 Adjusting for Bias

All trials, regardless of whether they are RCTs or observational studies can be prone to some forms of bias. In general RCTs should be free from internal biases, as both groups are treated in the same way. However, these trials may be subject to external bias. This may be because patients in the trial behave differently to how they would under regular clinical conditions, or as discussed previously, the patients are from a particular subset of the population. Turner et al. (2009) provide a method for eliciting, quanitifying and adjusting for bias, by constructing a prior distribution to represent the biases in each study. They build on work from Eddy et al. (1992) and Wolpert et al. (2004) by modelling bias from each trial individuallu, but also assume a direct form of bias in the target parameter, following Spiegelhalter & Best (2003). Welton, Ades, Carlin, Altman & Sterne (2009) also provide models for adjusting for bias. They find that in terms of precision, there is little benefit from including studies at high risk of bias, regardless of how large these studies are.

2.5 Individual Patient Data (IPD)

When carrying out an NMA we usually take results from the published literature. In the vast majority of cases only summary statistics (aggregate data (AgD)) are published. Therefore, most NMAs can only be carried out on AgD. For continuous outcomes an example would be the mean and standard deviation of the change reported in each group. In other circumstances a continuous outcome can be represented as a time-to-event, in which case we could use medians or hazard ratios (HRs). For binary outcomes it could be the number of people who have died or have been cured. By simply using AgD we will be losing some information. For example, we would not be able to analyse differing patient characteristics to see which characteristics may affect outcomes on various treatments.

For these reasons, an NMA using IPD for each trial is considered a more informative analysis. However, having a full network of IPD trials would be quite rare. In some cases, however, IPD may be available for a subset of trials in the network. Therefore, Donegan et al. (2013), Saramago et al. (2012), Hong et al. (2018), and Thom et al. (2015) have worked on methods for including both aggregate data and IPD in a Network Meta-Analysis. When a researcher has limited access to IPD it may be possible to obtain IPD from other sources. However, this can be a long and time consuming process. Attempting to estimate the additional benefit that IPD may contain can aid in considering whether or not to pursue this data. In Chapter 4 we analyse the strengths and weaknesses of including IPD.

2.6 Matching Adjusted Indirect Comparison (MAIC)

Sometimes the situation arises where there is IPD available for trials concerning a particular treatment, but only AgD is available for the other trials. This could be because investigators working in a pharmaceutical company or in collaboration with a pharmaceutical company may have access to the IPD from the trials which were carried out by that company, which of course will be focused on their own treatment. In this scenario the investigator can carry out an MAIC, by reweighting patients in the IPD trial so that the covariate make-up of the IPD trial matches that of the AgD trial (Signorovitch et al. (2010, 2012)). There are two main types of MAIC; anchored or unanchored. An anchored MAIC means that there is a common comparator in each trial, i.e. we have a connected network. An unanchored MAIC means that we have a disconnected network (Phillippo et al. (2016)). In this thesis we focus on an anchored MAIC only, but note that we could consider the techniques of an unanchored MAIC as an extension to Chapter 3.

There are numerous reasons why it may be desirable to carry out an MAIC. In some cases a pharmaceutical company may wish to assess the results of their trial in their competitor's trial population. This may be the case where a company is carrying out a HTA for the country in which their competitor's trial was carried out. Secondly, a company may wish to show that their trial results would hold in varying patient populations. Thirdly, a company may be aware that there are differences between effect modifiers between both trial populations, and may want to adjust for these so as to guard against inconsistency and make a fairer comparison with their competitor.

MAIC is an increasingly popular method, but as it is a post hoc analysis of data it can be more susceptible to bias than other methods. It therefore requires rigorous investigation. We therefore examine the strengths and weaknesses of an anchored MAIC in Chapter 5.

2.7 Fixed Effects versus Random Effects

The technical question of Fixed Effects (FE) versus Random Effects (RE) is often an issue in MA and NMA, which is discussed in Green & Higgins (2005). FE assumes that the true treatment effect is constant across trials and that differences in the results are due to sampling variability. Random Effects (RE) assumes that trial-specific treatment effects are not the same but come from a common distribution (Borenstein et al. (2010)).

For example, looking at the average test scores across schools, a fixed effects model assumes that the test scores will be the same across schools. However, a random effects model assumes that test scores will vary from school to school and, therefore, typically puts more weight on smaller studies when calculating the overall test scores of the population, i.e. 10 people surveyed from a new school would be considered more informative than 10 extra people in a school from where data has already been collected. FE models require a much more stringent assumption than RE models and therefore should be used with caution as they could potentially overestimate the certainty of any findings (Hunter & Schmidt (2000)).

Given that there can often be quite a lot of heterogeneity between trials in an NMA, it is important to assess whether an FE or RE analysis is more appropriate. The degree of heterogeneity can be quantified using statistical tests. For example, the I^2 statistic indicates the variation across studies that is attributable to genuine heterogeneity rather than chance (Higgins & Thompson (2002)). However, the heterogeneity should also be assessed by inspecting the difference between the studies themselves. If there is a concern that an FE model is too restrictive, then an RE model may be more appropriate (Welton et al. (2012)).

2.8 Types of Outcomes

2.8.1 Binary Outcomes

A typical NMA model for binary outcomes is constructed as follows. We observe r_{ij}, the number of events in the j^{th} arm of the i^{th} trial. Examples of events could be deaths or cures. Treatment 1 is considered the overall reference treatment, with all other treatments being compared to it. $r_{ij} \sim \text{Bin}(p_{ij}, n_{ij})$, where p_{ij} is the probability of the binary event in the j^{th} arm of the i^{th} trial and n_{ij} is the number of patients in the j^{th} arm of the i^{th} trial. n_{ij} is a fixed, observed quantity and p_{ij} is made up of δ_{ij}, the treatment effect in the j^{th} arm of the i^{th} trial and μ_i, the log odds of having an event in the baseline treatment (i.e. in arm one) for study i. We assume an RE model, $\delta_{ij} \sim N(d_{t_{ij}} - d_{t_{i1}}, \sigma_\delta^2)$, where $d_{t_{ij}}$ denotes the effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment and σ_δ represents the between trial standard deviation (SD) of the treatment effect. Let d_k denote the effect of treatment k relative to the reference treatment.

The model is written as:

$$\text{logit}(p_{ij}) = \begin{cases} \mu_i & \text{if j=1} \\ \mu_i + \delta_{ij} & \text{if j>1} \end{cases}, \tag{2.1}$$

Thus, δ_{ij} can be considered as a random, within study effect of the treatment, centred on the underlying mean value of the treatment, $d_{t_{ij}} - d_{t_{i1}}$. In this thesis the assumed priors are $\mu_i \sim N(0, 1.83^2)$, $d_k \sim N(0, 1.83^2)$. $\sigma_\delta \sim \text{Unif}(0, 2)$ in Chapter 3 and $\sigma_\delta \sim \text{Unif}(0, 5)$ in Chapters 4 and 5. Although different priors are used in different chapters, we found both priors to be sufficiently vague. We can adjust for trials with more than two arm,s by following Dias, Sutton, Ades & Welton (2013). In the case of an FE model δ_{ij} is simply set equal to $d_{t_{ij}} - d_{t_{i1}}$, instead of being centred on this value.

The hyper-parameters for $\boldsymbol{\mu}$ and $\mathbf{d}$ are chosen in order to have a rather flat distribution on the log odds ratio. Kass & Wasserman (1996) point out that the properties of a prior on one scale can differ when transformed to another scale. A seemingly vague prior such as $N(0, 100^2)$ is not vague on the inverse logit scale, as most of the distribution is close to either 0 or 1. However, the choice of $\sigma = 1.83$ meant that two standard deviations on each side of the mean covered 95% of the transformed (approximately flat) distribution. This is illustrated in Leahy et al. (2018) and in Chapter 4. For the prior for σ_δ, when transformed to the probability scale, two standard deviations covers the range (0.02, 0.98), which we deemed to be sufficiently vague. This prior has also been examined by Lambert et al. (2005). We compare the log odds ratios (LORs) for a dataset of RCTs in Hepatitis C using a commonly used WinBugs prior, $N(0, \frac{1}{0.001})$, versus the prior proposed here. The results are shown in Table A.6 in Section A.3.

2.8.2 Time to Event Outcomes

In this thesis we consider two different models used, either based on hazard ratios or median survival data.

Hazard Ratio Model

A simplified method of the models used in Woods et al. (2010) was employed using just hazard ratios (HRs). The model is:

$$\ln(H_{s,k}) = \mu_i + d_k - d_b + \mathrm{re}_{ik} - \mathrm{re}_{ib}, \tag{2.2}$$

where $H_{s,k}$ is the cumulative hazard for treatment k in study s, μ is the study effect and d_k and d_b are the treatment effect for treatment k and the baseline treatment effect (defined on a per study basis) respectively. For further information see Woods et al. (2010). The random effects on each of the treatment effect could result in the model being over-parameterised, leading to identifiability challenges. In practice, informative prior information for the variance parameters can resolve this. Nonetheless, an alternative model to consider could be to only use one random effects. The effect of this is not explored in this thesis, but could be considered in future work.

In Chapters 3 and 5 we assume proportional hazards when utilising the HR model. For the time to event simulation in Chapter 5 we know that the proportional hazards assumption is met as the data is simulated from an exponential distribution. However, when working with real-world data the proportional hazards assumption should be checked by methods such as visual inspection of the log-cumulative hazard plots, checking if time dependent covariates are significant in the Cox Model or analysing the Schoenfeld Residuals (Hess (1995), Schoenfeld (1982)).

Median Survival Model

In most cases we will only have AgD so our choice of model is limited to a model with one parameter. Therefore, we assume an exponential survival distribution and use the median survival time. In this case we take the fact that 50% of patients have had the event at this time point, and put this value in for the numerator and the total number of patients as the denominator. If we had IPD then other survival models would be available to use, which may produce more accurate results. Let r_{ij}, the number of events (i.e. 50%) in the j^{th} arm of the i^{th} trial. $r_{ij} \sim \text{Bin}(p_{ij}, n_{ij})$, where p_{ij} is the probability of an event in the j^{th} arm of the i^{th} trial and n_{ij} is the number of patients in the j^{th} arm of the i^{th} trial. As we are dealing with medians the outcome is t^*, the time when half of the patients have had the event. Therefore, we obtain μ_i and δ_{ij} from:

$$S_{ij}(t) = \begin{cases} \exp(-t^* \exp(\mu_i)) & \text{if j=1} \\ \exp(-t^* \exp(\mu_i + \delta_{ij})) & \text{if j>1} \end{cases}, \tag{2.3}$$

noting that for the median time p_{ij} is 50%. δ_{ij} is the effect of the treatment in the j^{th} arm of the i^{th} trial, μ_i is the baseline risk in study i and t^* is the median survival time. The treatment effects follow $\delta_{ij} \sim N(d_{t_{ij}} - d_{t_{i1}}, \sigma_\delta^2)$, where $d_{t_{ij}}$ denotes the effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment and σ_δ represents the between trial standard deviation (SD) of the treatment effect. The prior distributions chosen are $\mu_i \sim N(0, 1.83^2)$, $d_k \sim N(0, 1.83^2)$ and $\sigma_\delta \sim \text{Unif}(0, 2)$. The effect of treatment A (the reference treatment) is set to zero in the model with all other treatments being compared to treatment A.

2.9 Meta-Regression

Meta-regression is a tool for investigating the impact that variables other than the treatment of interest have on the outcome. There may be a number of covariates affecting the outcome of interest other than the treatment effect, for example age or severity of disease. When clinically plausible covariates have been identified these can then be tested through a meta-regression (Borenstein et al. (2009*a*)). The effect size in the studies is the dependent variable and the average covariate in each study is the independent variable. These covariates can either be discrete or continuous. It can be difficult for meta-regression to positively identify a covariate that affects outcome due to a lack of data points (i.e. a small number of studies), and therefore the Cochrane handbook only recommends that meta-regression is carried out if there are at least ten studies in the meta-analysis (Green & Higgins (2005)). However, IPD can potentially help to identify influential covariates.

Taking the binary outcome equation (3.4) as an example, this can be extended to include a covariate by:

$$\text{logit}(p_{ij}) = \begin{cases} \mu_i & \text{if j=1} \\ \mu_i + \delta_{ij} + (\beta_{t_{ij}} - \beta)x_{ij} & \text{if j>1} \end{cases}, \tag{2.4}$$

where β is the effect of the covariate, and x_{ij} is the proportion possessing the characteristic associated with the covariate of interest for a binary covariate, or the mean of a continuous covariate in the j^{th} arm of the i^{th} trial. Other outcomes can be extended using similar methods.

2.10 Covariate-Treatment Interactions

Covariates may be classified in a number of different ways, depending on their interaction with the treatments. In some cases, treatments will affect the outcome for all patients equally, regardless of the level of the covariate, i.e. the covariate has no interaction with the treatment. Therefore, the relative treatment effects will stay the same throughout the network. This means that the covariate is a *prognostic variable* (Phillippo et al. (2016)). In other cases there will be an interaction between the covariate and the treatment, which means the covariate is considered an *effect modifier*. This can make it more difficult to separate the underlying treatment effects from the covariate effects, so methods such as MAIC may be useful in adjusting for varying levels of the covariate. Hence, there are three ways to model the interaction between the covariate and the treatment:

1. Identical (Prognostic Variable): The covariate has no interaction with the

treatment. This model is described in Equation refIdCov in Section **??**.

2. Effect Modifiers:

   (a) Independent: The covariate has a separate underlying distribution for each treatment. In this case the model in Equation IdCov is extended to:

$$\text{logit}(p_{ij}) = \begin{cases} \mu_i & \text{if j=1} \\ \mu_i + \delta_{ij} + (\beta_{t_{ij}} - \beta_{t_{i1}})x_{ij} & \text{if j>1} \end{cases}, \tag{2.5}$$

   where $\beta_{t_{ij}}$ is the effect of the covariate interacting with each treatment. In this case each covariate-treatment interaction comes from a seperate distribution such as $N(0, 1.83^2)$.

   (b) Exchangeable: The covariate comes from a common distribution for each treatment. In this case the model is the same as Model 2.5. However, each $\beta_{t_{ij}} \sim N(\mu_\beta, \sigma_\beta^2)$, which in turn comes from $N(0, 1.83^2)$.

Identifying which assumption is most appropriate can have a large influence on results, but this may often be difficult to identify. In this thesis we investigate the extent with which IPD can help with this.

2.11 Consistency

The assumption of consistency is the most important assumption underlying NMA. If a network is consistent then we will obtain the same estimate from both the direct and the indirect evidence, i.e.:

$$\delta_3^{AC} = \delta_1^{AB} + \delta_2^{BC}, \tag{2.6}$$

where δ_1^{AB}, δ_2^{BC} and δ_3^{AC} are the effects from studies 1 (comparing A and B), 2 (comparing B and C), and 3 (comparing A and C), respectively (Dias, Welton, Sutton, Caldwell, Lu & Ades (2013)). Sometimes, however, the consistency assumption will not hold, in particular when there are differences in effect modifiers between trials. Some of the methods in this thesis, such as including IPD or undertaking an MAIC can attempt to account for these differences between studies.

2.12 Bayesian Methods

Bayesian methodology provides a flexible framework (O'Hagan & Forster (2004)) which has been used by numerous authors, especially in the field of medicine

(O'Hagan & Luce (2003), Ashby (2006), Spiegelhalter, Abrams & Myles (2004)). Bayesian methods are useful for a number of reasons, such as:

- Bayesian methods are a flexible modelling framework, which facilitate complex synthesis. In NMA this can be useful as we can model different types of evidence separately, as described in Chapter 3.
- We can include known information in the prior. This is especially important in medicine as we often have more knowledge than simply considering the data which we are analysing. By specifying the prior we can incorporate this information.
- It enables us to discuss the results in terms of direct probabilities and is therefore more easily interpretable. The frequentist notion of confidence intervals can often be quite confusing, whereas Bayesian confidence intervals, generally referred to as credible intervals are directly interpretable.
- Bayesian models are useful in predicting future benefits, in particular when a priori beliefs have been incorporated. (Prevost et al. (2000)).

Spiegelhalter et al. (2004) points out that Bayesian priors can be particularly useful in healthcare, as most trials produce incremental gains adding to the knowledge base as a whole, as opposed to "paradigm-shifting break throughs". Thus the prior can be particularly useful to incorporate the previous knowledge base.

Although Bayes theorem was published in 1763, it was not until the 20^{th} century that Bayesian techniques came to the fore. There were two main reasons for this. Firstly, many statisticians did not approve of the subjective nature of Bayesian statistics. Secondly, it was not practical to implement a Bayesian framework until the development of computers. Bayesian techniques were implemented at multiple times during the 20^{th} century. For example, during the second world war Alan Turing and others used Bayesian statistics to crack the Enigma code (Good (1950), Simpson (2010)). It was also used in medical research to demonstrate that smoking causes lung cancer (Cornfield (1951)), calculating insurance premiums (Bailey (1950), Hickman & Heacox (1999)), and predicting rare events such as a nuclear detonation (Ikle et al. (1958)). However, in a number of these applications Bayesian statistics were often used as a tool of convenience, rather than being considered as an academic concept in their own right. It was not until the introduction of Markov Chain Monte Carlo (MCMC) that Bayesian statistics were fully appreciated and routinely implemented. The results of the NMAs in this thesis are obtained using MCMC methods through OpenBUGS (Spiegelhalter et al. (2014)) or JAGS (Plummer (2012)).

2.13 Bayes Theorem

For completeness it is useful to introduce Bayes Theorem. Bayes theorem gives us the probability of an event conditional on another event having taken place. Let A and B be two events. Given that P(A) > 0 and P(B) > 0, then,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This can also be used to give a posterior distribution. Let $P(A)$ be the prior information. $P(A|B)$ is the posterior distribution that we are trying to find. B can then be thought of as the evidence. This can also be written in terms of random variables. Let $\theta = (\theta_1, ..., \theta_p)$ be the parameter of interest and $\pi(\theta)$ be the prior distribution. We observe some data $X = (X_1, ...X_n)$. Then,

$$f(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{f(X)} \propto f(X|\theta)\pi(\theta),$$

where $f(X|\theta)$ is the likelihood and $f(\theta|X)$ is the posterior probability. f(X) is a normalising constant given by:

$$f(X) = \begin{cases} \int f(X|\theta)\pi(\theta)d\theta & \text{for continuous random variables} \\ \Sigma_\theta f(X|\theta)\pi(\theta) & \text{for discrete random variables.} \end{cases}$$

2.14 Importance of Priors

In the frequentist or classical approach to statistics no prior information is taken into account. In classical statistics the objective is that we are trying to find a fixed unknown value by sampling.

We can consider the following contrived example, which illustrates that even in a classical framework, prior knowledge is still important. If a frequentist wanted to look at the average height of the adult population of Ireland, they would take a sample of the population and find the average, and use that as the best estimate of the sample of the population as a whole. However, if he got a value of three feet or eight feet then he would most likely conclude that there was either sampling error, or that the sample was too small, or sampling from a biased part of the population. This is because he has some prior knowledge of adult height.

In a Bayesian framework we can include this knowledge as a prior and it can influence the results (Spiegelhalter et al. (2004), Berger & Berry (1988)). In fact, Spiegelhalter states that Bayesian methods "explicity allow for the possibility that the conclusions of the analysis may depend on who is conducting it". The amount by which it influences the results is a function of the size of the data collected, and how informative the prior is. We can have non-informative, weakly informative, moderately informative or informative priors. Thorlund et al. (2013), Gelman (2006), and Lambert et al. (2005) have conducted simulation studies on looking at the effect of different types of priors. Ibrahim et al. (2000) discuss prior distributions for regression models. In this thesis we use non-informative priors in our analysis.

2.15 Bayesian Hierarchical Models

When conducting a meta-analysis we may be concerned with obtaining estimates at multiple levels, for example, on a study level and then on a higher overall level. We once again look at the random effects model in the school test scores example from Section 2.7. At the bottom level we obtain an estimate of the mean in each school, $\hat{\theta}_i$ where

$$\hat{\theta}_i \sim f(\theta_i, \xi_i^2). \tag{2.7}$$

Each $\hat{\theta}_i$ comes from a true mean for each school, θ_i. This in turns comes from an overall θ, which is the top level of the hierarchy. Each θ_i comes from a common distribution

$$\theta_i \sim f(\theta, \sigma^2), \tag{2.8}$$

and so are considered exchangeable.

Let n be the number of studies and let

$$Y = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n).$$

and

$$\Psi = (\theta_1, \theta_2, ..., \theta_n).$$

Let $f(y|\Psi)$ be a function for the data and $p(\Psi|\theta, \sigma^2)$ be the prior distribution for Ψ. Then,

$$p(\Psi|y) \propto f(y|\Psi)p(\Psi),$$

and the joint posterior distribution is given by:

$$p(\Psi, \theta, \sigma^2|y) \propto f(y|\Psi)p(\Psi|\theta, \sigma^2)p(\theta)p(\sigma^2),$$

where $p(\theta)$ and $p(\sigma^2)$ are the prior distributions for θ and σ^2 respectively. For further discussion see Whitehead (2002).

The Directed Acyclic Graph (DAG) for Equations 2.7 and 2.8 are shown in Figure 2.2. A DAG is a graphical representation of a set of quantities, G, in which each quantity, $v \in G$ is represented as a node in the graph. Nodes which have an arrow running into them mean that they have a stochastic dependency on the *parent node*, from which the arrow is pointing. Given that nodes have a stochastic dependency, this means that they are independent of all other nodes, except nodes from which they are descendants (Lunn et al. (2012)).

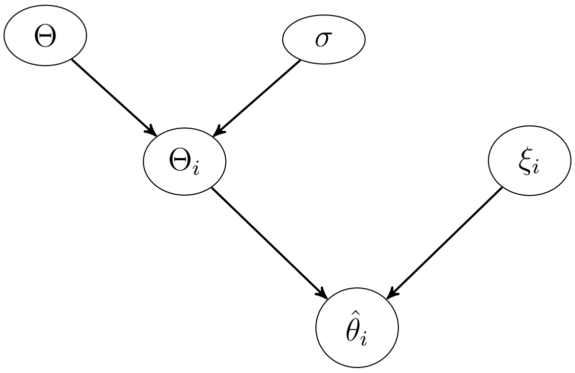


Figure 2.2: Directed Acyclic Graph for Meta-Analysis. The lowest level $\hat{\theta}_i$ gives the ***estimate*** *of the mean effect in each study. The middle level gives the* ***true*** *mean and standard deviation in* ***each*** *study. The top level gives the* ***overall true*** *mean and standard deviation.*

Bayesian hierarchical models can be particularly useful when we have evidence from different design types. In this case we can model each type of evidence seperately and combine these on the highest level of the hierarchy, as suggested in Welton et al. (2012), Schmitz et al. (2013). Additionally bayesian methods can be utilized to augment the data available from studies in humans, by also including animal studies (DuMouchel & Harris (1983)). Finally, these hierarchical model also use to group classes of treatments together, as shown in Dominici et al. (1999) and Owen et al. (2015).

2.16 Deviance Information Criterion (DIC)

The DIC (Spiegelhalter et al. (2002)) is a goodness of fit test which allows us to identify which model best fits a given dataset. In this thesis we employ the DIC in Chapters 4 and 5 to identify the nature of the covariate-treatment interaction and in Chapter 5 to test the appropriateness of FE models versus RE models. The DIC is a generalisation of the Akaike Information Criterion (AIC). The DIC is defined as:

$$DIC = p_D + \bar{D},$$

where $\bar{D}$ is the expectation of the deviance, $\bar{D} = D(\theta) = -2log(f(X|\theta)) + C$, and $f(X|\theta)$ is the likelihood function. As we always use the DIC to compare two or more models, the constant C will cancel out. $p_D = \bar{D} - D(\bar{\theta})$, where $\bar{\theta}$ is the expectation of θ. p_D is essentially a penalty for a larger number of parameters, which is required because more parameters make it easier for the model to fit the data. The model with the lowest DIC is the model which best fits the data. A variety of different thresholds, generally between two and ten, are used to conclude that one model is superior to another model. In this thesis we use a difference of three to indicate that there is a meaningful difference in model fit, as is consistent with (Welton, Caldwell, Adamopoulos & Vedhara (2009)).

2.17 Markov Chain Monte Carlo (MCMC)

For most Bayesian models it is not possible to identify the posterior as a known distribution, i.e. it is not of closed form. This posed a major problem for Bayesian statistics until the development of more powerful computers and the method of Markov Chain Monte Carlo (MCMC). This method allows us to sample from the posterior distribution by means of a *Markov Chain*, i.e., each step is dependent only on the previous step and is independent of all steps before that point. If we know the prior and the likelihood function we then have an expression for the posterior. We can pick an initial starting value and sample from the posterior a number of times until the samples converge. We can then disregard the unconverged samples (known as the burn-in) and get an average over the converged samples to give an estimate of the true value. If we are trying to find an estimate for more than one variable then we first use the initial sample of the other variables for our first sample and the most recent sample of each variable after that.

There are two popular types of MCMC:

- Metropolis-Hastings (MH) - in this case not all samples are accepted. The posterior distribution is used to decide whether to accept or reject each proposal. Suppose the Markov Chain is in state x. The following algorithm is used to sample from p:
  - Propose a move to y with probability $q(y|x)$.
  - Calculate the ratio:
  $$r = \frac{p(y)q(x|y)}{p(x)q(y|x)}. \tag{2.9}$$
  - Accept the proposed move with probability $\min(1, r)$. Otherwise, remain at x.
- Gibbs sampling - this is a special case of MH where the proposal distribution is a full conditional, and therefore the proposal is always accepted,

> as is outlined in Chib & Greenberg (1995). Let $\pi(y|x)$ be the full conditional distribution for y. This is proportional to $p(y)q(x|y)$, with a constant of proportionality being $p(x)$, and therefore it cancels. In Gibbs sampling $q(y|x) = \pi(y|x)$, and therefore the acceptance probability is one.

The MCMC processes carried out in this thesis are carried out in the software packages BUGS and JAGS, which implement Gibbs sampling, when the full conditional distribution is available.

2.18 Convergence Diagnostics

In order to decide the number of iterations to use for the burn-in we need to be able to assess whether or not a chain has converged. The Gelman-Rubin statistic (Gelman & Rubin (1992)) provides a method to assess whether the chain has converged. For this we need to run at least two chains. We then look at the ratio:

$$\frac{\text{between chain variance}}{\text{within chain variance}}. \tag{2.10}$$

If this ratio is close to 1 it has typically converged. Gelman et al. (2011) suggest a value of 1.1 shows adequate mixing between chains and indicates that the chains have converged. However, they warn that this could potentially lead to premature assumptions of convergence and therefore also recommend other methods for assessing convergence, for example using visual inspection. One way in which this can be done is by using “bgr diag” in BUGS as shown in Figure 2.3. The green line shows the between chain variability, the blue line shows the within chain variability and the red line is the ratio between the two. We can see that it has converged if the red line is close to one and the blue and green lines are stable.

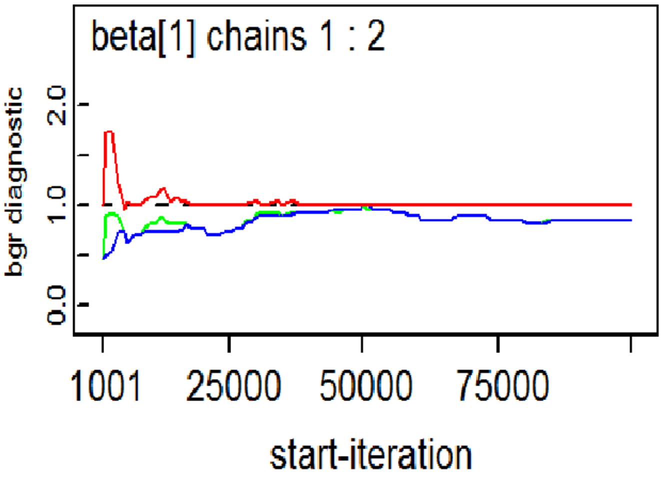


Figure 2.3: BGR Plot from OpenBugs

Chapter 3

Aggregate Level Matching of Single-Arm Evidence

The methodological work in this chapter and the work on the HCV infection dataset is currently under review for publication in “Statistics in Medicine” as a revised resubmission. It is based on collaboration with a number of different authors. Dr Aisling O’Leary and Dr Emma Gray undertook the systematic review and provided guidance on the clinical aspect relating to HCV infection. Dr Arthur White made substantial comments on the work during the internal academic PhD confirmation process, and in subsequent meetings thereafter. Dr Howard Thom, Dr Jeroen Jansen and myself were the co-organisers of an ISPOR workshop stemming from the original research. Some additional improvements to the work arose through discussions surrounding this workshop, namely including the plug-in estimator model (described in Section 3.2.1) and the scenario where the bias was varied (described in Section 3.3.1). All other aspects and all implementation of the additional improvements were carried by the author. This chapter also includes an application in melanoma. Claire Gorry undertook the systematic review for this application and provided guidance on the clinical aspect.

3.1 Introduction

RCTs are considered the gold standard of evidence, as their controlled approach minimises potential bias. However, sometimes not all treatments have been evaluated in an RCT, and the only available evidence on a treatment may be from sources such as observational studies or single-arm studies. These types of evidence may contain valuable information, especially when no other evidence is available, but they may potentially be biased (Sterne et al. (2016), Cameron et al. (2015), Valentine & Thompson (2013)). In a well conducted RCT we can be confident that the patients are exchangeable across treatment arms, as they have been randomly assigned. However, the same cannot be said for non-randomised evidence.

In one of the most recent reviews of it's kind, Griffiths et al. (2017) examined submissions to three Health Technology Assessment (HTA) agencies between 2010 and 2015; NICE (UK), CADTH (Canada), and IQWiG (Germany). The percentage of HTA submissions which considered non comparative evidence was 38%, 13%, and 12% for each agency respectively, although this may be a wider cohort than simply single-arm trials. Submissions based exclusively on non comparative evidence was 4%, 6%, and 4% respectively, making a total of 27 submissions, although some of these submissions may have included IPD. Positive outcome rates for non comparative evidence alone versus overall submissions were 60% versus 84% for NICE, 69% versus 68% for CADTH, and 17% versus 38% for IQWiG. From this analysis it is clear that when RCTs are unavailable, HTA decision-makers are willing to consider non-comparative evidence, despite its limitations. Given that single-arm evidence is being accepted by HTA agencies, it is crucial to minimise potential bias by providing clear guidelines for incorporating this evidence into an NMA. In 2016 Bell et al. (2016) identified "priority research requirements" such as exploration into "when data is drawn from intervention in one setting versus comparator patients from a different setting" and "the extent to which observational designs can complement or replace those of RCTs".

Matching-Adjusted Indirect Comparisons (MAIC) or Simulated Treatment Comparisons (STC) are emerging methods for reducing bias when incorporating single-arm trials into an NMA. However these require IPD for at least some of the studies. In practice IPD can be quite difficult to obtain and it is frequently the case that only aggregate data is available. Additionally, there are still questions on how applicable MAIC and STC are to a full evidence network, as opposed to just three treatments. The benefits and drawbacks of MAIC are investigated in Chapter 5. However, it is important to investigate techniques for including single-arm evidence with only aggregate data also.

Much research has been carried out on including non-randomised evidence into an NMA to date. Sutton & Abrams (2001) investigate incorporating observational evidence in a pairwise meta-analysis, while Schmitz et al. (2013) and Efthimiou et al. (2017) propose methods for including observational studies in an NMA in a manner that treats both evidence types seperately. Thom et al. (2015) and Goring et al. (2016) investigate methods of including single-arm evidence. Thom et al propose using a random effects model for the expected placebo effect in order to incorporate single-arm evidence into an NMA. However, this interferes with the randomisation of the RCT evidence. Goring et al estimate absolute treatment effects and compare the results of their models to the recommended relative effects models. Other work on absolute versus relative effects includes Hong et al. (2016*a*), and follow-up discussions by Dias & Ades (2016) and Hong et al. (2016*b*). However, it has been argued by Dias & Ades (2016) that absolute effects "effectively breaks randomisation, and in fact runs against the entire way in which randomised controlled trials are designed, analysed, and used."

Here we aim to address the research requirement proposed by Bell et al by assessing the appropriateness of including single-arm evidence in an NMA through matching to other arms or trials with similar patient covariates. As we are dealing with single-arm studies we match on both effect modifiers and prognostic variables, as recommended in Phillippo et al. (2016). Firstly, we consider a method of choosing another arm from the network. This means that we treat the single-arm and the chosen matched arm as if they are arms from the same study. We consider a naive pooled method and a more formal hierarchical model. The first approach was recently adopted by Jaff et al. (2017) in an application to assess the efficacy of endovascular interventions. They compare the NMAs which do and do not include single-arm trials using matching. Schmitz et al. (2018) also use a pooled model to include single-arm trials, and investigate the appropriateness of these models depending on how close the matches are to the single-arm trial. Secondly,we match to a chosen trial by pluging in the baseline odds of an event in a chosen matched trial for the reference treatment. We compare these different models and evaluate the performance in a simulation study. We also recommend a number of sensitivity analyses which can be used to detect bias. This is illustrated in Section 3.4 by an example in HCV infection and Section 3.5 by an example in melanoma.

There are a number of advantages to these methods over the methods described above:

1. The method only requires aggregate data.

2. Although including matched evidence adds non-randomised evidence into the network, the randomisation of the available RCTs is kept intact.

3. The methods fit within the relative effects framework.

4. We can apply this method to a network of any size.

The objectives of this chapter are to:

1. Assess which parameters influence the accuracy of the model's estimate in an NMA;

2. Assess under what circumstances it is appropriate to include single-arm evidence.

The remainder of this chapter is organised as follows: A detailed description of the matched arm methodology is provided in Section 3.2. Section 3.3 describes the construction and results of a comprehensive simulation study. Section 3.4 presents our methods applied to a HCV infection network and Section 3.5 presents our methods applied to a melanoma network. A general discussion and some recommendations are provided in Section 3.6. Table 2 at the beginning of this thesis gives a glossary of notation used throughout this chapter.

3.2 Methods

3.2.1 Model Development

In this Chapter we extend Equation 2.8.1 as we examine at a number of ways for incorporating single-arm trials into an NMA.

1. (a) Including the matched arm in data - pooled model: We treat the single-arm trial and its chosen match as if they come from the same trial. In this case we group all evidence types together in such a way that different forms of evidence are not distinguished by the model. This is the most straight-forward model to implement and is set up as in Equation 3.4. In this case the number of studies is the total number of RCT and matched studies.

   (b) Including matched arm in data - hierarchical model: Again, we treat the single-arm trial and its chosen match as if they come from the same trial. In this model we estimate the treatment effect, d, at each level of the study design, and then combine to give the overall treatment effect. This model provides more flexibility, at the cost of requiring a more stable network structure. This model is shown in Figure 3.2.

2. Plug in estimator model: We assume the log odds of having an event on treatment 1 (reference treatment) is the same for the single-arm trial and the chosen matched trial. This model has the advantage of not using any data more than once. Both the RCTs and the single-arm trials are pooled as in 1a above.

When using matching to incorporate single-arm trials into an NMA, the main concern is that the chosen comparator will include patients from a very different population to the single-arm trial. This could add bias to the model as one treatment could end up looking superior when it was simply allocated to a particularly healthy patient population. In order to minimise this potential bias we propose to choose matched comparators with the closest patient characteristics. When including a matched arm in the data (Model 1a or 1b), this match can be any other arm in the network, from either an RCT or another single-arm trial, provided that that the treatment is not the same as the treatment in the single-arm to which we are matching. Let M be the number of covariates considered, let x_{m_k} be the proportion of patients possessing the characteristic associated with the covariate in the single-arm trial with treatment k, and let $x_{m_{ij}}$ be the proportion of patients possessing the characteristic associated with the covariate in arm j of study i. The difference is: $\Delta_{ij,k} = \sum_{m=1}^{M} |x_{m_{ij}} - x_{m_k}|$. When matching to a trial using a plug-in estimator (Model 2) we can choose to match to any RCT. In this case the difference simplifies to: $\Delta_{i,k} = \sum_{m=1}^{M} |x_{m_i} - x_{m_k}|$. In our example, since we look at binary covariates, x is a proportion. However, this can easily extend to continuous covariates where x is the mean value of the covariate. The full steps and sensitivity analyses that can be carried out when matching single-arms is presented in a flow diagram in Figure 3.1.

For the hierarchical model the extra level on the treatment effect is modelled as follows:

$$\begin{aligned} d_{\text{RCT}[k]} &\sim N(d_{[k]}, \sigma^2_{\text{des}}), \\ d_{\text{MATCHED}[k]} &\sim N(d_{[k]}, \sigma^2_{\text{des}}), \end{aligned} \tag{3.1}$$

where σ_{des} is the between-study design SD which represents the variability between the RCT and matched studies. This is essentially a random effects model for the study design level. The prior distributions for d are the same as for the standard and pooled NMA models, and $\sigma_{\text{des}} \sim \text{Unif}(0, 2)$. A benefit of the hierarchical model is that we can adjust for overprecision in the matched arms by applying a multiplicative factor ω to the matched precision, thus inflating the variance. This represents our increased uncertainty in the evidence from the matched arms, and can be thought of as the weight given to the matched evidence. For example, if ω is small this indicates that we believe that the matched evidence is a poor estimate of the mean effect.

$$\begin{aligned} d_{\text{RCT}[k]} &\sim N(d_{[k]}, \sigma^2_{\text{des}}), \\ d_{\text{MATCHED}[k]} &\sim N(d_{[k]}, \tfrac{\sigma^2_{\text{des}}}{\omega}). \end{aligned} \tag{3.2}$$

Schmitz et al. (2013) investigate a number of different models for incorporating different types of evidence into an NMA and propose the hierarchical model as the

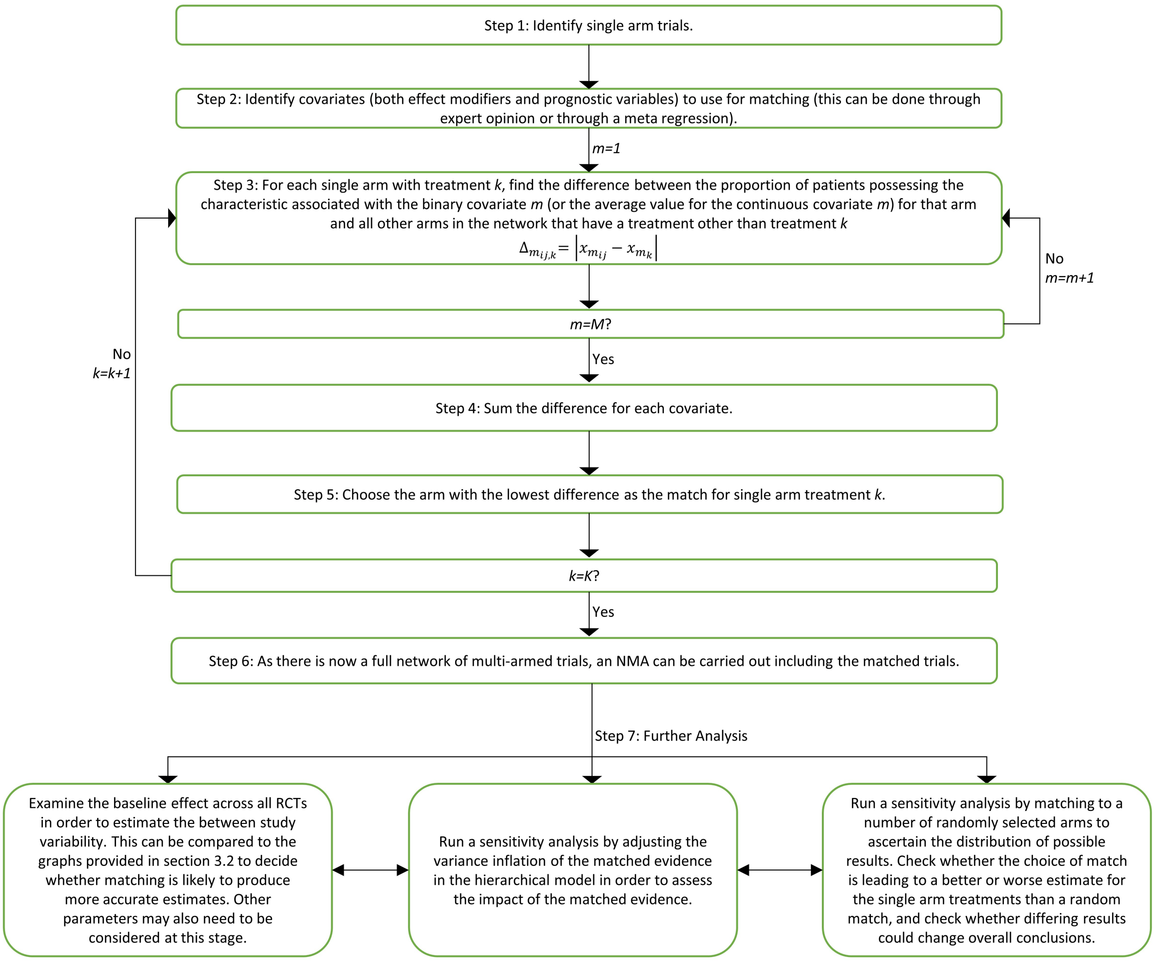

Step 1: Identify single arm trials.
Step 2: Identify covariates (both effect modifiers and prognostic variables) to use for matching (this can be done through expert opinion or through a meta regression).
m=1
Step 3: For each single arm with treatment k, find the difference between the proportion of patients possessing the characteristic associated with the binary covariate m (or the average value for the continuous covariate m) for that arm and all other arms in the network that have a treatment other than treatment k
$\Delta_{m_{ij,k}} = \left|x_{m_{ij}} - x_{m_k}\right|$
No
m=m+1
m=M?
Yes
No
k=k+1
Step 4: Sum the difference for each covariate.
Step 5: Choose the arm with the lowest difference as the match for single arm treatment k.
k=K?
Yes
Step 6: As there is now a full network of multi-armed trials, an NMA can be carried out including the matched trials.
Step 7: Further Analysis
Examine the baseline effect across all RCTs in order to estimate the between study variability. This can be compared to the graphs provided in section 3.2 to decide whether matching is likely to produce more accurate estimates. Other parameters may also need to be considered at this stage.
Run a sensitivity analysis by adjusting the variance inflation of the matched evidence in the hierarchical model in order to assess the impact of the matched evidence.
Run a sensitivity analysis by matching to a number of randomly selected arms to ascertain the distribution of possible results. Check whether the choice of match is leading to a better or worse estimate for the single arm treatments than a random match, and check whether differing results could change overall conclusions.

Figure 3.1: Steps to be taken when carrying out the matching by including an extra arm in the data, and recommended sensitivity analyses. $x_{m_{ij}}$ is the proportion possessing the characteristic associated with the binary covariate m (or mean of in the case of a continuous covariate m) in the j^{th} arm of the i^{th} trial, x_{m_k} is the proportion possessing the characteristic associated with the binary covariate m (or mean of the covariate in the case of a continuous covariate m) in the comparison arm, K is the total number of treatments examined in single-arm trials, and M is the total number of identified covariates.

best option, as we can obtain estimates for each study design levels, and down-weight certain types of evidence. However, due to the extra level in a hierarchical model, estimates may often be less certain and can be drawn closer to zero.

The plug-in estimator model is written for the RCT part as:

$$\text{logit}(p_{ij}) = \nu_i + \delta_{ij}, \tag{3.3}$$

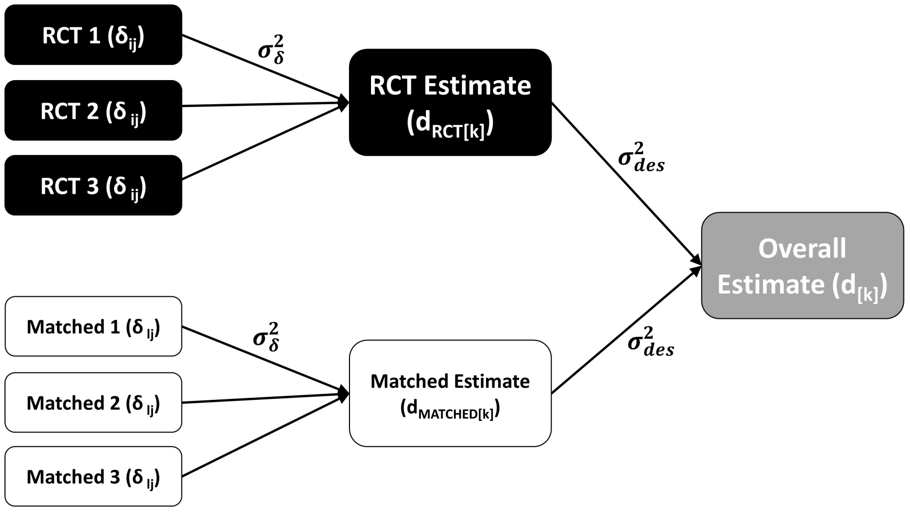

RCT 1 (δ_{ij})
RCT 2 (δ_{ij})
RCT 3 (δ_{ij})
σ^2_δ
RCT Estimate ($d_{RCT[k]}$)
σ^2_{des}
Overall Estimate ($d_{[k]}$)
Matched 1 (δ_{lj})
Matched 2 (δ_{lj})
Matched 3 (δ_{lj})
σ^2_δ
Matched Estimate ($d_{MATCHED[k]}$)
σ^2_{des}

Figure 3.2: Three level hierarchical model. The first level represents estimates of the treatment effect from each study (δ_{ij} for the RCTs (i=1 to number of RCTs) and δ_{lj} for the matched evidence (l=1 to number of matched studies)). The middle level has separate estimates of the treatment effect for the RCTs ($d_{RCT[k]}$) and the matched trials ($d_{MATCHED[k]}$). The final level combines these estimates to get an overall estimate of the treatment effect ($d_{[k]}$).

and the matched part is written as:

$$\text{logit}(p_l) = \nu_{\text{ChosenRCT[i]}} + \delta_l, \tag{3.4}$$

where ν_i is the log odds of having an event for treatment 1 in trial i, and ChosenRCT[i] represents the chosen trial to match to single-arm trial l. For the matched part of the model the j subscript is unnecessary on δ as this part includes single-arm evidence only.

3.3 Simulation Study

Our research was motivated by a network of treatment regimens for the treatment of HCV infection. Therefore, we loosely based the simulation study on the HCV infection network which we discuss in Section 3.4. In recent years there has been a major change in the way that newer HCV infection treatments have been formulated (Goring et al. (2016)). Several of these newer treatment regimens were not assessed for efficacy using comparative randomised trials and hence have not been compared in RCTs to older treatments. The network is therefore disconnected and in fact includes many treatment regimens that have only been assessed for efficacy in single-arm trials, but nonetheless have shown very promising results in such studies. Therefore, to assess comparative efficacy with existing treatment regimens, it is desirable to be able to make a comparison between any two treatment regimens using a connected network, hence the need to use the matching methods described in this chapter.

3.3.1 Methods

Set Up

We compare the results of the matching method to only using RCT evidence. We also include a scenario in which we randomly choose a matched arm, in order to assess how much value there is in finding the best match. We present results and analysis on the effect that varying four parameters have on including matched evidence. These are:

1. The SD of the baseline study effect, σ_μ. This is the measure of variability that will be used throughout this chapter. It is the standard deviation of the baseline risk of having an event in each study in the network. It can be thought of, for example, as the variability between published studies. In our simulations it is varied between 0 and 1.18. The highest value of 1.18 is approximately 3 times as large as the estimate of σ_μ in the HCV infection network. It should be noted that we choose to investigate this measure, instead of the more commonly investigated between-study heterogeneity, which quantifies how relative treatment effects varying between trials. When matching single arm trials we are interested in matching on prognostic variables and effect modifiers, whereas the between-study heterogeneity in RCTs is only affected by effect modifiers. Since σ_μ is driven by both unidentified prognostic variables and effect modifiers, we consequently use this as our measure of between study variability. Note that in the simulation the baseline risk refers to the risk when on treatment one. However, when including the matched arm as data (models 1a and 1b), this baseline risk is calculated for whichever treatment is first in each study. As we are only interested in relative effects this will not affect our results. In the case of the plug-in estimator model, a common baseline study effect is required and therefore the baseline risk of having an event on treatment 1 is calculated.

2. The bias, ξ, in the single-arm trials. In this case the mean study effect in the RCTs comes from a fixed distribution $N(0, 0.59)$, while the mean of the single-arm studies varies between 0 and 1. In the most extreme case the single-arm trials are simulated from $N(1, 0.59)$. This corresponds to a

baseline rate of 75% on the probability scale, so we believe that this is a sufficiently high value to use as the mean to cover plausible scenarios. We refer to this difference in the mean as bias in the single-arm trials.

3. The upper bound, z, of the uniform prior, $(0, z)$ on the between-study design effect in the hierarchical model, σ_{des}. z varied from 0.25 to 2. 2 is a sufficiently vague upper bound as two standard deviations cover the range (0.02, 0.98) on the probability scale.

4. The variance inflation on the matched evidence in the hierarchical model, ω. This has the effect of downweighting the matched evidence. It is varied between 0.1 to 1.

The parameters which we varied are highlighted on the directed acyclic graphs (DAGs) in Figure 3.3.

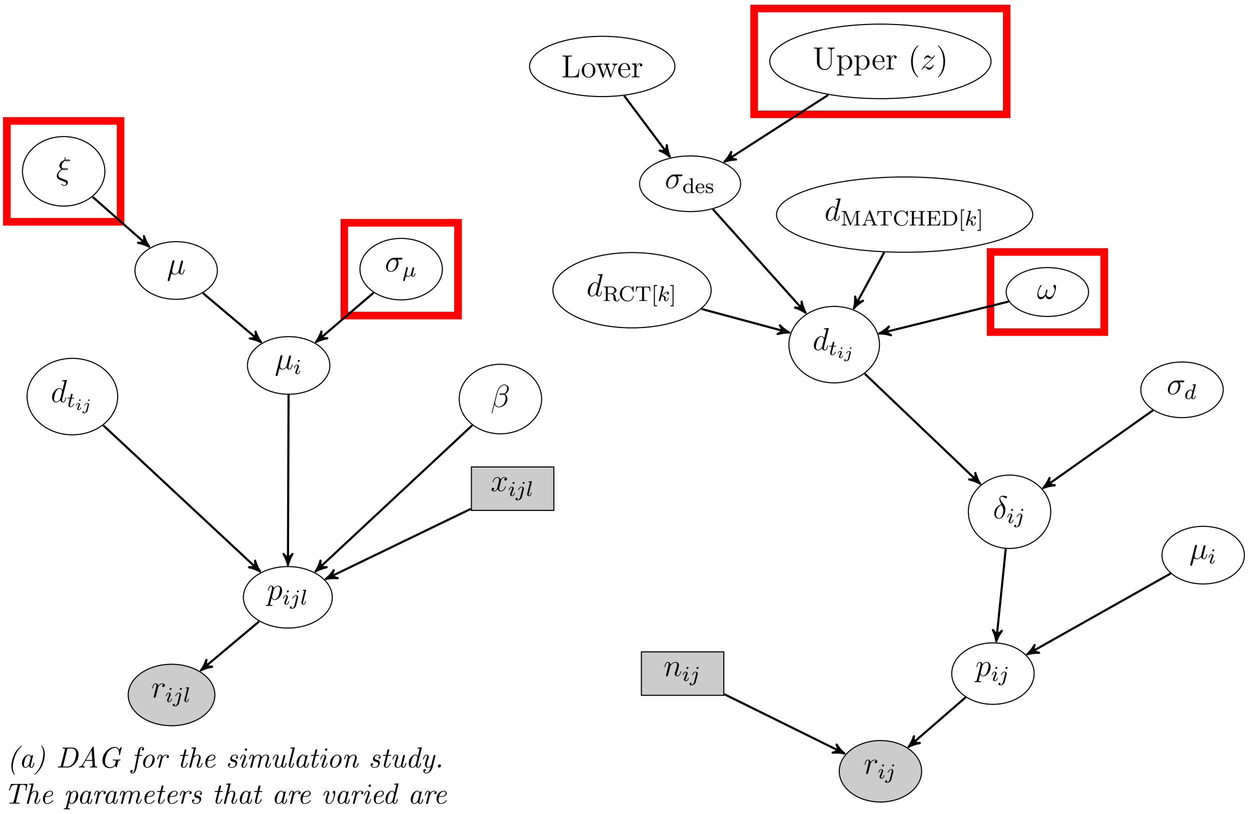
When examining the hierarchical model two further scenarios were considered:

- When examining the prior on the between-study design effect, σ_{des}, we also included a scenario with 12 RCT studies, but we assume that there are eight of one type and four of another. This is to quantify the effect of the prior on the hierarchical model versus the effects of genuine difference in study types.

- When examining the variance inflation of the matched evidence in the hierarchical model, ω, we also include ω acting on the RCTs instead of the randomly matched trials, $d_{\text{RCT}[k]} \sim N(d_{[k]}, \frac{\sigma^2_{\text{des}}}{\omega})$. Although we do not recommend that this is done in practice, this is examined in order to demonstrate that the effect of downweighting RCTs and downweighting single-arm trials will have different outcomes, which is due to the quality of the evidence types.

A network of 12 studies with seven treatments was simulated. The studies consisted of eight two-armed RCTs and four single-arm studies. The RCT studies consisted of treatments numbered 1-5 and the single-arm studies consisted of treatments 1, 2, 6 and 7. To ensure our results were applicable to a wide range of real world networks we kept our network as generic as possible, and randomly assigned treatments to the RCTs at each iteration. Additionally, the treatment effects, study effects and covariate effects were simulated at each iteration. An example of a typical network is shown in Figure 3.4.

Given we want to compare all possible treatment regimens in an NMA, it is necessary to form a connected network (Jansen et al. (2011)). In 3.4c the matched connection alone would not form a connected network, as treatments 1 and 5 are



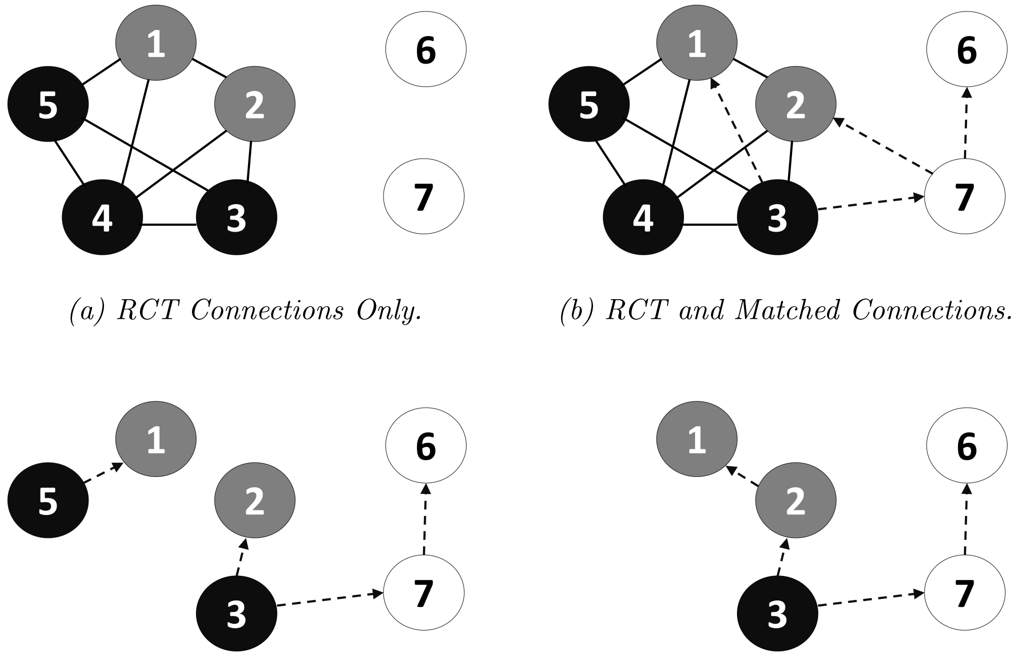
(a) DAG for the simulation study. The parameters that are varied are dependent on the data themselves, and the investigator has no input. Note: ξ is 0 for RCTs.

(b) DAG for Hierarchical Model. The parameters varied are modeling choices which are chosen by the investigator. Note: This DAG simplifies to the RCT or Pooled Model by excluding all hyper-parameters on $d_{t_{ij}}$.

Figure 3.3: Directed acyclic graphs for simulations and models. The red box around certain nodes indicates that these are varied in a simulation study. Square nodes represent fixed quantities while the ellipses are stochastic nodes. The shaded nodes represent observed quantities.

not connected to the other treatments. A disconnected network would rely on the remote evidence in the hierarchical structure and therefore we believe that connected networks are preferable. The impact of this is explored in Section 3.3.2. For the purpose of ensuring a connected network for each study design, when faced with this scenario after choosing the best matches, we would restrict the potential matches for treatment 1 to those are already included in the larger network, i.e. treatments 2, 3, 6, and 7 in this case, as shown 3.4d. In Section 3.3.2 we discuss how often our simulated networks were connected within each study type when using the best match.

Simulations were carried out by finding the probability of the event p_{ijl} for each individual patient l, in arm j of study i. This was computed by the study effect μ_i, the treatment effect $d_{t_{ij}}$, the effect of one baseline binary covariate β, and x_{ijl},



(a) RCT Connections Only.

(b) RCT and Matched Connections.

(c) Network is disconnected at the matched level. Therefore we restricted our choice of match for treatment 1 to one of the treatments already included in the larger network.

(d) Network is connected at the matched level.

Figure 3.4: Example Simulated Network: Black nodes denote treatments in RCT trials only, white nodes denote treatments in single-arm trials only, and gray nodes denote treatments in both single-arm trials and RCT trials. Black solid lines represent RCT connections and dashed lines represent matched connections. The treatment requiring the match has the arrow pointing towards it, i.e., an arm with treatment 3 has been chosen for the single-arm trial containing treatment 7.

a binary indicator variable for the presence of the characteristic associated with the covariate: $\text{logit}(p_{ijl}) = \mu_i + d_{t_{ij}} + \beta x_{ij}$. The event rate r_{ijl} was calculated from Bernoulli(p_{ijl}) for each patient l. These were then aggregated to give an event rate for each arm, r_{ij}, which was provided to our model. As we are examining aggregate data we summarize the binary covariates as the proportion of patients possessing the characteristic associated with that covariate in a given arm. In the case of continuous covariates our summary would be the mean, which takes the place of the proportion in the model. Thus our approach can be extended in a natural fashion to continuous covariates. We also included scenarios where there were three covariates. For brevity the description and results of the three covariate scenario has been included in the appendix.

Default Values

The default values were chosen based on the HCV infection network in Section 3.4 or based on vague prior distributions. The values were set as follows:

- The between-study variability, $\mu_i \sim N(0, 0.59^2)$. The default of 0.59 is half way through the range on which we are varying the between-study variability.

- The relative difference of treatments to baseline, $d_k \sim N(0, 1.83^2)$, which is a broad range for this parameter.

- The number of patients for single-arm trials was simulated from $n_i \sim \text{Unif}(75, 134)$, which reflected the inter-quartile range of the size of trials in the HCV infection network. The number of patients in the RCTs was twice this value.

- The probability of patients possessing the characteristic associated with each covariate was sampled from $\text{Unif}(0, 1)$ for each trial in order to cover the full range of possibilities. From this, each individual patient possessing the covariate was sampled from a Bernoulli distribution with the trial probability of possessing the characteristic associated with the covariate. This ensured that the RCT trials reflected the real world situation where treatment arms are exchangeable.

- The covariate effect size, $\beta = -1.04$, was set to be twice as large as the largest estimated covariate effect we found when analysising the RCT studies through a meta regression (Borenstein et al. (2009*b*)) in our HCV infection example, and was therefore thought to be adequately large to reflect possible real world covariates.

Implementation

We ran the models as described in Section 3.2 to assess how well they predicted the true relative treatment effects, d. Models were run using Markov chain Monte Carlo (MCMC) simulation in OpenBugs (Spiegelhalter et al. (2014)). A burn-in of 20 000 iterations was tested for convergence using the Gelman-Rubin statistic (Gelman & Rubin (1992)). Following this another 10 000 iterations were sampled for our estimates. If the convergence condition was not met the number of iterations was doubled (both for the burn-in and for the samples for estimation) and then tested again until the Gelman-Rubin statistic was less than 1.1. If the chains had not converged after a burn-in of 320 000 this iteration was excluded from the analysis. Given the amount of simulations required we decided it was not feasible to include chains that had not converged by this point. If the chains did not converge for one of the models in a particular simulation the results for all other methods in the simulation were excluded from the analysis in order to eliminate any potential bias due to differing simulations. Reasons for non-convergence could include identifiablity of certain parameters or numerical issues. In this chapter a potential cause could be the sparsity of information at each part of the hierarchical model, and the fact that we only have two study designs to help us estimate σ_{des}. We varied each parameter in turn and sampled 8-12 data points for each at least 200 times.

Let T be the total number of simulations. In order to assess whether including single-arm trials produces more accurate estimates than using RCT evidence alone, we look at the mean absolute error (MAE) for treatments 2 to 5: MAE $= \sum_{s=1}^{T} \left\{ \sum_{k=2}^{5} |d_{k_s} - \hat{d}_{k_s}|/4 \right\} /T$, where $\hat{d}_k$ is our model's estimate of the relative effect of treatment k, compared to treatment 1. A lower MAE means that the model produces a more accurate estimate of the treatment effect.

We can also compare the posterior SD as reported in the BUGS output to quantify the uncertainty in the resulting estimates. While we might expect a lower posterior SD when more evidence is added into the model, this might not necessarily hold due to the potential bias from the single-arm studies. Graphs showing the MAE and BUGS posterior SD are included in the results section.

3.3.2 Results

In a preliminary analysis we analysed the MAE for the hierarchical model with networks connected at all levels using the best match versus networks that were disconnected at the matched level using the best match, and found that using the connected network produces better results. We restricted our network matches to ensure a connected network on each study design if the best match did not produce a connected network on each study design. This replicates how a connected network could be ensured in a real world scenario. In a simulation study only 34% of the randomly matched networks and 24% of matches using the covariate were connected for each study types (i.e. RCT and matched single-arm sub-networks); as a consequence we restricted the choice of match for the hierarchical model for any network that was not connected using the best match. By contrast, 99.8% of networks were connected when pooling across study types. For simplicity, we use a pooled model for all analyses that do not directly investigate the nature of the hierarchical model.

The graphs in this section show the lines of best fit for the simulated data points, obtained by regression using both a linear and a quadratic term. Graphs showing the original data and the Monte Carlo Error (MC Error) of the simulations are included in the appendix.

Between-Study Variability

We examine the effect of the between-study variability, σ_μ, on the accuracy of the estimate of the treatment effect and the posterior SD in Figure 3.5. The study effect is centered at zero. As σ_μ increases, the accuracy of the estimates obtained by including the single-arm evidence decreases to a point where including single-arm evidence produces less accurate estimates than the RCT only model. The estimates are more accurate when the study effects are close together, as the information is more accurate in the model. However, as the difference increases noise is added into the model by assuming that treatment arms in the matched studies are exchangeable, when in reality they come from different distributions. The plug-in estimator model becomes worse than the pooled model when σ_μ is large.

The MAE will, of course, be larger for the treatments that are not in any RCTs. The error for treatments 6 and 7 alone when matching by covariate in the pooled model is between 0.57-1.52 as σ_μ varies from 0 to 1.18.

The crossing point is of particular interest as this is the decision point of whether to include the matched evidence or not. In Table 3.1 we estimate the between study variability for the HCV infection network, as detailed in Section 3.4, and a number of publicly available datasets with binary outcomes from the statistical software package, *R* (R Core Team (2013)). The estimates of σ_μ are of the order of, if not greater than the crossing point in the graph, indicating that matched evidence may often lead to an increase in bias. It may be worth noting that the dataset with the most objective outcome of mortality, i.e. the thrombolytic dataset has one of the lowest between study variability. In a previous analysis Turner et al. (2012) found that heterogeneity is lowest in networks where the outcome is an objective outcome. It may be possible that between study variability also follows this same pattern.

Figure A.3 examines the danger of incorrectly assuming that the single-arm studies come from the same distribution as the RCTs. We see that the estimates and posterior SDs increase slightly as the bias increases. However, the increase is quite small. We see that the bias has a low influence on the plug-in estimator model in particular.

Table 3.1: Between study variability in example networks

Dataset	$\hat{\sigma_\mu}$	Source	Outcome Description
Thrombolytic	0.36	gemtc in *R*	Mortality after 30-35 days.
Hepatitis C	0.39	details in Section 3.4	SVR after 12 or 24 weeks.
Certolizumab	0.35	gemtc in *R*	Improvement of at least 50% on the American College of Rheumatology scale (ACR50) at 6 Months.
Smoke	0.59	pcnetmeta in *R*	Successful cessation of smoking at 6-12 months.
Depression	0.61	gemtc in *R*	Reduction of at least 50% from the baseline score on the HAM-D or MADRS at week 8 (or, if not available, another time between week 6 and 12).

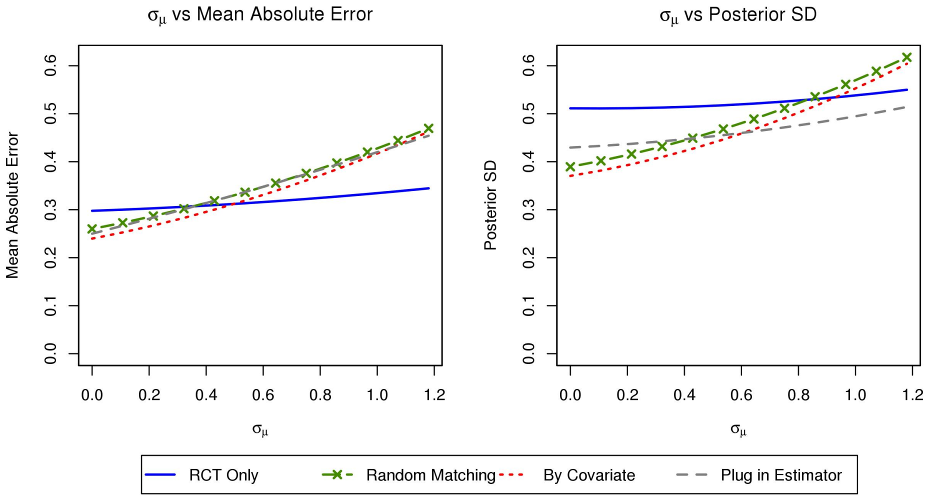


Figure 3.5: Effect of between-study variability, σ_μ, on the MAE and posterior standard deviation. Covariate effect, β=-1.04. The extreme left point on the graph shows the scenario where the study effect is set to zero for every study. The variability between the studies increases with the horizontal axis.

Hierarchical Model

We now look at the hierarchical model, which includes a between-study design effect. We exclude the plug-in estimator model from these results as this model is not be affected by the parameters that are varied in this section. Figure 3.7 shows how varying the upper bound of the prior on the between-study design effect affects the MAE and the posterior SD. The prior on the between-study design effect is given by Unif$(0, z)$ where z varies along the x-axis. The RCT only model has the same estimate at each point as there is only one study type in this model, therefore we have used the average value over all simulations for this line.

As the upper bound of the prior on the between-study design variance increases,

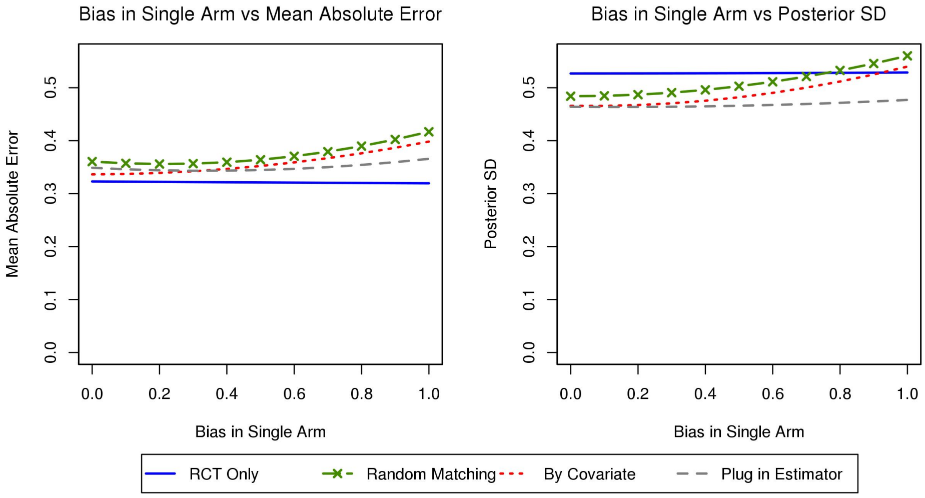


Figure 3.6: Effect of varying the bias in the single-arm trials on the MAE and posterior standard deviation. Between-study variability, σ_μ=0.59, covariate effect, β=-1.04. The study effect in the RCTs comes from N(0, 0.59). The value on the x-axis indicates the mean of the single-arm trials. Hence, at the extreme left point both the single-arms and RCTs come from the same distribution.

the MAE and posterior SD also increases. There is a fourth line on the graph corresponding to 12 RCT studies, where we assume that there are eight of one type and four of another. Here we see the same trend as before, with the posterior SD increasing as the prior on the between-study design σ_{des} increases. However, this time it happens to a lesser extent. The increase in posterior SD when including the extra RCTs is solely due to the prior. Any extra increase for the matching methods is due to actual differences between the study types.

Figure 3.8 shows the effect that down-weighting matched evidence, ω, has on the accuracy of our model's estimate of the treatment effect and the posterior SD. Again, the RCT only line is the average value over all simulations, as there is no weight on the matched evidence. The "RCT By Omega" line shows ω acting on the RCT evidence instead of the (randomly) matched evidence. Decreasing the weight of the high quality RCT evidence generally gives less accurate estimates and larger posterior SDs. However, decreasing the weight of the matched evidence actually gives a concave downwards shape. The MAE and the posterior SD is smallest when $\omega = 0.1$, i.e., the smallest weighting in the simulation study. However, weighting the matched evidence fully appears to be preferable to, or at least as good as, some of the values for ω in the centre of the graph.

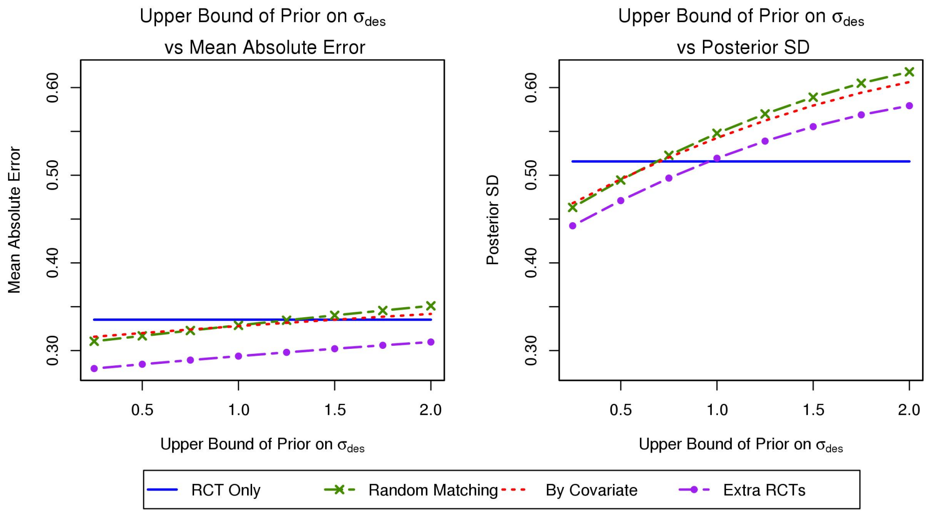


Figure 3.7: Effect of the prior on the between-study design effect (σ_{des}) on the MAE and posterior standard deviation. Between-study variability, σ_{μ}=0.59, covariate effect, β=-1.04. The horizontal axis shows the upper bound for the prior on the between-study design effect Unif(0, z). Taking two standard deviations for the largest value of an upper bound of 2 corresponds to (0.02, 0.98) on the probability scale.

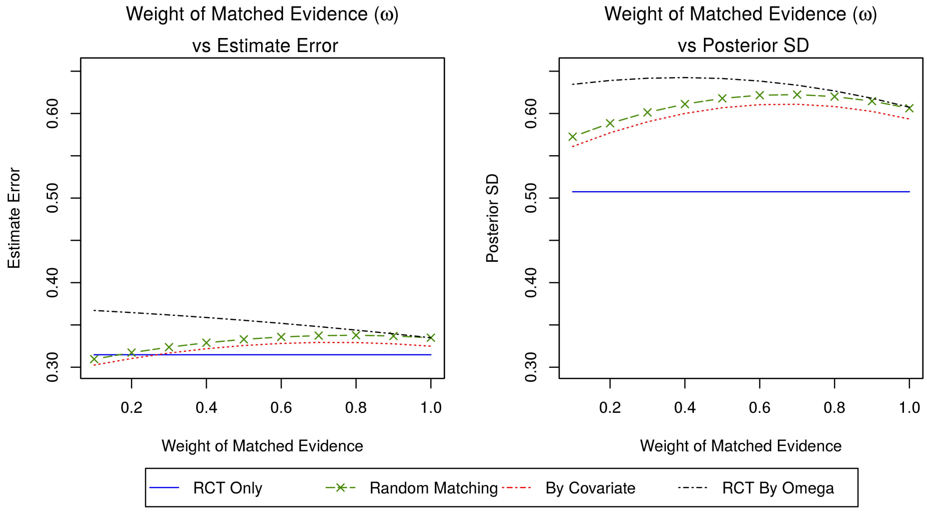


Figure 3.8: Effect of ω on the MAE and posterior standard deviation. Between study variability, σ_{μ}=0.59, covariate effect, β=-1.04, prior on between study design effect, $\sigma_{des} \sim$ unif(0, 2).

Other parameters considered

In an exploratory analysis a number of other parameters were examined through the simulation study. We analysed the magnitude of the covariate and found that for larger covariates effects there was an increased advantage of choosing a matched arm based on covariates over a randomly chosen match. However, the magnitude of the covariate effect chosen for the simulation study was based on how large we would reasonably expect a single standardised covariate to possibly be. Heterogeneity between treatment effects was also examined. When there was large heterogeneity all methods were less accurate and less precise at estimating the treatment effect. However the loss of accuracy and precision was more pronounced for the RCT only evidence than when matched evidence was included. Finally, we examined how trial size affected the accuracy of treatment effect estimation. We found that including matched evidence was most beneficial when trials were small.

3.4 Example: Hepatitis C Virus (HCV) Infection - Treatment Naive Patients - NMA

Chronic HCV infection is a global health burden of major concern. A number of treatment combinations are currently licensed for genotype 1 (GT1) HCV infection. At the time of this systematic review there was a wealth of clinical trial evidence available that compared single regimens in terms of treatment duration and with or without the addition of ribavirin. However, head-to-head comparative trials between all licensed regimens for GT1 infection were unavailable. There have been major advances in the treatment of HCV infection in the recent years, with a move away from non-specific anti-viral therapies, which had relatively low levels of cure rates, to antiviral combination therapies that directly target replication of the virus, with the ability to significantly enhance cure rates. While RCTs are the most appropriate method to directly assess the relative efficacy of all regimens from a methodological perspective, RCTs comparing newer treatment regimens to older (and most probably inferior) treatments may not be appropriate from an ethical perspective. Therefore, most of the evidence available on the newer HCV infection treatment regimens is disconnected from the older network, and in fact comes in the form of single-arm evidence. We apply the techniques discussed in Section 2 to indirectly estimate the relative treatment effect of licensed regimens for the treatment of GT1, by including single-arm trials, in treatment naive patients with chronic GT1 HCV infection.

3.4.1 Methods

A systematic review was conducted in accordance with the criteria of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses group (PRISMA) (Moher et al. (2009)) on 6th May 2015, and repeated on 24th November 2015 and on 14th May 2016. These cut off dates are based on the needs of informing the clinical programme in Ireland. We identified 13 RCTs that looked at two inter-

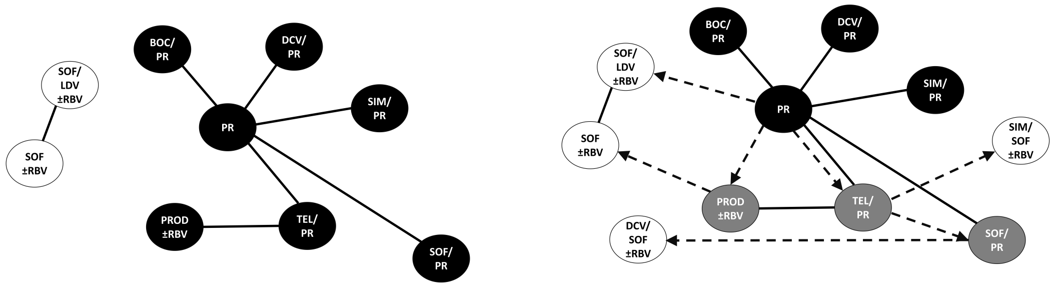
Table 3.2: List of treatment regimens with abbreviations

Abbreviation	**Treatment Regimen**
PR	Pegylated-interferon and ribavirin
DCV/PR	Daclatasvir (+ pegylated interferon and ribavirin)
BOC/PR	Boceprevir (+ pegylated interferon and ribavirin)
SIM/PR	Simeprevir (+ pegylated interferon and ribavirin)
TEL/PR	Telaprevir (+ pegylated interferon and ribavirin)
SOF/PR	Sofosbuvir (+ pegylated interferon and ribavirin)
PrOD±RBV	Paritaprevir boosted with ritonavir, ombitasvir and dasabuvir (with or without ribavirin)
SOF/LDV±RBV	Sofosbuvir and ledipasvir (with or without ribavirin)
DCV/SOF±RBV	Daclatasvir + Sofosbuvir (with or without ribavirin)
SIM/SOF±RBV	Simeprevir and sofosbuvir (with or without ribavirin)
SOF/RBV	Sofosbuvir (+ ribavirin)

ventions (Jacobson et al. (2011), Hézode et al. (2009), McHutchison et al. (2009), Poordad et al. (2011), Kwo et al. (2010), Kumada et al. (2012), Jacobson et al. (2014), Manns et al. (2014), Fried et al. (2013), Pol et al. (2012), Lawitz, Lalezari, Hassanein, Kowdley, Poordad, Sheikh, Afdhal, Bernstein, DeJesus, Freilich et al. (2013), Gane et al. (2014), Dore et al. (2016)) and 18 single-arm trials with no comparator. There were 11 different regimens in our network. In total the single-arm studies examined seven different treatment regimens. We choose one arm with no comparator from each treatment regimen to use in this example (Sherman et al. (2011), Feld et al. (2014), Osinusi et al. (2013), Afdhal et al. (2014), Kwo et al. (2015), Sulkowski et al. (2014), Lawitz, Mangia, Wyles, Rodriguez-Torres, Hassanein, Gordon, Schultz, Davis, Kayali, Reddy et al. (2013)). Where possible we choose single-arms that had full information on the covariates of interest. The full list of regimens is described in Table 4.1. The outcome of interest was a binary outcome, Sustained Virological Response (SVR).

We first ran an analysis which investigated the 13 RCT studies. We then matched the single-arms to an arm of another study to act as the comparator regimen. We then re-ran the meta-analysis using these matched arms. We used the pooled model, four hierarchical models, and the plug-in estimator model. The first hierarchical model gave equal weight to both study types, the other three hierarchical model provided sensitivity analyses by down-weighting the matched estimate by multiplying the matched precision by $\omega = 0.7$, $\omega = 0.4$ and $\omega = 0.1$.

The studies were matched according to the proportion of patients that were cirrhotic, had genotype 1a, and had viral load >800,000 IU/ml at baseline. These covariates were chosen, because according to clinical expert opinion, they were likely to influence SVR, and because these were reported in the majority of trials. Each single-arm study was compared to every individual arm (including other single-arm studies) and the differences in their baseline characteristics were determined. As we had more than one covariate to use for matching we added the difference in the three covariates together. The arm with the smallest difference in baseline characteristics was then chosen as the matched arm. Not all arms had information on each of the three covariates. The best match was chosen from the pool of arms that had at least the same amount of information as the single-arm trial. A network diagram of the RCT only network and a network including the single-comparator studies is shown in Figure 3.9. Note that for the plug-in estimator model the closest trial was chosen instead of the closest arm, and therefore a network diagram comparing matched treatments is not applicable for this model.



(a) RCT only

(b) Including matched evidence when matching to arms

Figure 3.9: Network Diagram for HCV infection. Black nodes denote treatments in RCT trials only, white nodes denote treatments in single-arm trials only, and gray nodes denote treatments in both single-arm trials and RCT trials. Black solid lines represent RCT connections and dashed lines represent matched connections. The treatment requiring the match has the arrow pointing towards it. Although SOF/LDV±RBV and SOF/RBV are in an RCT they are treated as only being in a single-arm trial as the RCT cannot be included as it is not connected to the rest of the network.

Random Matching Arms - Sensitivity Analysis

It is possible that the results of the analysis could change based on the choice of matched arms. As there are millions of combinations of arms, it is not possible to test all of them. Therefore, we ran over 1 000 simulations to check matching arms at random in order to assess the sensitivity to the choice of match.

Removing RCT arms

In order to assess the accuracy of matching on the HCV infection network we investigated how well our method would work on RCTs, which we turned into single-arm trials by removing arms. We found an estimate of the efficacy of each drug by using only the RCT evidence. We excluded those that did not have full information for the three covariates and were therefore left with nine RCTs and four treatments: PR, TEL/PR, BOC/PR and SIM/PR. We then selected studies at random and removed all the arms except one. We then matched these new "one-arm" studies to the remaining studies in the network. We compared our results to the results of the reduced dataset with the "one-armed" trials excluded to see which was closer to the estimate with nine trials. We compared these results making a varying amount of studies into single-arm trials, from one to six. In all cases we ensured that there was at least one study with each treatment remaining in an RCT so that we were comparing like with like.

3.4.2 Results

Table 3.3 shows the mean and posterior SD of the log odds ratios (LOR) for each treatment regimen versus PR (standard of care). For treatment regimens which have both RCT and matched evidence (TEL/PR, SOF/PR, and PrOD±RBV) including the matched evidence decreases the posterior SD of the LOR in all cases. The hierarchical model generally results in higher posterior SDs than the pooled model. This may be explained by the additional prior variance on the study effect. However, for all treatment regimens with RCT evidence only (DCV/PR, BOC/PR and SIM/PR) the adjusted hierarchical model with matched evidence down-weighted has the smallest posterior SD. The pooled model generally gives a more extreme LOR than the hierarchical model, since in the latter case the summary effect is shrunken toward zero.

We store the rank of each treatment at each iteration of the MCMC chain, and use these values to *a posteriori* estimate the probability of each regimen being in each position, and plot these on a rankogram as shown in Figure 3.10. We can then sum these probabilities to find the probability of each regimen being in the n^{th} position or better. Calculating the SUCRA (SUrface under the Cumulative RAnking curve) (Salanti et al. (2011)) gives us a one number summary for each regimen. Possible SUCRA scores range from 0 to 1. A treatment with a value of 1 means that it is the best treatment with no uncertainty, and a value of 0 mean that it is worst treatment with no uncertainty.

Table 3.4 shows the SUCRA for each regimen. All models that include the extra treatments rank these treatments in the same order. Including the matched evidence has had little impact on the RCT only treatment rankings. The hierarchical model and the hierarchical model with matched evidence down-weighted

Table 3.3: Log Odds Ratio versus PR. White represents largest posterior SD, darker shades of green represent smallest posterior SD within each row.

	RCT		**Pooled**		**Hierarchical**		**Matched Hier Down-weighted** $\omega = 0.7$		**Matched Hier Down-weighted** $\omega = 0.4$		**Matched Hier Down-weighted** $\omega = 0.1$		**Mu From Matched**	
Regimen	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**
DCV/PR	2.15	0.90	2.10	0.89	2.18	0.91	2.21	0.93	2.23	0.90	2.21	0.87	2.12	0.86
BOC/PR	1.10	0.23	1.10	0.26	1.11	0.23	1.11	0.22	1.11	0.23	1.11	0.23	1.11	0.23
SIM/PR	1.17	0.22	1.15	0.24	1.16	0.22	1.16	0.21	1.17	0.22	1.16	0.21	1.16	0.22
TEL/PR	1.06	0.21	1.12	0.20	0.93	0.57	0.97	0.55	0.94	0.59	1.03	0.43	1.12	0.18
SOF/PR	1.64	0.65	1.83	0.40	1.50	0.74	1.56	0.73	1.50	0.76	1.63	0.67	2.42	0.33
PrOD±RBV	3.24	0.61	3.56	0.44	3.12	0.79	3.15	0.78	3.09	0.79	3.25	0.69	3.82	0.44
SOF/LDV±RBV			3.94	0.60	3.08	1.16	3.25	1.16	3.17	1.19	3.30	1.12	4.15	0.64
DCV/SOF±RBV			1.90	0.95	1.52	1.32	1.62	1.36	1.59	1.39	1.78	1.42	2.98	1.05
SIM/SOF±RBV			2.76	0.66	2.12	1.22	2.20	1.24	2.16	1.22	2.36	1.23	2.96	0.61
SOF/RBV			0.86	0.69	0.62	1.16	0.71	1.21	0.59	1.17	0.81	1.23	1.10	0.54

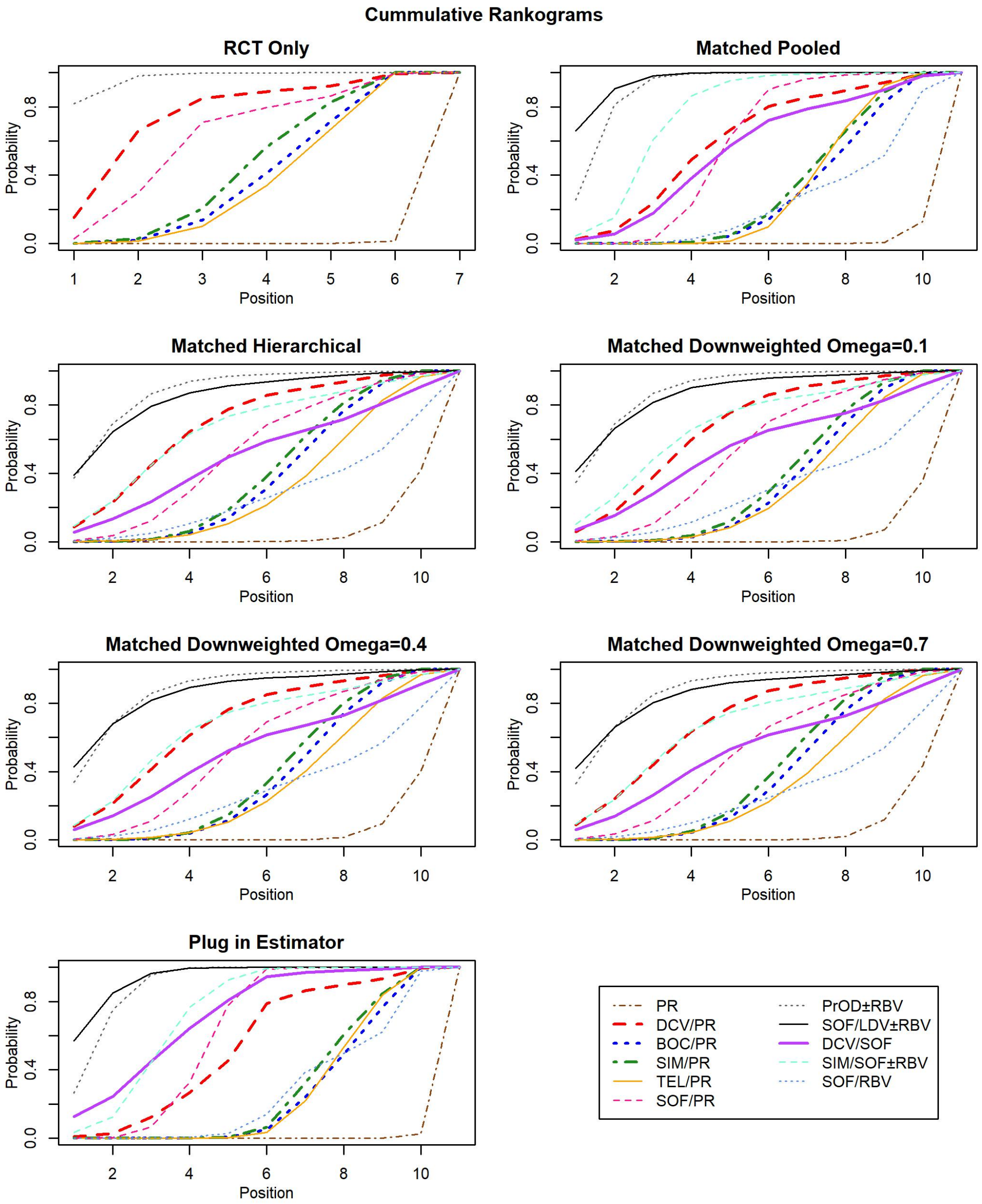


Figure 3.10: Rankogram for each of the seven models.

have the same treatment rankings as the RCT network. In the pooled model only TEL/PR and BOC/PR are switched. These were very close together anyway and are also unlikely to be the best treatments. In the plug-in estimator model SOF/PR and DCV/PR are switched, as well as TEL/PR and BOC/PR. We note that the plug-in estimator model is the only model that ranks DCV/ SOF ±RBV above DCV/PR, which would be considered clinically more plausible.

Table 3.4: SUrface under the Cumulative RAnking curve (SUCRA) score for each of the seven models. A treatment with a value of 1 means that it is best treatment with no uncertainty, and a value of 0 mean that it is worst treatment with no uncertainty.

RCT Only	PrOD ±RBV 0.97	DCV/ PR 0.74	SOF/ PR 0.61	SIM/ PR 0.44	BOC/ PR 0.38	TEL/ PR 0.35	PR 0.00				
Pooled	SOF/ LDV ±RBV 0.95	PrOD ±RBV 0.90	SIM/ SOF ±RBV 0.76	DCV/ PR 0.60	SOF/ PR 0.57	DCV/ SOF ±RBV 0.54	SIM/ PR 0.32	TEL/ PR 0.31	BOC/ PR 0.29	SOF/ RBV 0.24	PR 0.01
Hier	PrOD ±RBV 0.88	SOF/ LDV ±RBV 0.85	DCV/ PR 0.68	SIM/ SOF ±RBV 0.65	SOF/ PR 0.52	DCV/ SOF ±RBV 0.50	SIM/ PR 0.40	BOC/ PR 0.37	TEL/ PR 0.32	SOF/ RBV 0.27	PR 0.06
Hier Matched Down-weighted $\omega = 0.7$	PrOD ±RBV 0.87	SOF/ LDV ±RBV 0.86	DCV/ PR 0.67	SIM/ SOF ±RBV 0.66	SOF/ PR 0.52	DCV/ SOF ±RBV 0.51	SIM/ PR 0.39	BOC/ PR 0.36	TEL/ PR 0.32	SOF/ RBV 0.29	PR 0.05
Hier Matched Down-weighted $\omega = 0.4$	PrOD ±RBV 0.87	SOF/ LDV ±RBV 0.85	DCV/ PR 0.69	SIM/ SOF ±RBV 0.66	DCV/ SOF ±RBV 0.51	SOF/ PR 0.51	SIM/ PR 0.40	BOC/ PR 0.37	TEL/ PR 0.32	SOF/ RBV 0.26	PR 0.06
Hier Matched Down-weighted $\omega = 0.1$	PrOD ±RBV 0.88	SOF/ LDV ±RBV 0.86	SIM/ SOF ±RBV 0.68	DCV/ PR 0.66	DCV/ SOF ±RBV 0.53	SOF/ PR 0.52	SIM/ PR 0.37	BOC/ PR 0.34	TEL/ PR 0.31	SOF/ RBV 0.29	PR 0.04
Plug in estimator	SOF/ LDV ±RBV 0.94	PrOD ±RBV 0.90	SIM/ SOF ±RBV 0.73	DCV/ SOF ±RBV 0.72	SOF/ PR 0.62	DCV/ PR 0.54	SIM/ PR 0.28	SOF/ RBV 0.27	TEL/ PR 0.26	BOC/ PR 0.26	PR 0.00

Random Matching Arms - Sensitivity Analysis

Figure 3.11 shows the distribution of the point estimates for the LOR for each of treatment regimen vs PR. This can identify how sensitive treatments are to the choice of match. We see that the distribution of the point estimates for treatment regimens for which RCTs are available are quite small. However, the distribution of the point estimates for treatments regimens for which only matched evidence is available are much larger. We can see that on average newer treatments have the highest LOR (PrOD±RBV, which has both evidence types and SOF/LDV±RBV, DCV/SOF±RBV, and SIM/SOF±RBV which all have single-arm evidence only). This highlights the importance of including these treatments in an NMA as they are likely to be better than the older treatment regimens, which are included in RCTs. It is important to note here that most estimates from the chosen match give a smaller LOR than average. In particular we compare the best match hierarchical model (denoted by a red triangle) with the distribution shown, as these are the same models. This reassuringly highlights that, in this case, we are at least being conservative with our estimates of the single-arm treatments.

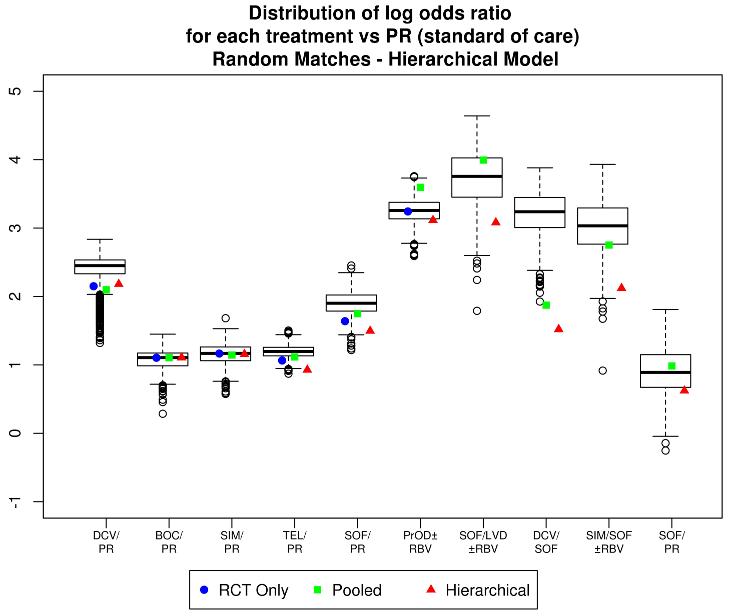


Figure 3.11: Distribution of the log odds ratio using randomly matching arms compared with results from RCT only and the best match chosen by the equal weights method.

Removing RCT arms

As we can see in Figure 3.12, in the HCV infection network if we change three or fewer RCTs to single-arm trials the reduced model is slightly closer to the full network than using our matched methods, on average. However, if we change five or six RCTs into single-arm trials then all methods gives results that are much closer to the full network than the results we obtain from the reduced dataset. We also see that the estimate of the reduced dataset is less precise than the pooled model, and thus the reduced model has the same or higher posterior SD than the plug-in estimator model. A “U-shaped” curved is noticeable for the hierarchical model for the posterior SD. The reason for the high uncertainty when we only match one or two studies is because we have very little evidence on the matched side of the hierarchical model. From this analysis the pooled model would be the most preferable.

3.5 Example: Melanoma - NMA

We have analysed a number of different networks of treatments for melanoma from a clinical perspective. The network on overall survival for second line treatments

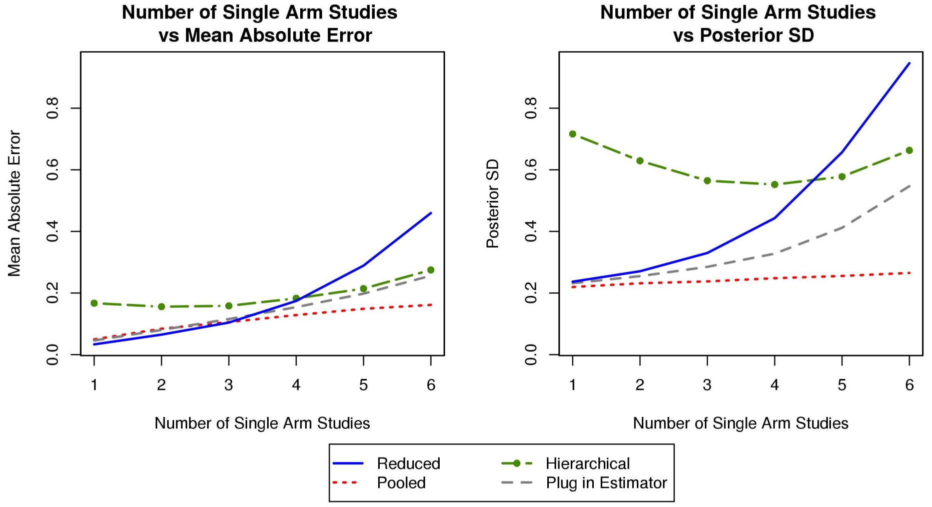


Figure 3.12: We use only one arm from a number of RCTs (as indicated on the x-axis). Results from the reduced RCT network or network obtained by matching the new "single-arm" studies are compared to our best estimate of the treatment effect, i.e. the full nine study network, where the number of new "single-arm" studies equals zero.

and the network for progression-free survival for patients with a BRAF mutation both contained single-arm trials. We therefore present the results of both of these networks, along with the sensitivity analysis of using random matches, as described previously. In both of these networks any single-arm treatments have also been compared in other multi-arm trials and therefore the single-arm trials are not required to connect the network. However, this provides us with an opportunity to assess any potential bias that single-arm trials may introduce.

Melanoma is a type of skin cancer. Metastatic melanoma means that melanoma cells have spread to distant sites in the body. There has been an influx of new drugs licensed for the treatment of metastatic melanoma in recent years, and there is significant uncertainty surrounding how to optimise the use of these new agents, and their relative efficacy. Some melanomas harbour the BRAF mutation, which is present in approximately 17-24% of cases of metastatic melanoma in Ireland. Treatment options vary according to the BRAF mutation status of the melanoma. Patients with BRAF positive melanoma may potentially receive any of the following as first line treatment; dabrafenib/trametinib in combination, vemurafenib/cobimetinib in combination, nivolumab, pembrolizumab or nivolumab/ipilimumab in combination. No RCTs exist to compare these treatment options in terms of clinical efficacy or safety. All of these treatments are licensed for use regardless of line of treatment; however some of the agents have never been trialled in a second line treatment cohort.

Much work has been done in recent years to identify the important prognostic factors in metastatic melanoma, to better inform clinical research and aid clinical decision making. In 2009, Balch et al published a revised staging system for melanoma, based on data from the American Joint Committee on Cancer (AJCC) Melanoma Staging database (Balch et al. (2009)). They found that there were two significant prognostic factors in Stage IV melanoma, elevated serum lactate dehydrogenase (LDH) levels and the sites of distant metastases (nonvisceral versus lung versus all other visceral sites).

Additional work has been published by Korn et al, in the form of a meta-analysis of Phase II trials in metastatic melanoma, to determine progression-free survival (PFS) and overall survival (OS) benchmarks (Korn et al. (2008)). They found that the Eastern Cooperative Oncology Group (ECOG) performance status, presence of visceral disease, and gender were all prognostic for OS. They found ECOG performance status, age and gender were prognostic for PFS although adjusting for these factors failed to eliminate between trial variability in PFS at 6 months. It should be noted that LDH values were not available for almost all of the 42 trials included, and so was not considered in the analysis.

Many treatments have been licensed for the treatment of advanced melanoma, in particular in the period since 2010. While most of these treatments have been licensed based on good quality Phase III RCTs, there is little evidence of their comparative efficacy, mainly because of the pace of change in the field over the last number of years. This systematic review was conducted in order to identify all relevant sources of efficacy information, in order to synthesise the outcomes in a network meta-analysis, to produce relative efficacy outcomes for the included treatments.

3.5.1 Methods

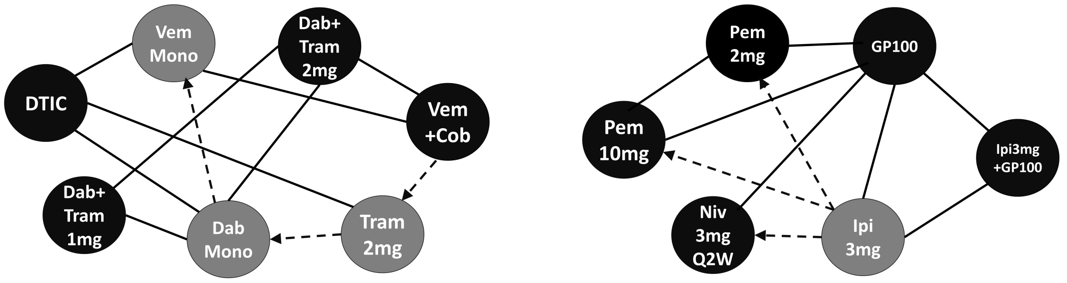
Systematic Review

The systematic review was conducted in accordance with the criteria of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses group (PRISMA) (Moher et al. (2009)). EMBASE, Medline and Central databases were searched for both networks. The databases were searched from inception to 9^{th} March 2017. Seven RCTs (Ascierto et al. (2016), McArthur et al. (2014), Long et al. (2015), Robert et al. (2015), Hauschild et al. (2013), Flaherty, Robert, Hersey, Nathan, Garbe, Milhem, Demidov, Hassel, Rutkowski, Mohr et al. (2012)) and three single arm trials (sosman2012survival, ascierto2013phase, kim2013phase) were identified for the BRAF inhibitor network. Four RCTs (Larkin et al. (2017), Hamid et al.

Table 3.5: List of treatment regimens with abbreviations

Abbreviation	Treatment Regimen
BRAF Inhibitor PFS Network	
DTIC	Dacarbazine 850mg/m^2 or 1g/m^2
Vem Mono	Vemurafnib 960mg twice daily (BD) monotherapy
Tram2mg	Trametinib 2mg once daily (OD) monotherapy
Dab Mono	Dabrafenib 150mg BD monotherapy
Dab+Tram1mg	Dabrafenib 150mg BD + trametinib 1mg OD dualtherapy
Dab+Tram2mg	Dabrafenib 150mg BD + trametinib 2mg OD dualtherapy
Vem+Cob	Vem+Cob = Vemurafenib 960mg BD + Cobimetinib 60mg OD dual therapy
Second Line OS Network	
GP100	GP 100 /Standard of Care
Pem2mg	Pembrolizumab 2mg/kg every 3 weeks (Q3W)
Pem10mg	Pembrolizumab 10mg/kg Q3W
Niv3mg	Nivolumab 3mg/kg every two weeks (Q2W)
Ipi3mg	Ipilimumab 3mg/kg Q3W
Ipi3mg+GP100	Ipilimumab 3mg/kg +GP 100

(2016), Hodi et al. (2010), Robert et al. (2014)) and three single arm trials (Wolchok et al. (2010), Hersh et al. (2011), Zimmer et al. (2015)) were identified for the second line network. Figure 3.13 shows the RCTs and matched trials for both networks. The full list of treatment regimens is described in Table 3.5.



(a) BRAF Inhibitors Progression-free survival Network Diagram

(b) Second Line Overall Survival Network Diagram

Figure 3.13: Network Diagram. Black nodes denote treatments in RCT trials only, gray nodes denote treatments in both single-arm trials and RCT trials. Black solid lines represent RCT connections and dashed lines represent matched connections. The treatment requiring the match has the arrow pointing towards it.

Statistical Methods

In this section we use the HR model (Equation 2.8.2) for our analysis. Single-arm trials were once again matched according to patient characteristics. Covariates identified to be used were LDH greater than the upper limit of normal (LDH

>ULN), ECOG ≥ 1, and disease at more than three sites. Disease at more than three sites was unavailable for the single-arm trial so this was excluded.

There was an extra complication using the survival data as there was insufficient information to obtain HRs on the matched trials based on the summary statistics. However, Kaplan Meier (KM) curves were published in the trials so IPD was reconstructed from these curves using a digitising method provided by Guyot et al. (2012). A second complication arose for the BRAF inhibitor network as Flaherty, Infante, Daud, Gonzalez, Kefford, Sosman, Hamid, Schuchter, Cebon, Ibrahim et al. (2012) contained a three armed trial. A HR was therefore required for Dab+Tram1mg vs Dab+Tram2mg and this was not provided. The KM curves were also reconstructed to obtain this HR using the same digitising process.

If using the HR model for the sensitivity analysis of matching random arms all curves would have to be digitised, which is quite labour intensive. However, another method would be to match arms using a median model, as described in Equation 2.8.2, instead of a HR model. Information on the median is available for each trial in both networks. Although HRs incorporate more information than simply using medians, we discuss in Chapter 5 that sometimes a median model can be preferable to a HR model.

For the second line network a relevant observational study was also identified through the systematic review. Although it may be desirable to run a sensitivity analysis with and without this study, for the purposes of this thesis we have chosen to include it in all scenarios. Therefore, in the second line network we consider scenarios of excluding single-arms versus including single-arms through aggregate level matching.

3.5.2 Results

Log Hazard Ratio

For both networks we present the results of four models. The first model excludes the single-arm trials. The next three models included the matched evidence by using a pooled model, a hierarchical model and an adjusted hierarchical model where $\omega = 0.1$. The log HRs are shown in Table 3.6 for the BRAF PFS network and Table 3.7 for the second line OS network. The posterior SDs are colour coded, with green indicating the smallest SD across the row and red indicating the largest SD. For both networks we see that including the matched evidence increases the precision of our estimates. The pooled model gives the smallest posterior SD, but the hierarchical gives a smaller posterior SD than excluding the single-arm evidence. For the BRAF network the ranking of the treatments stays the same no matter which model we're using. However, for the second line OS network, while Pem 10mg is the best treatment for each model there is a large variation between the ranking of the other treatments.

Table 3.6: Log hazard ratio for each treatment versus DTIC for the BRAF Inhibitor Network - Progression-free survival. Green cells indicate the smallest SD within each row, while red cells indicate the largest SD within each row.

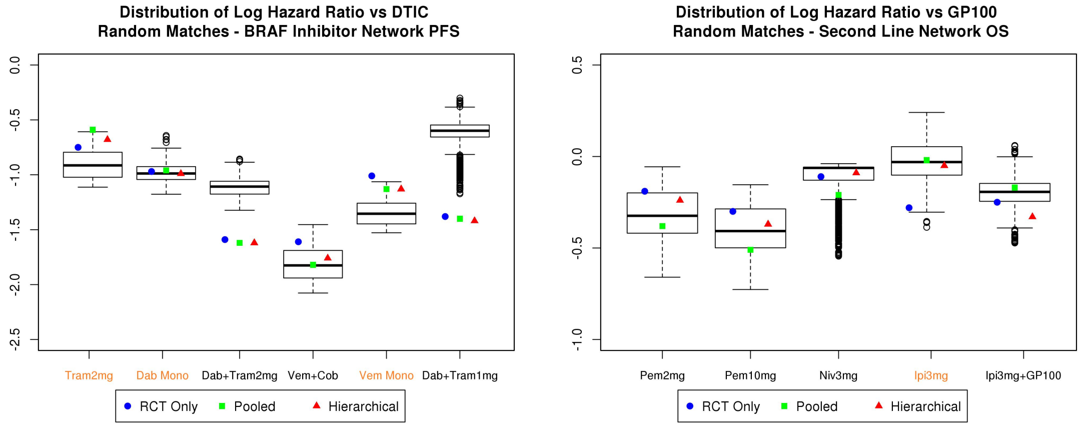
	RCT Only		**Matched Arms Pooled**		**Matched Arms Hierarchical Model**		**Matched Arms Down-Weighted**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Tram2mg	-0.75	1.11	-0.59	0.27	-0.68	0.45	-0.68	0.56
Dab Mono	-0.97	0.94	-0.96	0.26	-0.99	0.90	-0.99	0.83
Dab+Tram2mg	-1.59	0.98	-1.62	0.31	-1.61	0.48	-1.62	0.67
Vem+Cob	-1.61	1.56	-1.82	0.33	-1.76	0.97	-1.70	0.92
Vem Mono	-1.01	0.88	-1.13	0.26	-1.12	0.94	-1.06	0.81
Dab+Tram1mg	-1.38	1.30	-1.40	0.44	-1.41	0.66	-1.42	0.84

Table 3.7: Log hazard ratio for each treatment versus GP100 for the Second Line Network - Overall Survival. Green cells indicate the smallest SD within each row, while red cells indicate the largest SD within each row.

	Excluding Single Arms		**Matched Arms Pooled**		**Matched Arms Hierarchical Model**		**Matched Arms Down-Weighted**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Pem2mg	-0.19	1.27	-0.38	0.51	-0.24	0.50	-0.30	0.61
Pem10mg	-0.30	1.22	-0.51	0.51	-0.37	1.06	-0.42	0.96
Niv3mg	-0.11	1.66	-0.21	0.49	-0.09	1.08	-0.12	1.01
IPI3mg	-0.28	1.36	0.02	0.42	-0.05	1.01	-0.25	0.92
IPI3mg+GP100	-0.25	1.45	-0.17	0.63	-0.33	0.57	-0.30	0.73

Matching Random Arms

When examining Figure 3.14 we note that lower values indicate better treatments, which is the opposite of the corresponding graph for the HCV infection example. We see that two of the treatments included in single-arms (Tram 2mg and Vem Mono) would appear better if choosing a random match, while two of the other treatments would look worse (Dab+Tram 2mg and Dab+Tram 1mg). The other treatment analysed in a single-arm trial (Dab mono) is quite similar to the mean



(a) BRAF Inhibitor Network - Progression-free survival

(b) Second Line Network - Overall Survival

Figure 3.14: Distribution of the log hazard ratio using randomly matching arms compared with results from single-arms excluded, pooled, and hierarchical models, with the best match chosen by the equal weights. The names of the treatment regimens with single-arm trials are coloured in orange.

of the random matches, which is quite a small distribution to begin with. The one non single-arm treatment that looks better (Vem+Cob) is the only treatment which is not in a single-arm that is compared to a treatment in single-arm.

In the second line OS network we see a different scenario from what we have previously seen in this sensitivity analysis of single-arm trials. In this case same treatment (Ipi 3mg) has been analysed in three single-arm trials, which makes it more likely that the estimate from the best matches will correspond to the middle of the distribution of random matches. While this could give a more accurate reflection of the true relative effect of Ipi 3mg, it could also simply be adding noise to matched arms. This is plausible as we know that single-arm matching can produce quite large posterior SDs. Therefore an alternative analysis could be to identify which single arm trial has the closest match and to use this trial only.

In these scenarios we see that including the matched evidence has decreased the posterior SD in both the matched and the pooled models. We have also demonstrated how our method can be transferred to different outcome models. Finally we see once again in the BRAF inhibitor network that single-arms may consist of disproportionately healthy patient populations, while the second line network has no clear bias in either direction, possibly because of the fact that there are a number of single-arm trials which have been matched. Once again we recommend that single-arms can add strength to networks with limited evidence, but should always be analysed with caution.

3.6 Discussion

3.6.1 Summary

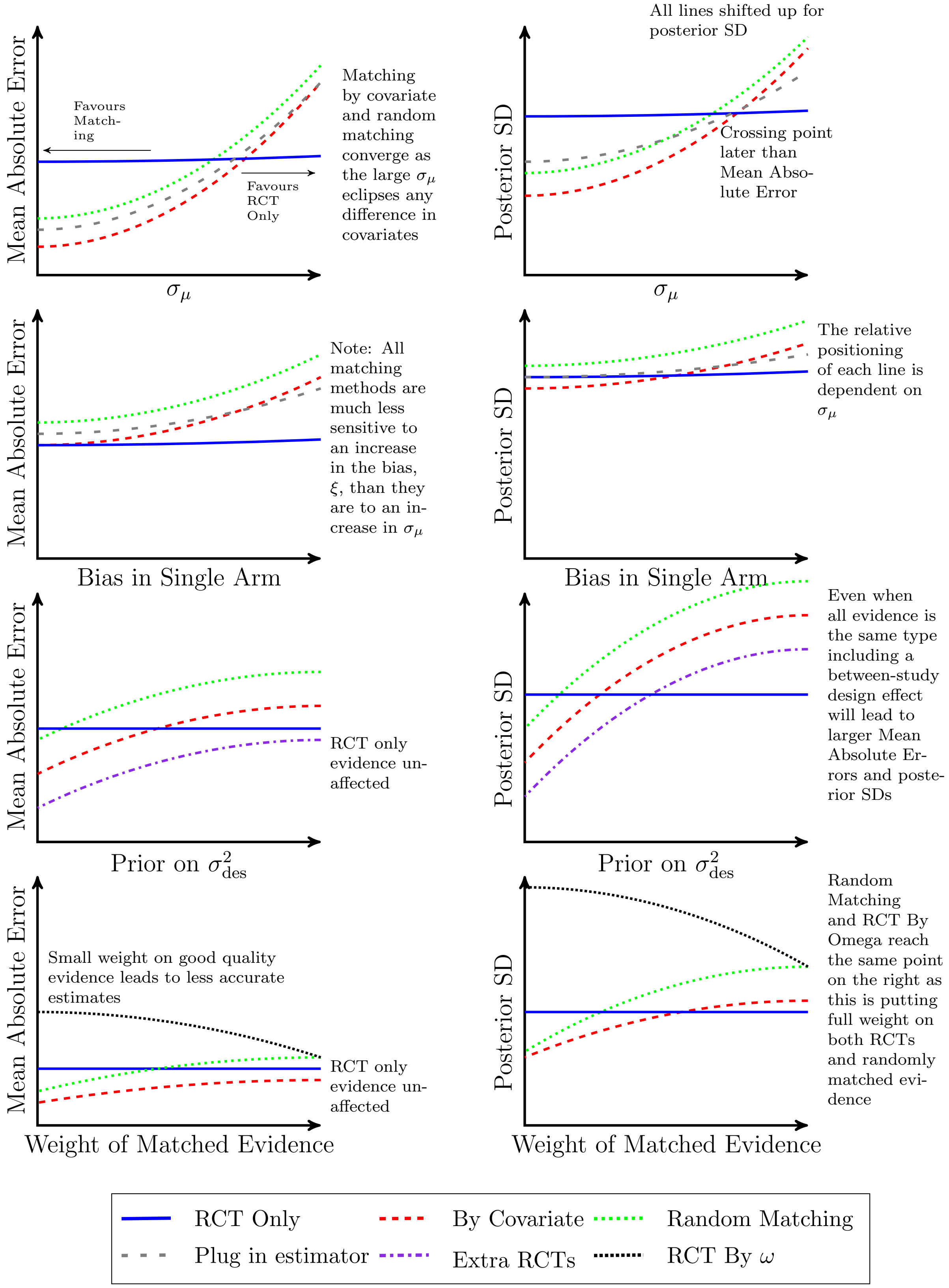
If there was sufficient RCT evidence there would be no need to include single-arm evidence as it is of lesser quality. However, sometimes only single-arm evidence is available. Single-arm evidence is currently being used in HTA, so it is important that we find the most suitable method, if any, for including single-arm evidence in an NMA, and that we understand the benefits and limitations of this type of evidence. Figure 3.15 shows a schematic representation of the association between some of the parameters considered in this chapter and the MAE and posterior SD.

Other factors such as the covariate effect, the treatment effect and the size of the RCTs can also affect the appropriateness of including matched evidence. For these we tried to choose the parameters in the simulation study so that they were as realistic as possible, and if anything they favoured RCT evidence. In the exploratory analysis we found that matching was most valuable when trials were small. Similarly, in the HCV infection example the matching method only showed superiority over the reduced method once four or five studies had been removed.

There is a concern that allowing the inclusion of single-arm evidence would disincentivise further RCT research from being undertaken. There is also the question of whether we are willing to accept a particularly biased estimate over no estimate at all. Therefore it makes sense to have the option to down-weight the single-arm evidence, so that it does not have the same impact as RCT evidence. Another possibility could be to only consider single-arm evidence for new treatments while RCTs are being undertaken. However, sometimes it may not be ethical to carry out the RCTs necessary to connect the network. For example, in the case of HCV infection where the new treatment regimens appear to be much better, it would not be in the interest of the patients to run an RCT involving new and old treatments regimens.

3.6.2 Recommendations

We recommend considering single-arm evidence when the variability between studies is small and when we can find an appropriate match. We can attempt to quantify this variability by examining the baseline effect across the RCTs. As the discrepancies between the studies increases, the incorrect assumption of exchangeablity between matched arms leads to a decreased accuracy and precision in the estimates. This form of evidence has a high possibility of being biased and should be used with caution. The estimates of between study variability for the networks



Matching by covariate and random matching converge as the large σ_μ eclipses any difference in covariates

All lines shifted up for posterior SD

Crossing point later than Mean Absolute Error

Note: All matching methods are much less sensitive to an increase in the bias, ξ, than they are to an increase in σ_μ

The relative positioning of each line is dependent on σ_μ

RCT only evidence unaffected

Even when all evidence is the same type including a between-study design effect will lead to larger Mean Absolute Errors and posterior SDs

Small weight on good quality evidence leads to less accurate estimates

RCT only evidence unaffected

Random Matching and RCT By Omega reach the same point on the right as this is putting full weight on both RCTs and randomly matched evidence

Figure 3.15: Schematic representation of the association between the parameters and the mean absolute error and posterior SD based on the results from the simulation study. The default parameters assumed in this graph (with the exception of when each parameter is varied) are based on the parameters used in the simualation study, namely $\sigma_\mu = 0.59$, no bias in single arm trials, prior on between-study design effect, $\sigma_{des} \sim Unif(0,2)$, full weight on single arm studies. Note that location and size of effects may depend on other parameters as described in the text. Therefore the relative positioning of some lines are subject to change.

analysed in this chapter were close to the crossing point where including matched studies produces more biased estimates, which indicates that this method may not be suitable for many networks where RCT data is also available. For treatments that are not in any RCTs we cannot quantify the bias versus using RCT only as this information simply is not available. However, in some cases including matched evidence will be the only option available to connect the network.

When undertaking single-arm matching there are a number of sensitivity analyses that should be carried out:

1. Quantify the variability between RCTs;
2. Compare the results using the best match to the results using random matches;
3. Adjust the weight of the matched evidence in the hierarchical model by increasing the variance inflation.

It is important to emphasise that any method for incorporating single-arm evidence cannot replace the unbiased approach of an RCT. Given the large potential monetary gain for a pharmaceutical company arising from a positive HTA recommendation, there is a clear incentive for pharmaceutical companies to ensure that their treatment looks as effective as possible. Single-arm trials are much more exposed to bias or manipulation than RCTs (Grieve et al. (2016)). Therefore, it is imperative that we remember this when making decisions based on single-arm evidence. While aggregate level matching may be helpful in estimating treatment rankings when no other evidence is available, we would not consider this type of evidence to be convincing enough to be the primary source of evidence when making reimbursement decisions. Although there are advantages in using single-arm evidence, the burden of proof must be on the pharmaceutical company to demonstrate the benefits of their treatments using methods that are as free from bias as possible.

3.6.3 Limitations and Extensions

When matching arms, we are making the assumption that we have considered all relevant effect modifiers and prognostic variables. However, this may not necessarily be the case. As we have seen in the case of Hepatitis C, not all studies record the same covariates, nor might all relevant covariates even be known to clinician when the trial is carried out. Therefore, making use of observational evidence to explore covariate effects further may be useful when checking the validity of the assumptions made during the matching process.

We analysed five networks of RCTs to estimate the between study variability in order to identify the typical amount of variability in a network. However, this may not be a representative sample as four of the five networks were taken from *R*, and these may have been chosen by the package authors as they have well behaved properties. In addition, only datasets with binary outcomes were chosen, as these were the most relevant to our simulation study. Finally, if we estimate the variability between studies by using the RCTs alone, there is no guarantee that the between-study effect σ_μ in the single-arm studies will follow the same distribution as the RCTs. As there are often multiple single-arm trials available for each treatment regimen, another method would be to study the variability between these single-arm trials.

In the hierarchical model we were concerned that choosing a disconnected network at each study design would lead to excessive influence from remote parts of the hierarchy. Since we have flexibility in which study arms to match we can force the matched network to be connected, as we had to do most of the time in the simulation study. However, doing this means we are using a less similar arm for matching. Whether we want to do this or not would depend on the difference between the two potential matches, which may require further exploration. However, we believe that this was the most conservative option to take in this scenario. A further issue is that the prior can add in extra uncertainty and the hierarchical model can lead to less accurate estimates than the pooled model as well, especially when each level is too sparse. In addition, given that we have only two study types, it can be quite difficult to estimate σ_{des}. Therefore, it may be worth considering an informative prior distribution formed by expert opinion as suggested by Efthimiou et al. (2017). A further option could be to include a bias term in the hierarchical for the matched data.

Although our worked example only uses binary covariates, this method can easily be applied to continuous covariates by using the mean in each trial. This work could be extended to match the SD of continuous covariates also. Issues with scale may arise when matching to multiple continuous covariates. In this situation the standardised mean difference should be used as in Austin (2011) and Flury & Riedwyl (1986). In this analysis we have only considered differences in patient populations. However, there are other factors that can account for differences between studies, for example differences in how treatment regimens were administered. This work could be extended to include these differences as well.

In this chapter we choose the closest match for each treatment regimen. However, there can be uncertainty regarding the choice of best match, as there may be other studies or arms that are nearly as similar as the chosen match. On the other hand, there may be some treatments regimens for which all potential matches are very dissimilar to the single arm trial. In this case we may want to consider excluding this treatment regimen from the analysis. Criteria for assessing a sufficiently similar match is discussed in Schmitz et al. (2018). In addition, the sensitivity analysis suggested in this chapter of choosing random matches gives us some idea of the variability that arises in the conclusions based on the choice of match.

In both the HCV infection network and the melanoma BRAF inhibitor network the best match is more conservative than a random match in the network, possibly due to single-arm trials having a disproportionately healthier patient population. The second line melanoma network has no clear bias in either direction, possibly because of the fact that there are a number of single-arm trials which have been matched, and are therefore closer to the average overall. It would be worth checking if the best match is usually more conservative than a random match in various networks. If so, it may indicate that single-arm trials are more likely to give better results than RCTs and reinforce the point that single-arm evidence should be used with caution and covariates should be matched as much as possible.

3.6.4 Conclusion

There is an increasing desire to use single-arm evidence where available. However, single-arm evidence is never going to be a perfect substitute for RCTs. Methods such as those explored in this chapter are already being implemented in clinical practice, as seen in Jaff et al. (2017). There is a high risk of bias in the conclusions from including this evidence type. Therefore it is imperative that there is a clear method for including this evidence, backed up by systematic investigation, and an indication for how much bias could potentially be introduced by using this evidence. We believe that this work provides such an approach method, in particular by following the steps and sensitivity analyses in Figure 3.1.

Chapter 4

Individual Patient Data (IPD)

The following chapter is based closely on Leahy et al. (2018), which has appeared in Research Synthesis Methods. It is based on joint work with Dr Aisling O'Leary, Dr Emma Gray, Dr Nezam Afdhal, Dr Scott Milligan, and Dr Malte Wehmeyer. Dr O'Leary and Dr Gray undertook the systematic review and provided guidance on the issues relating to HCV infection. Dr Afdhal, Dr Milligan, and Dr Wehmeyer provided IPD and provided guidance on the relevant datasets.

4.1 Introduction

While most trials publish only the summary statistics (Aggregate Data (AgD)) for patients as a whole, some may also be able to provide IPD. There are both benefits and drawback to IPD NMAs (Debray et al. (2016), Veroniki et al. (2016), Tierney et al. (2015), Jansen (2012), Van Walraven (2010), Sud & Douketis (2009), Stewart & Tierney (2002)). On the positive side IPD allows for a more in-depth and accurate analysis of the data, as we can explore how patient covariates influence the treatment effect, d_k, as defined in Figure 4.1. We can therefore account for differences in the patient covariates between arms. It also allows separation of within study associations of the covariates from across study associations. It facilitates harmonisation between trials, both in terms of outcome and analysis, and encourages input from the clinical investigators. However, IPD can be quite time consuming to obtain, and only certain types of IPD may be available which could potentially lead to bias. For example, IPD may only be available for observational studies, or perhaps certain investigators may be more likely to share their data than others. There are also a number of data protection issues that arise from using IPD, and as such it can take a long time before IPD can be released to secondary researchers. Finally, even when the IPD are obtained, much more computational power is required to perform the analysis. As it is unlikely that investigators will obtain a full IPD network, the best way to make use out of the data is to be able to combine both IPD and AgD in the same NMA. Therefore, a number of models have been developed for this purpose (Hong et al. (2018), Thom et al. (2015), Donegan et al. (2013), Saramago et al. (2012)).

A number of articles have reviewed the use of IPD in MAs and in NMAs. Debray et al. (2016) and Poppe et al. (2011) found that there are differences in results between IPD and AgD, and encourage researchers to consider using IPD. However, there remains a question as to whether the marginal benefit is worth the additional time involved. Tudur Smith et al. (2016) also found IPD-MAs to be beneficial, but note that "in many cases, similar results and conclusions can be drawn from IPD-MA and AgD-MA. Therefore, before embarking on a resource intensive IPD-MA, an AgD-MA should initially be explored and researchers should carefully consider the potential added benefits of IPD."

Therefore, this chaper aims to quantify the impact of including IPD studies in NMA. A simulation study is carried out to examine the impact of additional IPD studies on:

1. The accuracy of the point estimate of both the treatment effect and the covariate effect,

2. The posterior Standard Deviation (SD) from the model's estimate of both the treatment effect and the covariate effect,

3. The coverage probability, which is the proportion of the time that the Credible Interval (CrI) contains the true effect,

4. The DIC, which assesses model fit.

We compare the results of this simulation study for both RCTs and observational studies. We also analyse a HCV infection network consisting of three IPD observational studies to ascertain how different amounts or combinations of IPD studies affect results.

A simulation study has previously been undertaken by Jansen (2012). He found that the use of IPD does increase precision and reduce bias. We extend his work by examining the coverage probabilities and the DIC. We also explore a number of different model assumptions for the interaction between covariates and treatments.

This chapter is organised as follows: Section 2 describes the model development, construction of the simulation study, and the HCV infection network. Section 3 presents the results of both the simulation study and the HCV infection network. A general discussion and some recommendations are provided in Section 4.

4.2 Methods

4.2.1 Model Development

All notation used throughout this chapter is described in Table 3 at the beginning of this thesis. We extend the models described in Sections 2.9 and 2.10 to include IPD. We examine models relating to a binary outcome (e.g. death or cure), and a covariate which can either be binary or continuous. Following on from Donegan et al. (2013), we use the following nomenclature for the models we examined, that is independent, exchangeable, and identical interactions of the treatment with the covariates. Formally we note that independent refers to the situation where the effects are independently and identically distributed (i.i.d). The exchangeable model is one where the effects are conditionally i.i.d., and the identical model assumes that the effects are the same. This is shown in Figure 4.1f. As described in Chapter 2, an identical interaction can also be referred to as a *prognostic variable*, while independent and exchangeable interactions can be referred to as *effect modifiers*. To illustrate the interaction modelling options we take the example of a binary covariate that reduces the probability of being cured. If there is an identical interaction with all treatments then, in this case, one's probability of being cured decreases by a fixed amount regardless of which treatment is taken. However, another possibility is that certain treatments may target this covariate more effectively than others, and therefore possessing this covariate may be worse on one treatment than on another. In this scenario interactions can be considered either completely independent, or exchangeable, which means that they are still different but come from the same underlying distribution.

The first model assumes independent treatment by covariate interactions. The effect of treatment 1 (the reference treatment) is set to zero in the model with all other treatments being compared to treatment 1. For AgD the model is:

$$\text{logit}(p_{ij}) = \begin{cases} \mu_i & \text{if j=1} \\ \mu_i + \delta_{ij} + (\beta_{t_{ij}} - \beta_{t_{i1}})x_{ij} & \text{if j>1} \end{cases}, \tag{1a}$$

and for IPD the model is:

$$\text{logit}(p_{ijl}) = \begin{cases} \mu_i + \beta_{0_i} x_{ijl} & \text{if j=1} \\ \mu_i + \beta_{0_i} x_{ijl} + \delta_{ij} + (\beta_{t_{ij}} - \beta_{t_{i1}})x_{ijl} & \text{if j>1} \end{cases}, \tag{1b}$$

where p_{ij} is the probability of an event in the j^{th} arm of the i^{th} trial. This probability comes from the effect of the i^{th} study, μ_i, the treatment effect in the j^{th} arm of the i^{th} trial, δ_{ij}, the covariate effect interacting with each treatment, $\beta_{t_{ij}}$, and the proportion possessing the characteristic associated with the covariate of interest for a binary covariate, or the mean of a continuous covariate in the j^{th} arm of the i^{th} trial, x_{ij}. For the IPD part, let p_{ijl} be the probability of an event for the l^{th} patient in the j^{th} arm of the i^{th} trial and x_{ijl} is a binary indicator variable for the presence of the characteristic associated with the covariate of interest for patient l in the j^{th} arm of the i^{th} trial. In the case of the IPD dataset we can also include the trial-specific covariate effect β_0. As this model assumes that the interaction with the covariate is independent for each treatment, there is a separate prior for the covariate associated with each treatment. The covariates are centred in the model, as was done in Donegan et al. (2013).

The second model assumes that treatment by covariate interactions are exchangeable, i.e. that the interactions are different for each treatment, but that they come from a common distribution. With this assumption the model stays the same as above but the priors differ, as detailed later in this section.

The third model assumes that treatment by covariate interactions are identical. Hence the models are now:

$$\text{logit}(p_{ij}) = \begin{cases} \mu_i & \text{if j=1} \\ \mu_i + \delta_{ij} + \beta x_{ij} & \text{if j>1} \end{cases}, \tag{2a}$$

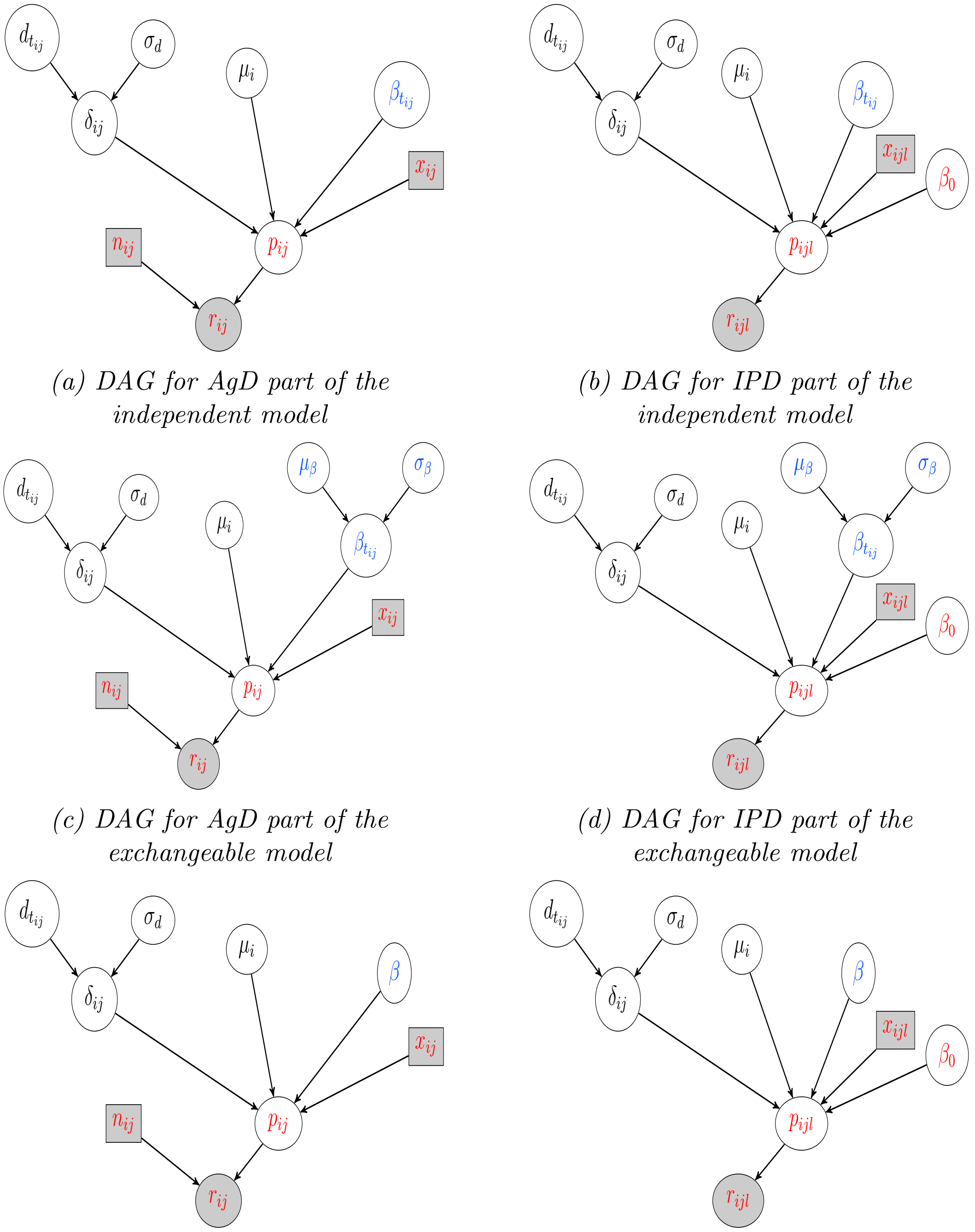
and:

$$\text{logit}(p_{ijl}) = \begin{cases} \mu_i + \beta_{0_i} x_{ijl} & \text{if j=1} \\ \mu_i + \beta_{0_i} x_{ijl} + \delta_{ij} + \beta x_{ijl} & \text{if j>1} \end{cases}. \tag{2b}$$

All three models are represented as directed acyclic graphs (DAGs) in Figure 4.1.

The prior distributions for the parameters chosen for this model are $\mu_i \sim N(0, 1.83^2)$, $\delta_{ij} \sim N(d_{t_{ij}} - d_{t_{i1}}, \sigma_\delta^2)$, $d_k \sim N(0, 1.83^2)$, $\sigma_\delta \sim \text{Unif}(0, 5)$, $\beta_0 \sim N(0, 1.83^2)$. We can adjust for trials with arms greater than two by following Dias, Sutton, Ades & Welton (2013). For the independent model the prior on each β is $\beta_k \sim N(0, 1.83^2)$ for each treatment by covariate interaction. For the exchangeable model the distribution for each β is $\beta_k \sim N(\mu_\beta, \sigma_\beta^2)$, with $\mu_\beta \sim N(0, 1.83^2)$ and $\sigma_\beta \sim \text{Unif}(0, 5)$. For the identical model the prior is $\beta \sim N(0, 1.83^2)$

The prior distributions for μ are chosen in order to have an approximate uniform distribution on the log odds ratio. Kass & Wasserman (1996) point out that the properties of a prior on one scale can differ when transformed to another scale. A seemingly vague prior such as $\mu_i \sim N(0, 100^2)$ is not vague on the inverse log odds scale, as most of the distribution is close to either 0 or 1. This is illustrated in Figure 4.2. However, $\sigma = 1.83$ was chosen such that two SDs on each side of the mean covered 95% of the distribution. σ_δ represents the between trial SD of the treatment effect, which covers (0, 1) on the probability scale (rounded to four



(a) DAG for AgD part of the independent model

(b) DAG for IPD part of the independent model

(c) DAG for AgD part of the exchangeable model

(d) DAG for IPD part of the exchangeable model

(e) DAG for AgD part of the identical model

(f) DAG for IPD part of the identical model

Figure 4.1: DAGs for the identical, exchangeable and identical models. Square nodes represent fixed quantities while the ellipses are stochastic nodes. The shaded nodes represent observed quantities. Nodes written in red change depending on whether the model is for AgD or IPD. Nodes written in blue change depending on which treatment by covariate interaction is assumed.

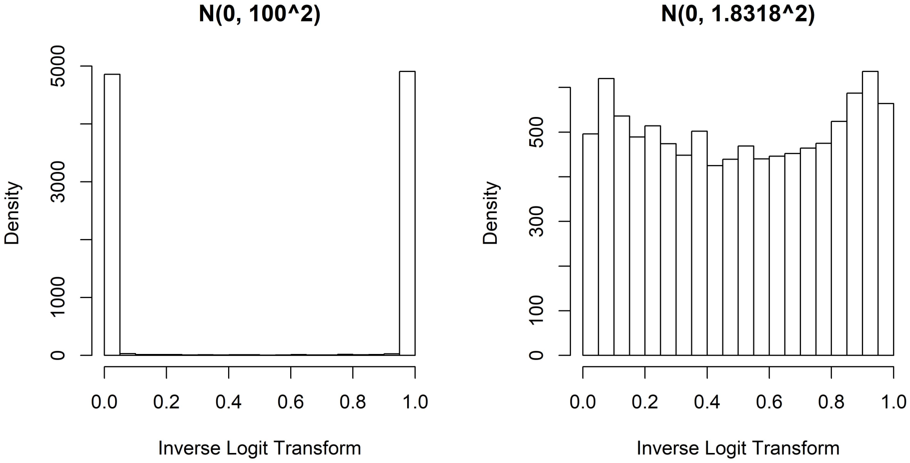


Figure 4.2: Simulated values from transformed probability distribution after applying the inverse log odds function

figures).

4.2.2 Simulation Study

A simulation study was performed in order to assess the impact of using IPD studies. IPD was simulated for eight studies and this was aggregated. We then used each data point in a nine iteration loop to compare the results from using the AgD from the eight studies with no IPD, to the results obtained from analysis including one more IPD study each time until we eventually used a full IPD network. This was repeated at least 100 times in order to reduce Monte Carlo error. The accuracy and precision of our model's estimate of the treatment effect and covariate effect, the coverage probabilities, and difference in DIC between models were assessed as the number of IPD studies increased. Models were run using Markov Chain Monte Carlo (MCMC) simulation in the JAGS software package (Plummer (2012)).

Two different interactions between treatments and covariates were examined:

- The covariate effects are identical for each treatment and one value is simulated from $N(0, 1.83^2)$.
- The covariate effects are exchangeable between treatments and the value for each treatment is simulated from $N(0, 1.83^2)$.

This distribution was chosen for the covariate effects as it is rather flat in the transformed space as described in Section 5.2.4. These were examined for both RCT and observational studies, as well as for a binary and a continuous covariate, which came to eight scenarios in total. For each scenario the three modelling assumptions of independent interactions, exchangeable interactions and identical interactions were explored. A tree diagram that illustrates all scenarios and modelling options for the simulation study is shown in Figure 4.3.

There were five treatment regimens in the network. Each study had two arms. In order to keep the studies as generic as possible the treatments were randomly assigned to each study arm. Other values were simulated as follows:

- μ was simulated from $N(0, 1.83^2)$ as it was a vague distribution that is rather flat on the inverse logit scale as seen in Figure 4.2.

- d was simulated from $N(0, 1.83^2)$ as it was a vague distribution that is rather flat on the inverse logit scale as seen in Figure 4.2.

- Number of patients generated from a rounded Unif$(250, 350)$ for the first arm of each trial. This was then doubled and patients were then randomly assigned to one of the arms so both arms did not necessarily have the same number of patients. This method was chosen to more accurately reflect the reality of patients being recruited first and then assigned to an arm. This was applied for both the RCTs and the observational studies. As a consequence of this approach the number of patients in each arm of a trial will be similar, although the covariate make-up of the patients will differ for observational studies as highlighted in the next step.

- In the case of binary covariates the probability of possessing the characteristic associated with the covariate was generated from Unif$(0.1, 0.9)$. The covariate for each individual patient was then generated from a Bernoulli distribution using the simulated probability. For RCTs one probability was generated per study. As a consequence the underlying distribution for both arms is the same. However, a known issue with observational studies is that there can be systematic differences between arms as they have not been randomly assigned (Schmitz et al. (2013)). In order to reflect these differences the probability of possessing the characteristic associated with the covariate was generated separately for each arm.

- In the case of continuous covariates the mean was simulated from Unif$(0.1, 0.9)$. The covariate for each individual patient was then generated from a truncated normal distribution. This was done to allow the continuous and binary scenarios to be as similar as possible in terms of the effect of the covariate, so as to allow comparison between the two scenarios. The covariate for each individual patient was bounded either below by 0 (if the mean was less than 0.5) or above by 1 (if the mean was greater than 0.5). The other side was

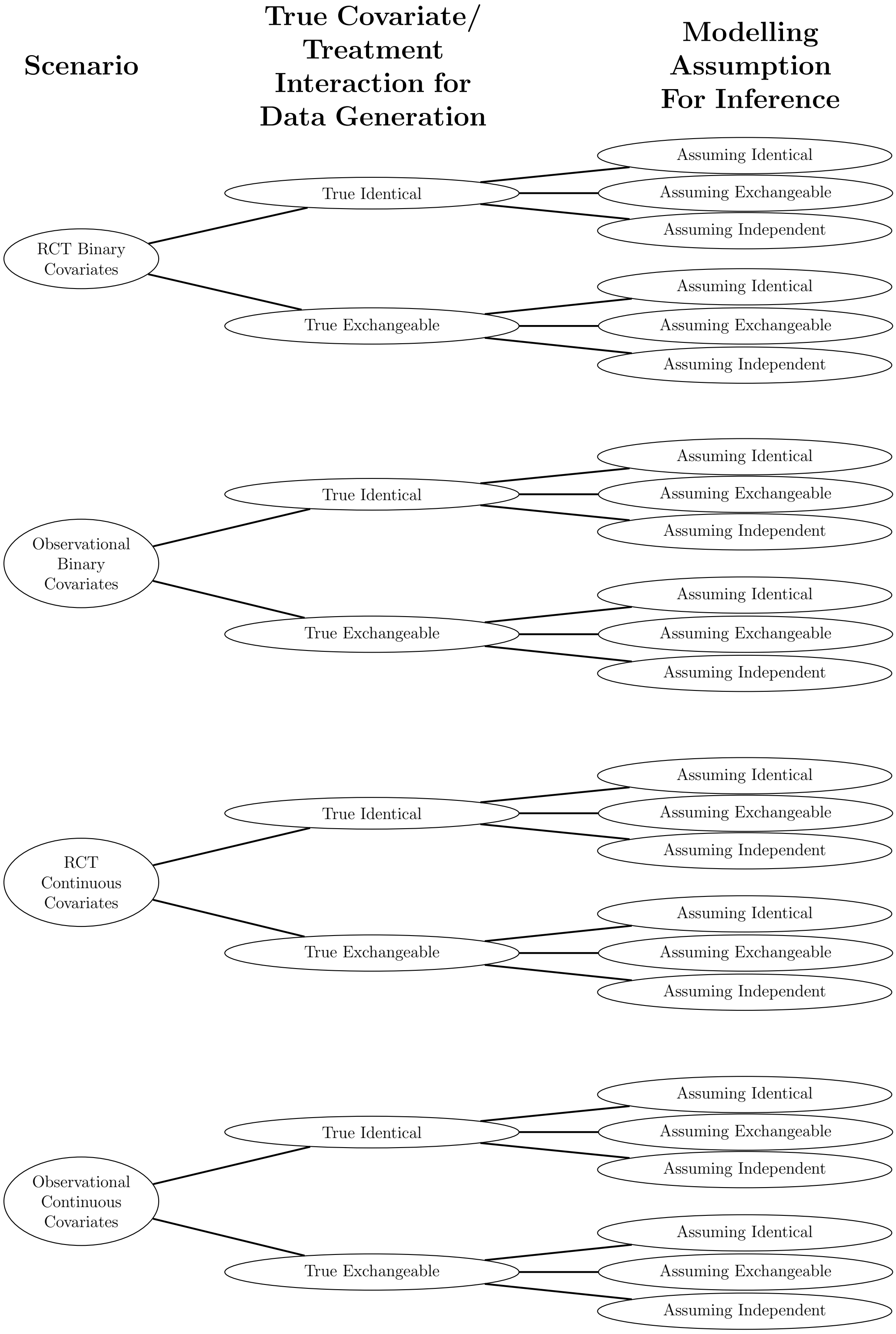
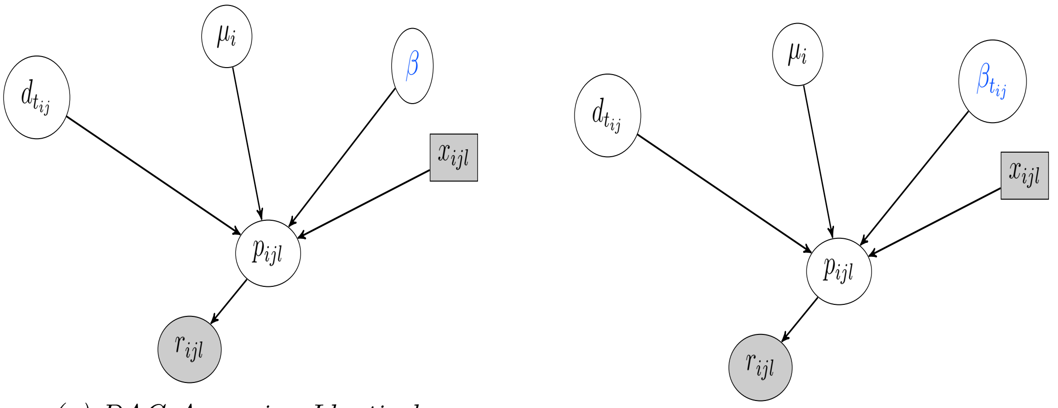


Figure 4.3: Tree diagram for all scenarios and modelling assumptions for the simulation study as described in Section 5.2.3



(a) DAG Assuming Identical interactions for the covariates

(b) DAG Assuming Exchangeable

Figure 4.4: DAGs for simulation study. Square nodes represent fixed quantities while the ellipses are stochastic nodes. The shaded nodes represent observed quantities. Nodes written in blue change depending on which treatment by covariate interaction is assumed.

bounded by the same distance from the mean, in order to maintain the symmetry of the distribution. For example, if the mean of the distribution was 0.3 then the individual patient covariates were bounded between (0, 0.6). The SD was set at a sixth of the range of the distribution, in order to limit the amount of truncated values. For RCTs we used the same mean for both arms in a study; however for observational studies we generated the mean separately for each arm. As was the case with the binary covariates, the distribution was generated at the study level for the RCTs and at an arm level for observational studies.

The probability of an event, p_{ijl}, in each study i and each arm j, was computed on the basis of the i^{th} study effect, μ_i, the treatment effect, $d_{t_{ij}}$, and the effect of each of the baseline covariates, $\beta_{ij}(x_{ijl})$, where x_{ijl} is a indicator variable for the presence of the characteristic associated with the covariate for patient l in arm j of study i:

$$\text{logit}(p_{ijl}) = \mu_i + d_{t_{ij}} + \beta_{t_{ij}}(x_{ijl}).$$

The observed event, r_{ijl}, is then calculated from a Bernoulli(p_{ijl}) for each patient l, in arm j of study i based on the above probability. DAGs describing the simulation study are shown in Figure 4.4.

The three models were tested to assess how well they predicted the true treatment effects and covariate effects. A burn-in of 20 000 iterations was tested for convergence by checking if the Gelman-Rubin statistic (Gelman & Rubin (1992)) was less than 1.1. Following this another 10 000 iterations were sampled for our estimates. If the convergence condition was not met the number of iterations was doubled (both for the burn-in and for the samples for estimation), and then tested again until the Gelman-Rubin statistic was less than 1.1. If the chains had not converged after a burn-in of 320 000 the corresponding simulation was excluded from the analysis. For the full IPD dataset it can take approximately five minutes to complete 20 000 iterations with binary covariates, and 17 minutes to complete 20 000 iterations with continuous covariates. We were prepared to allow the model run 16 times longer than this to converge. Given the amount of simulations required we decided it was not feasible to include chains that had not converged by this point. This occurred for less than 1% of chains. Reasons for non-convergence could include identifiablity of certain parameters or numerical issues. In this chapter a potential cause could be the indentifiability of the covariate-treatment interaction, particularly in the exchangeable model. For each set of simulated values we analysed all three models and all nine possibilities for the proportion of IPD studies. Therefore, if the chains did not converge for one of the models in a particular simulation the results for the two other models in the simulation, and all other eight possibilities for proportions of IPD studies were excluded from the analysis, in order to eliminate any potential bias due to differing simulations. The simulations took approximately 18 000 computing hours, run over parallel sessions. The exact number of simulations for each of the eight scenarios are detailed in the appendix. There are more simulations completed for the scenarios that involve binary covariates since each simulation for the continuous covariate required longer computation time.

Coverage probability was examined to ascertain how often the estimate of the treatment effect was in the 95% CrI. The mean absolute error (MAE) was also assessed by looking at the mean absolute difference between the estimates and the true values for treatments 2 to 5 as defined as follows:

$$\text{MAE}(\hat{d}) \equiv \frac{\sum_{s=1}^{S} \frac{\sum_{k=2}^{5} |d_{k_s} - \hat{d}_{k_s}|}{4}}{S}, \tag{4.3}$$

where $\hat{d_k}$ is our model's estimate of the effect of treatment k and S is the total number of simulations. Treatment 1 was excluded as this is our reference treatment, which was set to 0. When identical covariate interactions were simulated the accuracy of the estimates for β were assessed by:

$$\begin{aligned} \text{MAE}(\hat{\beta}) &\equiv \frac{\sum_{s=1}^{S} |\beta_s - \hat{\beta}_s|}{S}, \quad \text{assuming identical interactions,} \\ \text{MAE}(\hat{\beta}) &\equiv \frac{\sum_{s=1}^{S} \frac{\sum_{k=1}^{5} |\beta_s - \hat{\beta}_{k_s}|}{5}}{S}, \quad \text{assuming non-identical interactions,} \end{aligned} \tag{4.4}$$

where $\hat{\beta}$ is our model's estimate of the effect of the covariate under the identical interaction assumption, and $\hat{\beta_k}$ is our model's estimate of the effect of the interaction of the covariate with treatment k under the exchangeable and independent interaction assumptions. When exchangeable interactions were simulated the accuracy of the estimates for β was assessed by:

$$\text{MAE}(\hat{\beta}) \equiv \frac{\sum_{s=1}^{S} \frac{\sum_{k=1}^{5} |\beta_{k_s} - \hat{\beta}_{k_s}|}{5}}{S}. \tag{4.5}$$

Note that for the case of the identical model each $\hat{\beta_k}$ will be the same. Precision is obtained from the posterior SD. The DIC was calculated and models were ranked according to which had the lowest DIC. We highlight differences greater than three, as is consistent with Welton, Caldwell, Adamopoulos & Vedhara (2009).

4.2.3 Applied Example in HCV Infection

We used studies from a HCV infection network to test our conclusions. 36 potential studies, composed of 20 RCTs and 16 observational studies, were identified by a systematic review. We attempted to make contact with the authors of all the studies in order to collect as much IPD as possible. Our intention was to conduct a full IPD NMA. We received anonymised IPD from three observational studies (TRIO (Flamm et al. (2017)), Wehmeyer (Wehmeyer et al. (2014)), and ICORN (Gray et al. (2017))), which highlights the difficulty in obtaining IPD. We limited the network used in this chapter to these three observational studies in order to allow us compare a full AgD network to a full IPD network. Notwithstanding the risk of *data accessibility bias* in the type of studies for which IPD are available (Debray et al. (2015)), we proceed to analyse these data in order to illustrate our approach.

Using these three studies we had a network of 10 treatment regimens in total. However, most treatment regimens had been considered in only one study in our IPD network so we decided to restrict our network to treatments that appeared in more than one study. We restricted our analysis to patients with a subtype of the disease known as HCV infection genotype 1 (and therefore treatment regimens which are indicated for genotype 1), who had received treatment for at least 12 weeks. This resulted in a network of 4 treatment regimens as listed in Table 4.1. The treatment regimens in each study, as well as the number of patients per arm are detailed in Table 4.2. Figure 4.5 shows the network diagram.

Table 4.1: List of treatment regimens with abbreviations

Abbreviation	**Treatment Regimen**
TEL/PR	Telaprevir (+ pegylated interferon and ribavirin)
BOC/PR	Boceprevir (+ pegylated interferon and ribavirin)
PrOD±RBV	Paritaprevir boosted with ritonavir, ombitasvir and dasabuvir (with or without ribavirin)
SOF/LDV±RBV	Sofosbuvir and ledipasvir (with or without ribavirin)

Table 4.2: Number of patients per arm for each of the three studies. Dashes indicate that the treatment was not included in the study.

	TEL/PR	BOC/PR	PrOD±RBV	SOF/LDV±RBV
TRIO	-	-	459	3149
Wehmeyer	65	37	-	-
ICORN	203	94	183	315

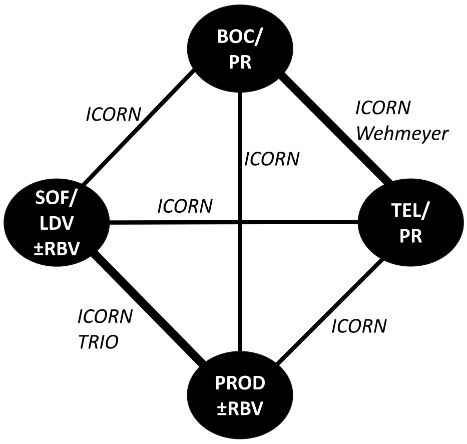


Figure 4.5: HCV infection Network of IPD Studies

4.3 Results

4.3.1 Simulation Study

In this section we compare the three models in terms of coverage probability, MAE of the estimates, and posterior SD. Ideally, we would like the coverage probability to be as close to the nominal CrI (in this case 95%) as possible, while minimising the MAEs of the estimates and posterior SDs. We also assess the effect of additional IPD on these outcomes and explore when the DIC can be used to choose between models. A summary of all results in this section is presented in Tables 4.8 and 4.9 in the discussion.

The results obtained when using a binary covariate and a continuous covariate followed the same trend. However, the effect of IPD was sometimes more pronounced for binary covariates. For the most part, there was no noticeable difference between the results from the simulated RCTs and observational studies. For clarity and brevity, only the results from the scenario using RCTs with a binary covariate are presented in the main text. Graphs for all four scenarios are included in the appendix.

Coverage Probability

Figure 4.6 shows the coverage probability. Low coverage probability indicates that we have underestimated the uncertainty that is present; whereas coverage probability that is too high indicates that we have overestimated uncertainty. In our results we use the term *misspecified model* to mean that in the inference phase of our simulation the model that was fitted differed from the model that was used to generate the data from that same simulation.

Note that the coverage for some misspecified models is quite far away from the nominal 95%. When the identical model is used incorrectly the coverage of both the treatment effect and the covariate effect is much lower than the nominal CrI. Meanwhile, for the covariate effect, the coverage of the independent model is much lower than the nominal CrI when the true interactions are identical. The coverage of both of these misspecified models becomes even worse when extra IPD are included. A decrease in coverage probability is due to either a less accurate estimate or smaller posterior SD. We will see in Section 4.3.1 that MAEs of the estimates do not tend to increase with extra IPD, so most of the decrease in coverage is due to smaller posterior SD. This indicates a potential issue with IPD; it can cause over confidence when the incorrect model is chosen.

When we only have AgD, the coverage of the covariate effect for the identical model is too low regardless of whether it is correct or not. However, when the interactions are in fact identical the coverage increases as we include more IPD. When most of the studies have IPD then the coverage is too high. An increase in coverage probability can be caused by a more accurate estimate or a larger posterior SD. We will see in Section 4.3.1 that posterior SDs do not tend to increase with extra IPD, and most of the increase in coverage is due to smaller MAEs of the estimates.

Excessively large posterior SDs can also be seen in most of points for the exchangeable model, especially for the coverage of the covariate effect when it is the correct model. Here the coverage is 100% in all cases, which indicates that the posterior SD should be much smaller. However, in this scenario the independent model is almost exactly on the 95% line.

For the treatment effect the inclusion of IPD shows evidence of a very slight increase of coverage, but sometimes this is further away from the nominal CrI.

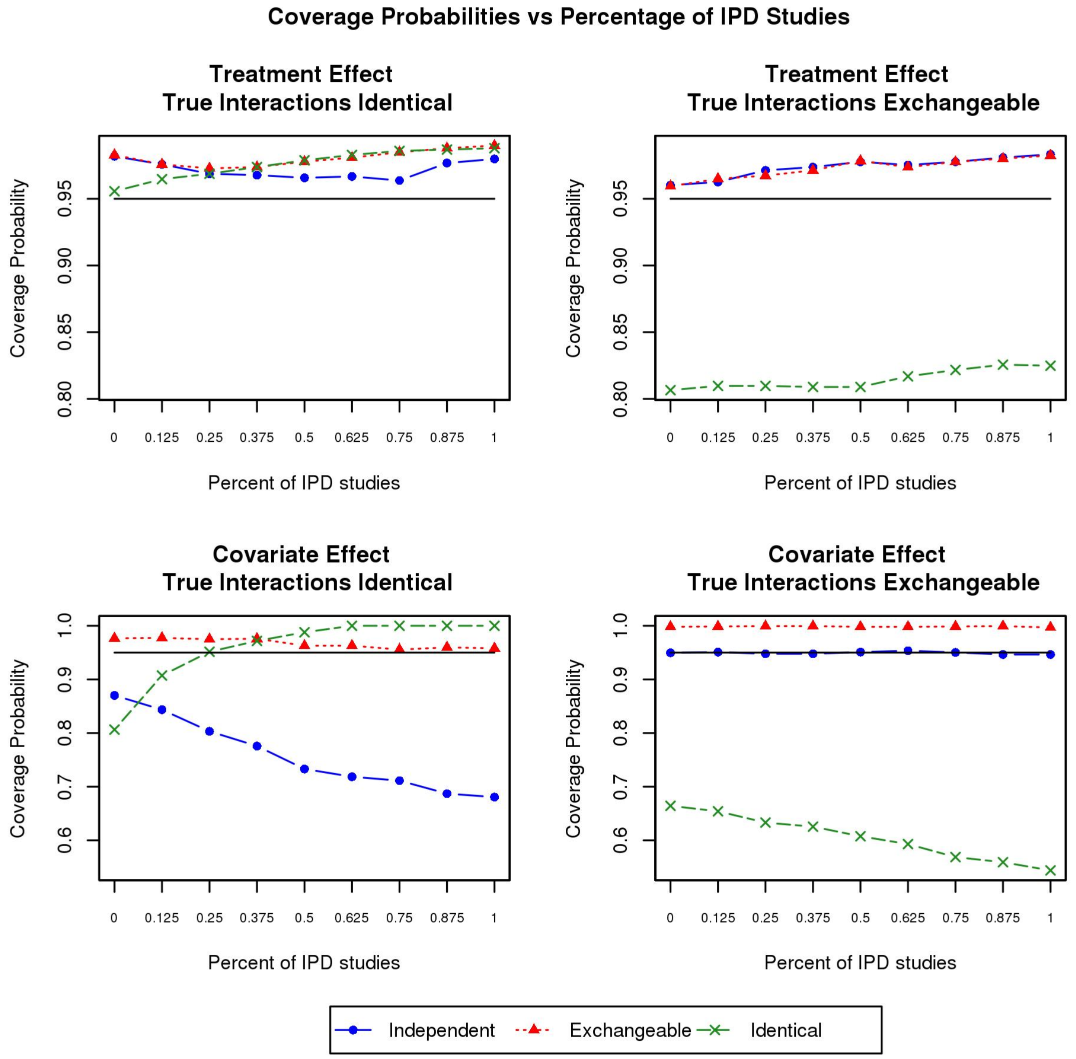


Figure 4.6: Coverage Probabilities of the 95% credible intervals for treatment effect and covariate vs percentage of IPD studies. Coverage above the solid black 95% line often indicates that posterior standard deviations are too conservative, while coverage below this often line indicates that posterior standard deviations are too precise. IPD can cause over confidence when the incorrect model is chosen.

Mean Absolute Errors and Posterior SDs

Figure 4.7 shows the MAEs of the estimates and posterior SD from the JAGS output of the treatment effect and covariate effect vs percentage of IPD studies. It demonstrates the importance of choosing the correct model, or at the very least, making a distinction between an identical interaction model (prognostic variable), as defined in Equations 2a and 2b, or a non-identical interaction model (effect modifier), as defined in Equations 1a and 1b. As we may have assumed, correctly choosing between an identical or non-identical model gives the smallest MAEs for both the estimates of the treatment effect and the covariate effect. However, the identical model produces the smallest posterior SDs for the covariate effect regardless of which model is correct, even though it has the highest MAE for both the estimate of the covariate and the treatment effect when the true interactions are in fact exchangeable. In this situation the posterior SD is quite a bit smaller than the MAE of the estimate, which would explain the poor coverage of the model. Again when true interactions are exchangeable, the posterior SD of the identical model for the treatment effect is also smaller than the MAE of the estimate, and is in fact the smallest posterior SD when the majority of studies are AgD. This discrepancy between the MAE of the estimate and posterior SD highlights the danger of assuming an identical model when this is not the case. The MAEs of the estimate from the non-identical models do not differ much between the two scenarios simulated from models assuming identical and non-identical interaction effects, but the error of the estimate obtained from the model assuming identical interaction is shifted quite far up when it is chosen incorrectly.

The posterior SD of the treatment/covariate effect for the exchangeable model is much higher than the corresponding MAE of the estimates in all cases. This is the reason why the coverage probability for the exchangeable model is too high. For covariate effects, the posterior SD is much higher for the exchangeable model than for the other two models. The exchangeable model incorporates all five interactions into one distribution, whereas the other models need to just estimate one effect (identical), or estimate five completely separate effects (independent).

The covariate effect is generated from a $N(0, 1.83^2)$ distribution. The mean of the absolute values from this distribution is 1.46, which is quite close to the MAE of the estimate of the covariate effect when the incorrect model is chosen. This implies that we gain very little information about the covariate when a non-identical model is chosen incorrectly.

All models have some MAEs of the estimates and posterior SD which have been decreased by including extra IPD. This can particularly be seen for the treatment effects and the non-identical models. The effect of IPD mainly follows a slightly convex downwards slope, which means that most of the benefit of IPD comes from the first few studies, with the marginal benefit decreasing the more IPD that is included in the NMA.

Deviance Information Criterion (DIC)

As we have demonstrated the importance of choosing the correct model, we need to check how well this can be undertaken by comparing the DIC across models, and whether IPD can be of benefit here. Examining Figure 4.8 we can see that

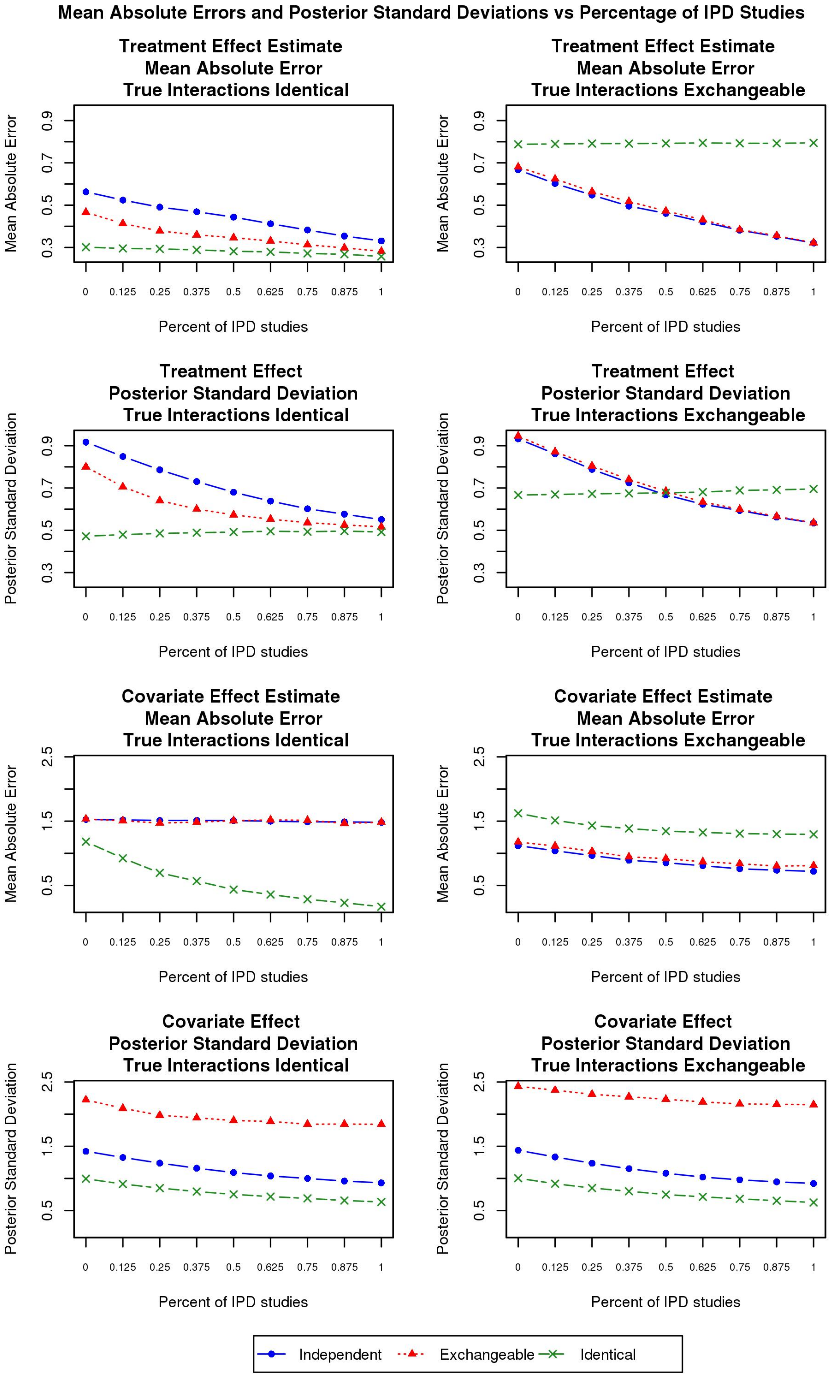


Figure 4.7: Mean absolute error (MAE) and posterior standard deviation (SD) of the estimate of the treatment effect and covariate effect vs percentage of IPD studies. As the amount of IPD increases, the MAEs of the estimates and posterior SDs decrease for a number of models. While the correct assumption will produce the lowest MAEs of the estimates, an incorrect model may produce smaller posterior SDs which are overly precise.

there is quite a large difference in our ability to choose between models, even when we include a small number of IPD studies, as compared to none. When the true interactions are identical, using the DIC we can correctly choose the identical model over the independent interaction model approximately 60% of the time with a full IPD dataset, whereas this can rarely be done with a full AgD dataset. In a small number of instances the DIC can also choose identical over exchangeable or exchangeable over independent. This is what we may have expected as the exchangeable assumption is somewhere between identical and independent. When the true interactions are exchangeable, the DIC is even more powerful. While there is no difference between models with the AgD dataset, using the DIC allows us to differentiate correctly between models up to 100% of the time with the full IPD dataset. Even including one IPD study has a considerable effect on the DIC. Again, we see the marginal benefit decreasing as more IPD are included in the NMA. There seems to be a clear distinction between either using the identical model or using one of the other two models. However, there is no difference between the exchangeable and independent models.

While we see some of the same patterns with continuous covariates it is less pronounced. As was the case with the binary covariate, when using the full AgD dataset we cannot distinguish between models. While there is more of a difference in DIC when IPD are included, differences in DIC do not occur with the same frequency as when using binary covariates (see Figure B.7 in the appendix).

To understand the reason why the DIC is more powerful for true exchangeable interactions, we examine the SD of the interaction of the covariate with the five treatments for the independent and exchangeable models, as seen in Figure 4.9. There is clearly a larger SD when the interactions are truly different. Therefore, if the model incorrectly assumes that there is a difference between treatments, the estimates will not be as diverse as in a truly exchangeable scenario, and so there is not as much of a difference between an identical and non-identical model in a truly identical scenario. We can also note that the SD in the exchangeable scenario increases with more IPD studies. So as the model gets more information about the covariates it can estimate them to be further apart. The true effects are simulated with an SD of 1.83, which is quite close to the estimate from a full IPD dataset. Conversely, in the truly identical scenario, the SD slightly decreases with more IPD, i.e., it gets closer to the true SD of 0.

Finally, we examine the MAE for the chosen model. In this case we only include simulations where the difference between the highest and lowest DIC is greater than three. We then choose the model with the lowest DIC as the chosen model. In this case we see that the MAE of the chosen model performs almost the

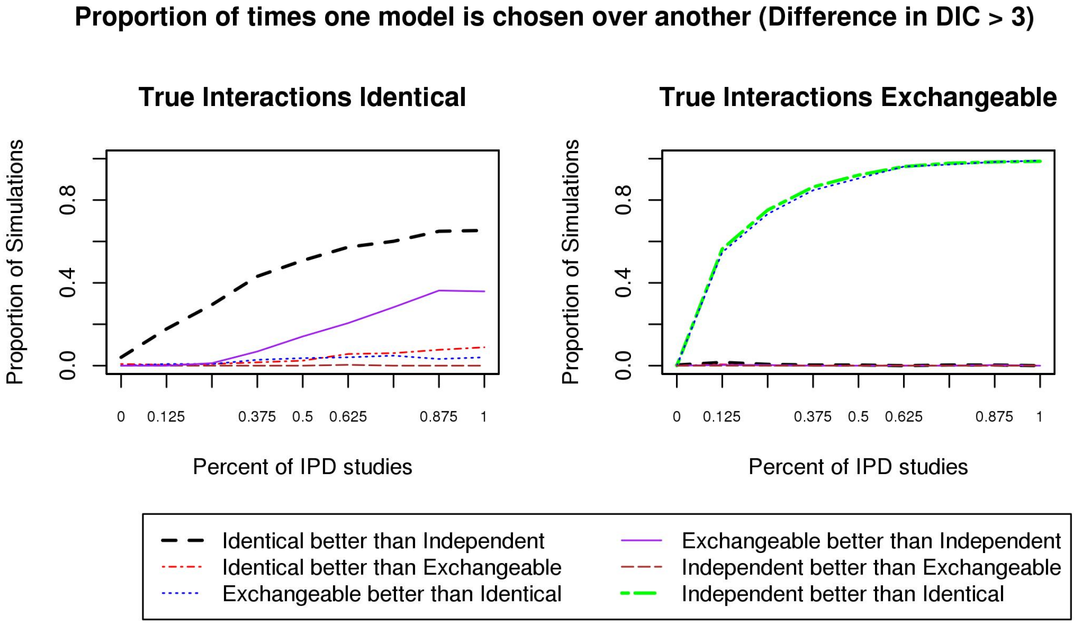


Figure 4.8: Proportion of deviance information Criterion (DIC) differences greater than 3 vs percentage of individual patient data (IPD) studies. The lines track the number of iterations when there is a meaningful difference between two models. There are seldom differences between the models with a full aggregate dataset. However, as the amount of IPD increases the correct model is identified more often (up to 100% of the time in the case of an exchangeable model).

same as the effect modifier models when the interactions are truly exchangeable, and somewhere between the effect modifier models and the identical model when the true interactions are truly identical. This echos what has been observed in Figure 4.8, in that we are unlikely to choose the incorrect model when interactions are truly exchangeable, but we may choose the incorrect model when interactions are identical.

RCT vs Observational Studies

The main difference between RCTs and observational studies is that the observational studies produce a more accurate estimate of an identical covariate effect when we have mainly AgD studies (see Figure B.5). Under the identical model the observational studies provide estimates at different covariate levels in the two arms and thus have a smaller MAE of the estimates when compared to RCTs, which will measure both arms at approximately the same level of the covariate because of randomisation.

We may have expected a greater difference in the accuracy of the treatment effect between the RCTs and the observational studies, as the RCTs are better quality trials that balance patient level covariates by design. While the MAEs of

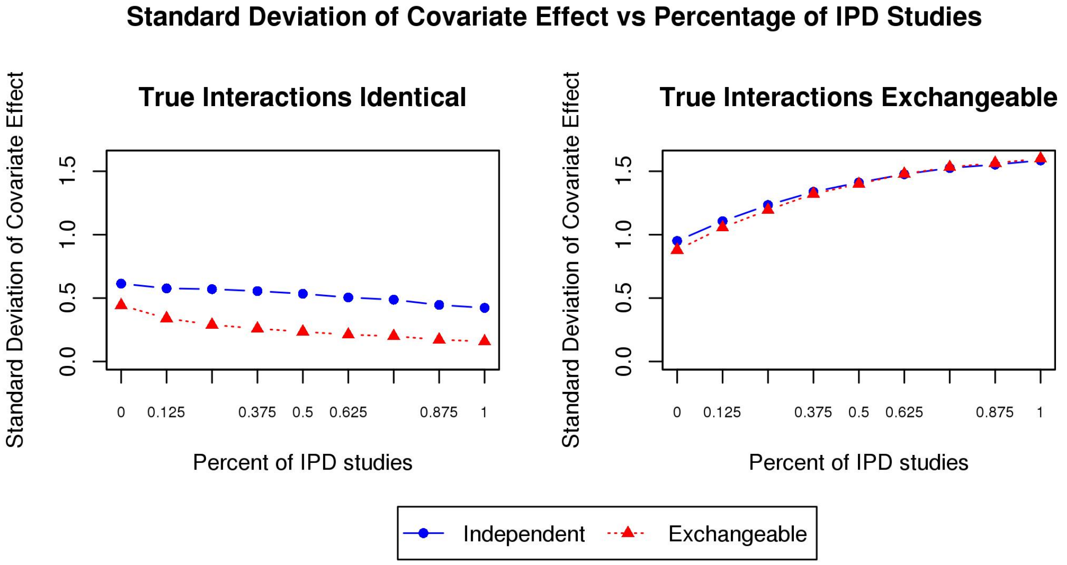


Figure 4.9: Standard deviation of the interaction of the covariate with the five treatments for the independent and exchangeable models vs percentage of individual patient data (IPD) studies. Exchangeable and independent models will estimate the covariate effects to be more different from each other when they are actually exchangeable as opposed to truly identical. As the amount of IPD increases the standard deviation comes closer to the true standard deviation of zero when true interactions are identical, while the standard deviation tends to the prior when true interactions are exchangeable.

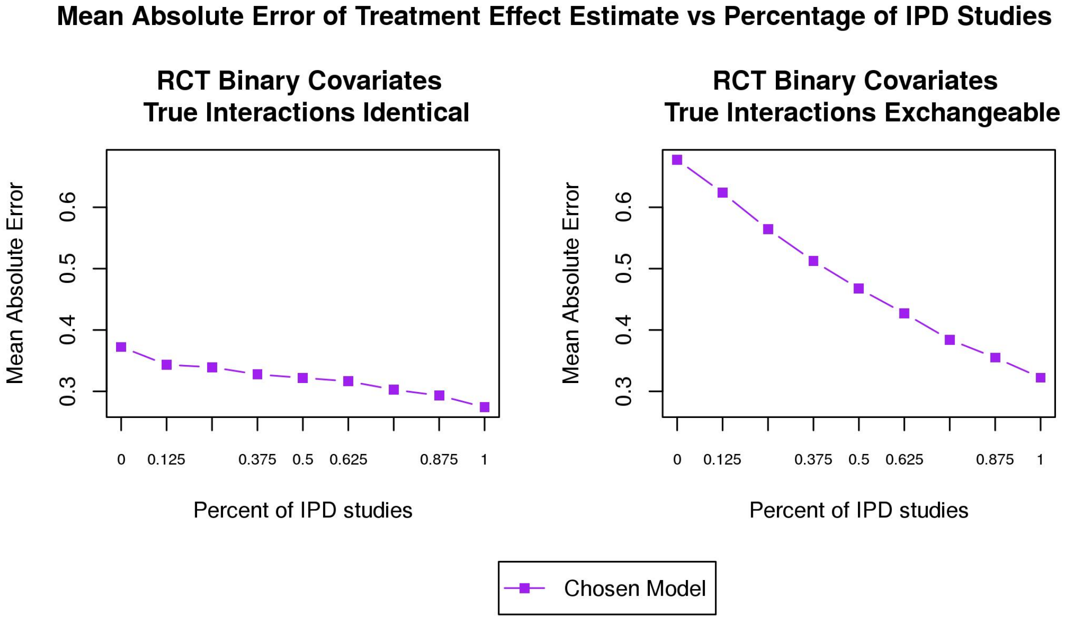


Figure 4.10: MAE for chosen model vs percentage of IPD studies. In this case we see that the MAE of the chosen model performs almost the same as the effect modifier models when the interactions are truly exchangeable, and somewhere between the effect modifier models and the identical model when the true interactions are truly identical.

the estimates are slightly higher in all scenarios for the observational studies as compared to the RCTs, the difference is not large enough that we can confidently conclude that it is not due to chance (see Figure B.3).

4.3.2 Applied Example in Hepatitis C Virus (HCV) Infection

We now present the results of all eight possible datasets for the HCV infection network. These consist of a full AgD dataset, a full IPD dataset, and the six mixed combinations of either one or two IPD studies out of the three possible IPD studies. We will also present the results using the three model assumptions of independent, exchangeable and identical interactions. A tree diagram that illustrates all dataset combinations and modelling options is shown in Figure 4.11.

Deviance Information Criterion (DIC)

Table 4.3 shows the DIC for the three models using the eight different datasets. Models with a smaller DIC are considered better. Examining the difference in DIC for one IPD study, we can see that ICORN is the only study that shows a difference between models; this is the only study to include all four treatments. In this case the DIC from the independent and exchangeable models are at least three lower than identical, which indicates a better model fit. Using IPD from either of the other two studies on their own does not allow us to distinguish between models. In fact, we only see a difference of greater than 3 between any models when we include IPD from ICORN. However, when IPD from TRIO is included in addition to IPD from ICORN the difference in DIC decreases. In fact, the difference between identical and exchangeable falls below 3 in this case. This possibly demonstrates some discrepancy between TRIO and ICORN. Overall we conclude that one of the non-identical (effect modifier) models is more appropriate in this case.

Credible Intervals and Posterior SDs

Results of all models and datasets are shown in Figures 4.12 and 4.13. All of these CrIs span zero. Of course, it is recognised that this is not the full evidence base for HCV infection, but it is used here for illustrative purposes. Including all AgD studies could lead to smaller CrIs, but this is outside the scope of this chapter. Figure 4.12 shows the treatment effects. Although there are some noticeable differences in point estimates for the efficacy of PrOD±RBV and SOF/LDV±RBV versus

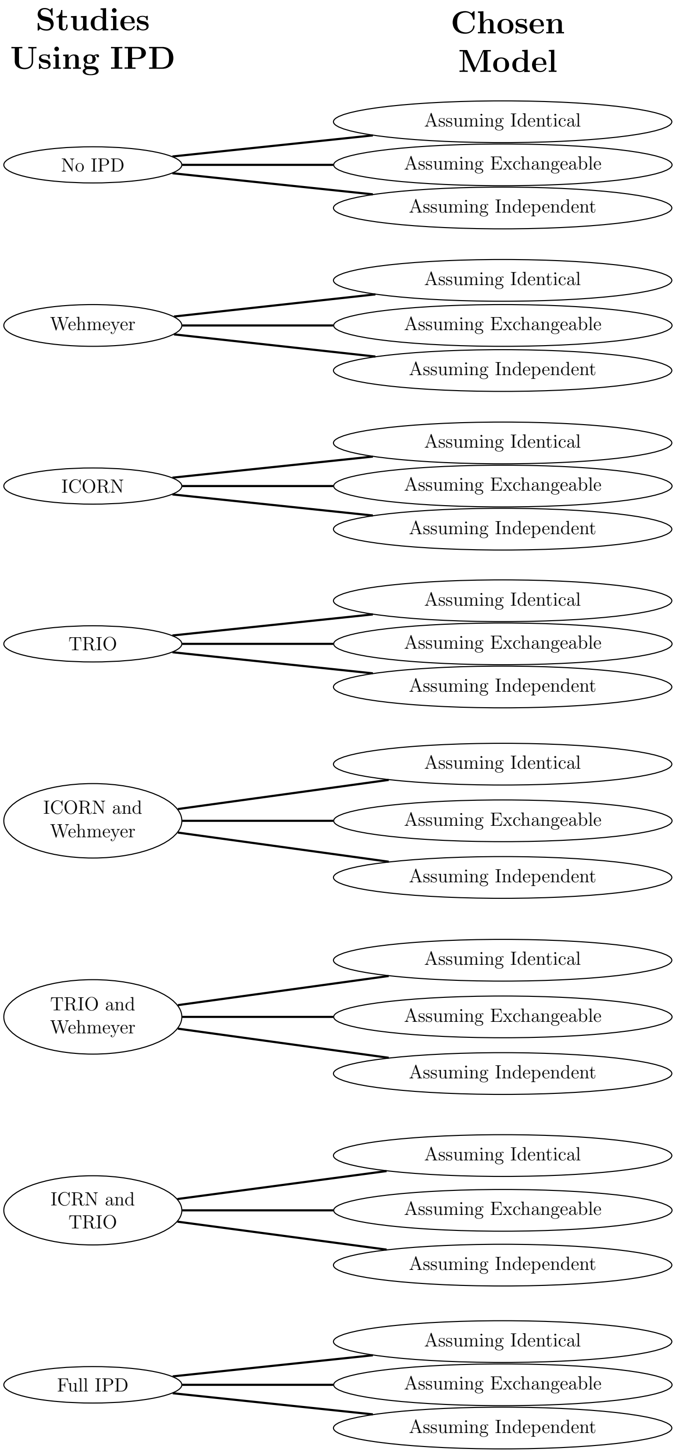


Figure 4.11: Tree diagram for all IPD and AgD combinations and modelling options for the real world HCV infection network as described in Section 4.2.3

Table 4.3: Deviance Information Criterion (DIC) for each model and dataset. Models with lower DIC are considered more appropriate. A full aggregate dataset does not allow us to differentiate between the models. The identical model is never chosen for any dataset.

DIC		**One IPD**			**Two IPD**			
	AgD	**TRIO**	**ICORN**	**WEH**	**IC&WE**	**TR&WE**	**IC&TR**	**IPD**
Identical	57.41	2,720.21	678.79	186.08	807.17	2,849.06	3,341.78	3,470.00
Exchangeable	56.85	2,720.36	674.14	186.18	802.86	2,849.24	3,339.51	3,467.67
Independent	56.42	2,720.16	672.85	186.37	801.05	2,850.07	3,337.30	3,465.49
Identical minus Exchangeable	0.56	0.16	4.65	-0.10	4.31	-0.18	2.28	2.32
Identical minus Independent	0.99	0.04	5.94	-0.29	6.12	-1.02	4.49	4.50
Exchangeable minus Independent	0.43	0.20	1.29	-0.19	1.81	-0.84	2.21	2.18

Notation: AgD (aggregate data), WEH (Wehmeyer), IC&WE (ICORN and Wehmeyer), TR&WE (TRIO and Wehmeyer), IC&TR (ICORN and TRIO), IPD (individual patient data)

TEL/PR, in each of the 24 model and dataset options, the ranking of the treatment regimens, in terms of attainment of SVR, is unchanged. SOF/LDV±RBV is ranked highest, followed by PrOD±RBV, TEL/PR, and finally BOC/PR. This is consistent with Gray et al. (2015), which uses all available AgD evidence, which also found SOF/LDV±RBV to be the best treatment, followed by PrOD±RBV. Figure 4.13 shows the covariate effects. The difference in the length of the CrIs between identical and exchangeable covariate interactions again highlights the need to check model choice if we wish to draw conclusions on the effect of the covariate. There is considerably more discrepancy between models for the covariate effects, with the identical models in particular being closer to the line of no effect. Depending on the model and dataset there are estimates on either side of this line.

The size of the credible intervals is solely dependent on the posterior SDs which is detailed in Tables 4.4 and 4.5. For all three model assumptions the posterior SD of each covariate effect decreases when including IPD. As we have identified that the identical model would not be appropriate for this network, the identical model may produce overly precise SDs, as shown in the coverage probability of the simulation study in Figure 4.6. The posterior SD of the treatment effects are sometimes increased and sometimes decreased. In cases where including extra IPD increases posterior SD of the estimated treatment effect, this may indicate between-study variability. It should be noted that the main decrease in posterior SD of the treatment effect is for the identical model, which is most likely not the correct model for this network.

Tables 4.6 and 4.7 show the effect of IPD and model choice on the posterior SD for each combination of IPD and AgD. Cells are colour coded based on the size of the posterior SD within each model (row). In Table 4.6 we see that including IPD

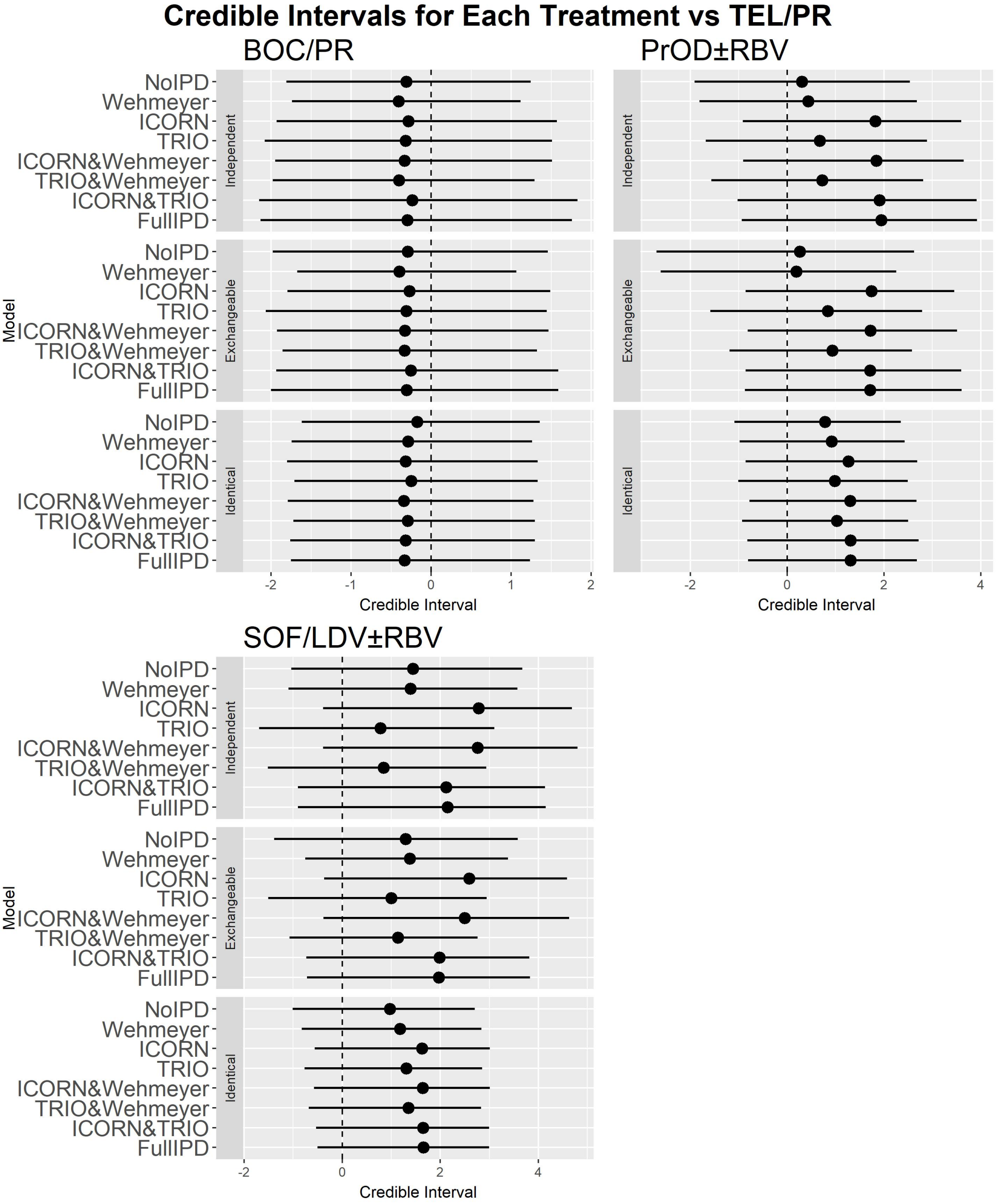


Figure 4.12: 95% credible intervals for each treatment relative to TEL/PR. The panel on the left hand side specifies the studies for which IPD are included. Points to the right of the vertical line of no effect indicate that the treatment is superior to TEL/PR. Treatment rankings are the same for all models and datasets. All lines span zero which indicates a non statistically significant difference.

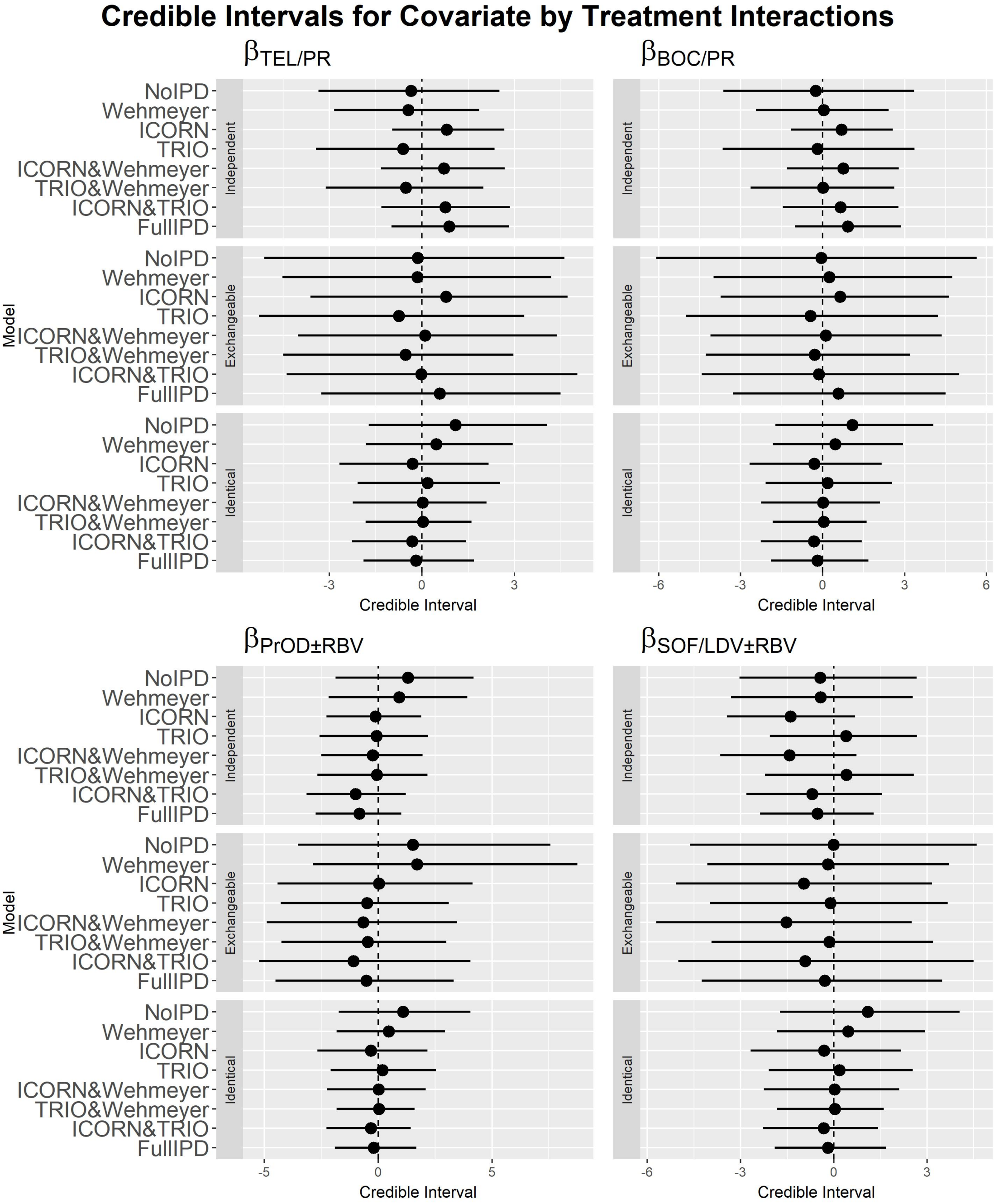


Figure 4.13: 95% credible intervals (CrI) for covariate interaction with each treatment. The panel on the left hand side specifies the studies for which IPD are included. The credible interval for the identical assumption is the same in each graph. CrIs are smaller when the amount of IPD are increased. The estimate of the covariate effect is quite dependent on which model or dataset is chosen.

Table 4.4: Posterior Standard Deviation (SD) for covariate effects. Percentage change shows the reduction between the full aggregate dataset and full individual patient dataset. Smaller posterior SDs within each row are coloured in green.

		AgD	**IPD**	**Percent Change**
Identical	β	1.48	0.91	-39%
Exchangeable	$\beta_{TEL/PR}$	2.58	1.82	-30%
	$\beta_{BOC/PR}$	2.90	1.99	-31%
	$\beta_{PrOD\pm RBV}$	2.85	1.98	-30%
	$\beta_{SOF/LDV\pm RBV}$	2.30	1.96	-15%
Independent	$\beta_{TEL/PR}$	1.53	0.99	-35%
	$\beta_{BOC/PR}$	1.80	0.98	-45%
	$\beta_{PrOD\pm RBV}$	1.57	0.95	-39%
	$\beta_{SOF/LDV\pm RBV}$	1.44	0.94	-35%

Table 4.5: Posterior Standard Deviation (SD) for treatment effects versus TEL/PR. Percentage change shows the reduction between the full aggregate dataset and full individual patient dataset. In the case of treatment effect this is sometimes negative to represent an increase. Smaller posterior SDs within each row are coloured in green.

		AgD	**IPD**	**Percent Change**
Identical	**BOC/PR**	0.71	0.72	1%
	PrOD±RBV	0.85	0.84	-1%
	SOF/LDV±RBV	0.93	0.85	-9%
Exchangeable	**BOC/PR**	0.84	0.84	0%
	PrOD±RBV	1.35	1.10	-19%
	SOF/LDV±RBV	1.23	1.12	-9%
Independent	**BOC/PR**	0.77	0.91	18%
	PrOD±RBV	1.14	1.23	8%
	SOF/LDV±RBV	1.19	1.27	7%

from any one study decreases the posterior SD of the covariate effects for identical and exchangeable. For the independent assumption, some of the posterior SDs are not decreased with IPD from only Wehmeyer or TRIO, as neither one of these has studied all four treatments. We again see some evidence of disagreement between studies as some posterior SDs are increased when IPD are included from more than one study. Overall, however, the full IPD dataset has posterior SDs that are quite close to the smallest posterior SD, which indicates that there is not considerable disagreement. Once again we see that the exchangeable model produces the highest posterior SDs, especially for covariate effects. In Table 4.7 we see that the use of IPD does not make as much difference to the posterior SD of the treatment effect. As highlighted in the simulation study, this may indicate heterogeneity between studies and therefore we should use the full IPD where possible so as not to rely on overly precise posterior SDs.

Table 4.6: Posterior Standard Deviation (SD) for covariate effects for all eight datasets. Smaller posterior SDs within each row are coloured in green.

			One IPD			Two IPD			
		AgD	**WEH**	**ICORN**	**TRIO**	**IC&WE**	**TR&WE**	**IC&TR**	**IPD**
Identical	β	1.48	1.23	1.26	1.19	1.13	0.89	0.98	0.91
Exchangeable	$\beta_{TEL/PR}$	2.39	2.16	2.06	2.07	2.24	1.91	2.27	1.98
	$\beta_{BOC/PR}$	2.90	2.17	2.08	2.32	2.25	1.92	2.28	1.99
	$\beta_{PrOD\pm RBV}$	2.85	2.72	2.15	1.81	2.27	1.86	2.32	1.98
	$\beta_{SOF/LDV\pm RBV}$	2.30	2.03	2.14	1.83	2.23	1.85	2.32	1.96
Independent	$\beta_{TEL/PR}$	1.49	1.20	0.92	1.49	1.01	1.30	1.06	0.97
	$\beta_{BOC/PR}$	1.80	1.24	0.94	1.78	1.04	1.33	1.08	0.98
	$\beta_{PrOD\pm RBV}$	1.57	1.58	1.03	1.17	1.15	1.24	1.08	0.95
	$\beta_{SOF/LDV\pm RBV}$	1.44	1.48	1.04	1.17	1.12	1.22	1.07	0.94

Notation: AgD (aggregate data), WEH (Wehmeyer), IC&WE (ICORN and Wehmeyer), TR&WE (TRIO and Wehmeyer), IC&TR (ICORN and TRIO), IPD (individual patient data). White represents largest posterior SD, darker shades of green represent smallest posterior SD within each row .

Table 4.7: Posterior Standard Deviation (SD) for treatment effects relative to TEL/PR for all eight datasets. Smaller posterior SDs within each row are coloured in green.

			One IPD			Two IPD			
		AgD	**WEH**	**ICORN**	**TRIO**	**IC&WE**	**TR&WE**	**IC&TR**	**IPD**
Identical	BOC/PR	0.71	0.71	0.75	0.72	0.73	0.72	0.73	0.72
	PrOD±RBV	0.85	0.83	0.87	0.87	0.84	0.84	0.85	0.84
	SOF/LDV±RBV	0.93	0.90	0.89	0.90	0.86	0.86	0.86	0.85
Exchangeable	BOC/PR	0.84	0.66	0.76	0.84	0.78	0.76	0.83	0.84
	PrOD±RBV	1.35	1.26	1.04	1.08	1.06	0.92	1.08	1.10
	SOF/LDV±RBV	1.23	1.03	1.23	1.11	1.23	0.94	1.12	1.12
Independent	BOC/PR	0.77	0.68	0.81	0.89	0.80	0.79	0.93	0.91
	PrOD±RBV	1.14	1.12	1.11	1.16	1.11	1.11	1.23	1.23
	SOF/LDV±RBV	1.19	1.19	1.27	1.20	1.27	1.14	1.27	1.27

Notation: AgD (aggregate data), WEH (Wehmeyer), IC&WE (ICORN and Wehmeyer), TR&WE (TRIO and Wehmeyer), IC&TR (ICORN and TRIO), IPD (individual patient data). White represents largest posterior SD, darker shades of green represent smallest posterior SD within each row.

4.4 Discussion

Through our simulation study we found a variety of benefits of using IPD, even when IPD are not available for all studies. Therefore, when conducting an NMA one should not abandon an IPD approach, even if IPD are unavailable for some studies. The proportion of IPD that is available, in combination with any available AgD, could have great benefits to an NMA. It decreases MAEs of estimates and posterior SDs in most cases. It also increases our ability to choose between model assumptions through the DIC. This is almost impossible using just AgD. When the model is misspecified, IPD may cause overconfidence in posterior SD leading to poor coverage probabilities. However, using IPD to choose between models should limit the frequency with which an incorrect model is used. Therefore we recommend using IPD when and where possible.

Tables 4.8 and 4.9 summarise the findings of the simulation study for both model comparisons and the use of IPD, in order to identify situations when model choice is particularly important and when IPD are most useful. Ideally we wish to have the coverage probability of the CrI as close as possible to the nominal probability (in this case 95%), while keeping the MAEs of estimates and posterior SD as small as possible.

Table 4.8: Summary of findings from simulation study on differences between models.

	Assuming Identical	**Assuming Exchangeable**	**Assuming Independent**
Correct coverage probability for treatment effects?	Above when identical, below when exchangeable	Close or above in both cases	Close or above in both cases
Correct coverage probability for covariate effects?	Below when exchangeable, below with AgD for identical and increases to above with inclusion of IPD	Too high when exchangeable, correct or above when identical	Correct to slightly below when exchangeable, poor when identical (decreases with IPD)
Correct model can be identified by using the DIC?	Not using full AgD, up to 60% of the time (over independent) with IPD (less when covariate is continuous)	Not using full AgD, up to 100% of the time with IPD (less when covariate is continuous)	NA
Ranking of MAE of treatment effect estimates?	Smallest when identical, largest when exchangeable	Middle when identical, joint smallest when exchangeable	Largest when identical, joint smallest when exchangeable
Ranking of MAE of covariate effect estimates?	Smallest when identical, largest when exchangeable	Joint largest when identical, joint smallest to middle when exchangeable	Joint largest when identical, smallest or joint smallest when exchangeable
Size of posterior SD of treatment effects?	Smallest when identical and when using a continuous covariate, smallest using majority AgD for exchangeable with a binary covariate (but overly confident), other models decrease with inclusion of IPD	Middle when identical, joint largest for continuous, joint largest for continuous using majority AgD but decreases with inclusion of IPD	Largest when identical, joint largest for continuous, joint largest for continuous using majority AgD but decreases with inclusion of IPD
Size of posterior SD of covariate effects?	Smallest for both (but overly confident when exchangeable)	Largest for both	Middle for both

Table 4.9: Summary of findings for all models from simulation study on effect of Individual Patient Data (IPD)

	True Identical	**True Exchangeable**
Improve coverage probability for treatment effects?	Very slight increase (often further away from nominal)	Slight improvement (often further away from nominal)
Improve coverage probability for covariate effects?	Decrease when model is misspecified as independent (especially for binary covariates), increases identical model which eventually surpasses nominal probability.	No noticeable effects
Improves ability for DIC to identify correct model?	Increases correct model choice	Increases correct model choice
Improve estimate of treatment effects?	Yes, especially for non-identical model (but to a lesser extent when using a continuous covariate)	Yes, especially for non-identical model (but to a lesser extent when using a continuous covariate)
Improve estimate of covariate effects?	Yes, especially for identical (prognostic variable) model	Yes, especially for non-identical (effect modifier) model
Increase precision of treatment effects?	Yes, for non-identical models	Yes, for non-identical models
Increase precision of covariate effects?	Yes, for non-identical models	Yes, for non-identical models

Dias, Sutton, Welton & Ades (2013) have highlighted that identical models are often used in practice and for this reason their chapter explored the properties of only the identical model in greater depth. However, as we have seen from the simulation study, assuming identical interaction effects when they are not identical can be quite problematic due to the overly precise estimates. On the other hand, in the simulation study, through examining the DIC, the identical model is very rarely incorrectly identified as the best model, so if this model has the lowest DIC it is quite likely to be correct.

If a model has been identified as non-identical it appears that an independent model may be preferable over an exchangeable model, due to the unnecessarily high SDs from the exchangeable model. However, as the independent model can produce quite low coverage of the covariate effect, it is worth investigating whether an identical model is more appropriate. Additionally, we have only considered cases with five treatments in the simulation study and four treatments in the HCV infection network. Perhaps if there were many more treatments, an exchangeable model may be more appropriate than independent. Similarly, we have only examined the scenario when all the treatments are either identical or non identical. There are possibly many cases where there are a few different classes of treatments. Therefore these could possibly be modelled as a joint distribution with independence or exchangeability between groups of treatments and exchangeable or identical interactions within groups.

It should be noted that the covariate-treatment interactions in the exchangeable model can also be considered as independent and identically distributed. In particular, given the vague hyper-priors on these parameters in this chapter, there is only a small difference between the exchangeable and the independent models. It could be worth examining more informative priors for the hyper-parameters on the exchangeable interaction to increase precision.

In the HCV infection example the ranking of treatments remained the same regardless of the amount of IPD included. This is consistent with the conclusion of Tudur Smith et al. (2016), that in many cases similar conclusions can be drawn from both AgD and IPD NMA. However, the use of IPD also allowed us to conclude that we should use a model which assumes non-identical interactions between the treatments and the covariates. The use of IPD also reduced the posterior SDs for the covariate effects, and occasionally reduced posterior SDs for the treatment effects, which may indicate some heterogeneity between studies. By including the IPD we can gain more information about covariates, (although it is worth pointing out that the covariates are centred in our models, and therefore interpretation of the results should be done with this in mind). Increased information on the covariates either produce more precise estimates of effect, or allow us to identify between-study heterogeneity. Of course, part of the motivation for meta-analysis is that no one study can provide complete information about treatments. Therefore, identifying and including heterogeneous studies is crucial as it provides a more accurate picture of the effect of treatments on a wider scale.

4.4.1 Conclusion

Overall, we would recommend using IPD when available, particularly due to the added benefit of choosing the correct model. In the case where no IPD are available, we would recommend trying to obtain IPD from at least one or two datasets to ensure that the model is specified correctly. This could also lead to more accurate and precise estimates. While any extra IPD study is valuable, the marginal benefits decrease as more IPD are included in the NMA. In practice all studies should be included when undertaking an NMA, including those with just AgD.

Chapter 5

Matching Adjusted Indirect Comparison (MAIC)

The work in this chapter has been reviewed by Research Synthesis Methods and referees comments have been addressed for resubmission there.

5.1 Introduction

As discussed in Chapter 4, if IPD is available for some, or all, trials in an NMA, then incorporating this IPD into an NMA is considered the gold standard of evidence synthesis, as it allows a more in-depth analysis of the data, and accounts for differences in covariates between trials. However, the situation can often arise where a researcher has IPD for trials concerning a particular treatment (for example from a sponsor), but none for other trials. In this case one can reweight the IPD so that the covariate characteristics in the IPD trials match that of the aggregate data (AgD) trials, in what is known as a matching adjusted indirect comparison (MAIC) (Signorovitch et al. (2010, 2012)). MAIC allows one to account for the differences in covariates between trials, and provides insight into the potential outcome of the researcher's IPD trial, if it had been carried out in the trial population to which it is being matched. However, there are many potential downsides to this method, and the subjective nature that can be involved in identifying covariates for matching can potentially leave this method open to bias or even manipulation. Given that this is an increasingly popular method, it is important to be able to identify situations when it is appropriate and situations when it is not appropriate. This importance is also expressed in Phillippo et al. (2016), as they identify the need for comprehensive simulation studies to explore the properties of the method.

We undertake an MAIC for a connected network of treatments for multiple myeloma in newly diagnosed patients (ndMM) post-Autologous Stem Cell Transplant (ASCT), where the outcome is progression-free survival. The results of this MAIC (i.e. the aggregated data of each reweighted IPD study) are then treated as data in a Bayesian NMA. We investigate the reliability of the methods and results through a simulation study, which mirrors the ndMM network. MAIC can be carried out using a number of different outcome models. Tremblay et al. (2016), Van Sanden et al. (2017), and Kühnast et al. (2017) have all used survival models, with Van Sanden et al also using the reweighted MAIC data in a Bayesian setting. Kähnast et al evaluate the use of MAIC using two studies in a simulation study in a frequentist setting and found MAIC to be particularly useful when effect modifiers are present in dissimilar populations. Belger et al. (2015) conducted a simulation study with more than two studies in a frequentist setting using a continuous outcome. We extend these works to a Bayesian NMA with a time-to-event outcome, while also assessing the options available when we have multiple IPD studies.

This chapter aims to quantify the impact of reweighting IPD studies before running an NMA in a Bayesian setting with a time-to-event outcome. A simulation study is carried out to:

1. Quantify the effect of MAIC in NMAs using median and hazard ratio (HR) models.

2. Investigate two different options of weighting covariates. Firstly, within separate IPD trials, or secondly, using one weighting method across all IPD trials.

We compare the results of the MAIC method to using a standard NMA, and to using a mixed Agd/IPD model, which allows us to incorporate extra information about the covariates of interest, without the requirement for a researcher to carry out a post hoc weighting of the data.

This chapter is organised as follows: Section 2 describes the model development, construction of the simulation study, and the ndMM network. Section 3 presents the results of both the simulation study and the ndMM network. A general discussion and some recommendations are provided in Section 4.

5.2 Methods

In this section we will describe the method of MAIC, an application to ndMM and details of the simulation study. As a number of different models are implemented, we describe the details of the NMA models at the end in Section 5.2.4. All notation used throughout this chapter is described in Table 4 at the beginning of this thesis.

5.2.1 Matching Adjusted Indirect Comparison

In this chapter we consider binary covariates only. Suppose we have an AB-IPD trial with IPD comparing treatment A to treatment B, and a BC-AgD trial with aggregate data only comparing treatment B to treatment C. To match to the target (BC-AgD) trial we reweight the IPD trial such that the proportion of patients possessing the characteristic associated with each covariate in the IPD trial match the proportion of patients possessing the characteristic associated with each covariate in the target trial. The outcome for the reweighted AB-IPD trial to the target BC population is given by:

$$\hat{Y}_{k(BC)} = \frac{\sum\limits_{l=1}^{N_{k(AB)}} Y_{lk(AB)} w_{lk}}{\sum\limits_{l=1}^{N_{k(AB)}} w_{lk}},$$

where the weight w_{lk} assigned to the patient l receiving treatment k is equal to the odds of being enrolled in the BC trial versus the AB trial, $Y_{lk(AB)}$ is the outcome for patient l receiving treatment k in the AB-IPD study and $N_{k(AB)}$ is the number of patients assigned to treatment k in the AB-IPD study. For further information see Phillippo et al. (2016).

This method may be appropriate when the effect of the treatment is dependent on the characteristic that the patient possesses. In this case the covariate is an *effect modifier*, which means there is a non-zero covariate-treatment interaction. When the covariate-treatment interaction is zero this means the covariate is a *prognostic variable* and it is not recommended to adjust for these types of variables (Phillippo et al. (2016)) in a connected network (also known as an anchored comparison), due to the unnecessary increase in variance associated with MAIC.

5.2.2 Applied Example in Newly Diagnosed Multiple Myeloma

The ndMM network consists of three IPD studies comparing Lenalidomide to Placebo/Observation (Len-Placebo), one AgD study comparing Len-Placebo, and one AgD study comparing Thalidomide to Placebo (Thal-Placebo) (McCarthy et al. (2012), Attal et al. (2012), Palumbo et al. (2014), Morgan et al. (2012), Jackson et al. (2016)). Figure 5.1 shows the network diagram.

IPD information was available for the following binary covariates:

- Age: < 60 vs ≥ 60.

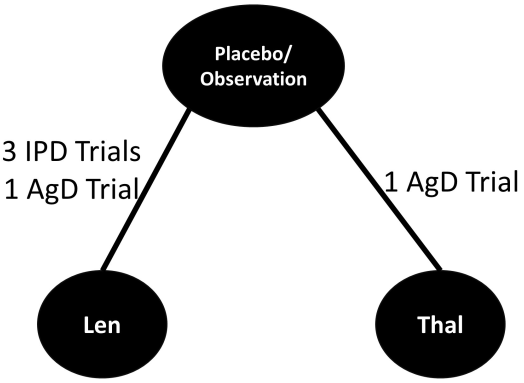


Figure 5.1: Newly diagnosed multiple myeloma network. Abbreviations: IPD = individual patient data; AgD=aggregate data; Len=lenalidomide; Thal=thalidomide.

- International Staging System (ISS) stage: I/II vs III.
- Adverse Risk Cytogenetics: Present vs Absent.
- Response post-ASCT (Response): Complete Response/Very Good Partial Response (CR/VGPR) vs Other.
- Gender: Male vs Female.

To ascertain which covariates were effect modifiers and which were prognostic variables we compared models assuming identical, exchangeable or independent covariate-treatment interactions using the DIC. It has been shown that it can be difficult to distinguish between models using AgD alone (Leahy et al. (2018)). Consequently, we used IPD models to aid in this decision-making. In some cases, studies were missing patient level covariate information for a proportion of the patients in the study. In this case we imputed the missing covariate information by conditioning on the other covariates.

As there are multiple IPD studies comparing Lenalidomide to Placebo/Observation, there are two different reweighting options:

1. MAIC Separate Trials: The IPD within each of the three studies are reweighted such that the AgD of each reweighted study matches the AgD trial as follows:
   (a) Within each IPD study reweight the IPD, such that the proportion of patients possessing the characteristic associated with each covariate in each IPD study matches the Thal-Placebo study.
   (b) Generate aggregate data from each reweighted IPD study.
   (c) Combine the aggregate data from the reweighted Len-Placebo studies and the AgD studies using NMA.

2. MAIC Pooled Trials: All IPD studies are pooled together for the reweighting element, such that the AgD of the studies combined matches the AgD trial, but they are put into the NMA as separate trials as follows:
    (a) Pool IPD from the three IPD studies together.
    (b) Reweight the IPD irrespective of study, such that the IPD from the three studies combined matches the Thal-Placebo trial.
    (c) Separate the reweighted IPD back into the three original studies.
    (d) Generate aggregate data from each reweighted IPD study.
    (e) Combine the aggregate data from the reweighted Len-Placebo studies and the AgD studies using NMA. Note that this contrasts with a naive pooling approach where studies are not separated out again before the NMA.

We wish to stress the importance of including the reweighted studies in the NMA as separate studies in both methods to ensure the original randomisation is still intact. Note that MAIC pooled trials is a less stringent requirement than MAIC separate trials, as only the combination of three studies need to reflect the covariate distribution of the AgD study, rather than each study having to reflect the covariate distribution of the AgD study, as is the case for MAIC separate trials. For all models when carrying out the NMA we included all five studies. These are the three IPD Len-Placebo trial (reweighted in the case of MAIC), the AgD Len-Placebo trial, and the AgD Thal-Placebo trial. When carrying out the MAIC, although we could plausibly match the IPD trials to the AgD Len-Placebo trial, in this case we are assuming that the population of interest is the Thal-Placebo trial. The NMA was carried out using Markov Chain Monte Carlo (MCMC) simulation in the OpenBUGS package (Spiegelhalter et al. (2014)).

The proportion of patients possessing the characteristic associated with each covariate in each trial is detailed in Table 5.1. In order to decide which covariates we should adjust for we obtained an estimate for the extent of bias from each covariate, as recommended in Phillippo et al. (2016). We calculated the difference in the covariate-treatment interaction for lenalidomide versus thalidomide from the fixed effect IPD models. We also calculated the degree of imbalance for which we need to adjust, by calculating the difference in the proportion of the patients possessing the characteristic associated with the covariate in the Thal-Placebo trial (Morgan), versus the average proportion of patients possessing the characteristic associated with the covariate in the Len-Placebo IPD trials. The imbalance of each covariate is multiplied by the difference in the covariate-treatment interaction for

Table 5.1: Proportion of patients with the characteristic associated with each covariate in each of the Len-Placebo trials and the Thal-Placebo AgD "target" trial in the newly diagnosed multiple myeloma network for patients post-autologous stem cell transplant. All trials compared an active treatment to placebo. For trials with IPD, covariates are computed from the available data, which includes some missing data, hence proportions may differ slightly to what was reported at the aggregate level.

	Active Treatment	**Age: <60**		**ISS Stage: III**		**Response: CR/VGPR**		**Cytogenetics: Present**		**Gender: Male**	
		Treatment	Placebo	Treatment	Placebo	Treatment	Placebo	Treatment	Placebo	Treatment	Placebo
McCarthy	Len	0.57	0.58	0.26	0.22	0.61	0.71	-	-	0.52	0.56
Attal	Len	0.64	0.63	0.22	0.17	0.59	0.58	0.17	0.10	0.55	0.59
Palumbo	Len	0.64	0.74	0.12	0.14	0.40	0.38	0.40	0.32	0.48	0.60
Jackson (AgD Only)	Len	-	-	0.29	0.26	0.73	0.73	0.51	0.44	-	-
Morgan	Thal	0.56	0.60	0.31	0.36	0.75	0.72	0.39	0.46	0.63	0.66

lenalidomide versus thalidomide to obtain the bias. These results are shown in Table 5.3

5.2.3 Simulation Study

A simulation study was carried out in order to assess the impact of using MAIC. Results were assessed by examining:

1. the mean absolute error (MAE) between the estimated effects and the true simulated effects.

2. the posterior SD, as reported in the JAGS output.

3. the effect on the estimate of the between study heterogeneity, i.e. the estimate of the difference in the relative treatment effects between trials.

4. the coverage probability, which is the proportion of the time that the CrI contains the true effect.

Four trials were simulated. Three of these trials were IPD trials comparing treatment A to treatment B (AB-IPD trials), and the fourth trial compared treatment B to C, with only the AgD available for the NMA (BC-AgD trial). This was set up to mimic the ndMM network. However we did not include any AB-AgD trial as this could dilute the effect of the MAIC, thus making it more difficult to analyse the results of the simulation study and to understand the true effect of MAIC. The network diagram for the simulation study is shown in Figure 5.2. We varied the covariate-treatment interaction (scenario 1) and the distribution of covariates in the trials (scenarios 2-4) as detailed below. Monte Carlo error was examined and is illustrated on the graphs in Section 5.3.1. The accuracy and precision of our model's estimate of the treatment effect was assessed as the parameters

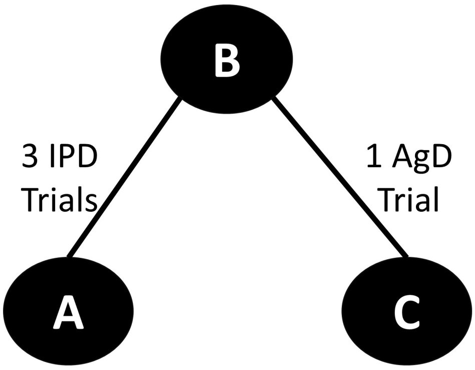


Figure 5.2: Network diagram for simulation study. Abbreviations: IPD = individual patient data; AgD=aggregate data.

varied. The simulation study was run in the JAGS software package (Plummer (2012)), rather than OpenBUGS, as it was carried out through a high performance computing cluster, which better supports JAGS.

Parameters were set in the simulation study as follows:

- The baseline risk in each study, the treatment effect and the overall covariate effect were simulated from $N(0, 0.5)$.
- The SD of each covariate effect between treatments (i.e. the covariate-treatment interaction) varied from 0-0.8. When this was set to zero it means that the treatment effect was the same for all levels of the covariate (or, in this case of a binary covariate, the treatment effect is the same regardless of whether or not the patient possesses the characteristic associated with the covariate), i.e. the covariate is a *prognostic variable.* Any other value for the standard deviation (SD) means that the treatment has a differing effect depending on whether the patient possesses the covariate or not, i.e. the covariate is an *effect modifier.* For simulations where this is not varied this value was set to 0.8. This means that two standard deviations of the covariate-treatment interaction range from 0.17-0.83 on the probability scale, which we believe sufficiently covers the difference in effect that a covariate could plausibly have on a range of treatments.
- When the covariate-treatment interaction was varied (scenario 1), the simulation set-up mimicked the number of patients in each trial and the proportion of those patients possessing the characteristics associated with the covariates in the ndMM network Section 5.2.2. We included three covariates; ISS stage, response, and age. We excluded cytogenetics as this was not recorded in the McCarthy trial.

- For scenarios where we varied the proportion of patients possessing the characteristics associated with the covariates (scenarios 2-4) we considered only two covariates for simplicity. These are detailed in Table 5.2. The covariates have the same proportion in each trial arm. However, these are assigned to patients individually.

- For scenarios 2-4 the number in each IPD arm was set to 200, and the number in each AgD arm was set to 240. These figures were similar to the numbers in the ndMM network, when the total number of patients in AB-IPD trials was averaged over the three trials.

Any parameter that was static in a scenario was only simulated at the beginning of each loop and held constant until all data points in the loop had been sampled, in order to reduce variance between the data points.

Table 5.2: Proportion possessing the characteristic associated with the binary covariate in each arm for the simulation study. The parameters were sampled in a loop in the range detailed below at the specified increments. Parameters where the increments are statics are fixed for that scenario. The values in scenarios 2 and 3 were chosen so that the fixed proportions were in the centre of the considered range. The values in Scenario 4 were chosen to allow the average proportion possessing the characteristic associated with the covariate to differ sufficiently between the AgD and IPD trials and within the IPD trials.

		Range	**Increments**
Scenario 2: Varying BC-AgD Covariate	**AB-IPD(1)**	0.4	Static
	AB-IPD(2)	0.5	Static
	AB-IPD(3)	0.6	Static
	BC-AgD	0.1-0.9	0.1
		Range	**Increments**
Scenario 3: Varying AB-IPD Covariate	**AB-IPD(1)**	0.1-0.9	0.1
	AB-IPD(2)	0.1-0.9	0.1
	AB-IPD(3)	0.1-0.9	0.1
	BC-AgD	0.5	Static
		Range	**Increments**
Scenario 4: Varying AB-IPD Between Studies	**AB-IPD(1)**	0.45	Static
	AB-IPD(2)	0.45-0.9	0.05
	AB-IPD(3)	0.45-0.1	-0.05
	BC-AgD	0.9	Static

The time-to-event (T_{ijl}) for each individual l, in study i and arm j, was simulated by $T_{ijl} \sim \text{Exp}(\lambda_{ijl})$ where the rate λ is given by:

$$\log(\lambda_{ijl}) = \mu_i + d_{t_{ij}} + \beta_{1,t_{ij}}(x_{1,ijl}) + \beta_{2,t_{ij}}(x_{2,ijl}),$$

where μ_i is the study effect in trial i, $d_{t_{ij}}$ is the treatment effect in arm j of study i, and the effect of covariate 1 is given by $\beta_{1,t_{ij}}(x_{1,ijl})$, where $\beta_{1,t_{ij}}$ is the covariate interaction with the treatment in arm j of study i, and $x_{1,ijl}$ indicates whether patient l in arm j of study i possesses the characteristic associated with covariate

1. The same notation follows for covariate 2. In order to imitate a real world trial, censoring time was simulated by $C_{ijl} \sim \text{Unif}(0, \max(T_i))$. Whether or not an individual was censored was decided by the minimum of time-to-event or censoring. This resulted in approximately 20% of patients being censored.

In total nine models were tested to assess how well they predicted the true treatment effects. These were:

1. HR model:
   - (a) Standard NMA Model: Unadjusted weights.
   - (b) Standard NMA Model with Covariate: Including a term for the average covariate per arm in the model. In this model we only use aggregate data and assume that the covariate is a prognostic variable due to limited data points.
   - (c) MAIC Separate Trials: The IPD within each of the three studies are reweighted such that the AgD of each reweighted study matches the AgD trial.
   - (d) MAIC Pooled Trials: All IPD studies are pooled together for the reweighting element, such that the AgD of the studies combined matches the AgD trial, but they are put into the NMA as separate trials.
2. Median model:
   - (a) Standard NMA Model: Unadjusted weights.
   - (b) Standard NMA Model with Covariate: Including a term for the average covariate per arm in the model. In this model we only use aggregate data and assume that the covariate is a prognostic variable due to limited data points.
   - (c) MAIC Separate Trials: The IPD within each of the three studies are reweighted such that the AgD of each reweighted study matches the AgD trial.
   - (d) MAIC Pooled Trials: All IPD studies are pooled together for the reweighting element, such that the AgD of the studies combined matches the AgD trial, but they are put into the NMA as separate trials.
   - (e) Mixed AgD/IPD Model: IPD is used in the model where possible, (i.e. for the AB IPD trials), and we use AgD otherwise (i.e. for the BC AgD trial). We also model each covariate as an effect modifier, assuming an independent interaction between each treatment and covariate. This

is the similar to the independent model described in Chapter 4, but applied to a time-to-event outcome, rather than a binary outcome.

The models were tested to assess how well they predicted the true treatment effects. A burn-in of 20 000 iterations was tested for convergence by checking if the Gelman-Rubin statistic (Gelman & Rubin (1992)) was less than 1.1. Following this another 10 000 iterations were sampled for our estimates. If the convergence condition was not met the number of iterations was doubled (both for the burn-in and for the samples for estimation), and then tested again until the Gelman-Rubin statistic was less than 1.1. If the chains had not converged after a burn-in of 320 000 the corresponding simulation was excluded from the analysis. If the chains did not converge for one of the models in a particular simulation, or if there were numerical problems, for example, if a trap error was reported in JAGS, the results for all other models compared in the simulation, and all other data points in the loop were excluded, in order to eliminate any potential bias due to differing simulations. Less than 5% of simulations could not be used for each scenario for these reasons, which mainly affected the HR model.

The accuracy of the estimate was assessed by looking at the mean absolute error (MAE) between the estimates and the true values for treatments B and C. It was necessary to adjust for differing levels of covariates due to the interaction with the treatments. Given a population with M binary covariates, there are 2^M distinct covariate groups. For illustrative purposes we consider a population with 2 covariates, for example, age (< 60 vs ≥ 60) and ISS stage: I/II vs III. Let $x_{m_1 m2}$ denote the proportion of the population in each covariate group. In this case m_1 is either a 1 to indicate that a patient is under 60, and 0 otherwise. Likewise, m_2 is either a 1 to indicate that a patient has ISS stage III, and 0 if a patient has ISS stage I or II. Thus, there are four distinct groups: x_{00} for patients ≥ 60 with ISS stage I or II, x_{10} for patients < 60 with ISS stage I or II, x_{01} for patients ≥ 60 with ISS stage III, and x_{11} for patients < 60 with ISS stage III. The true average efficacy of treatment k is estimated as:

$$\bar{d}_k = d_k x_{00} + (d_k + \beta_{1,t_k})x_{10} + (d_k + \beta_{2,t_k})x_{01} + (d_k + \beta_{1,t_k} + \beta_{2,t_k})x_{11}. \tag{5.1}$$

Then, the MAE for each treatment is given by:

$$\text{MAE}(\hat{d_k}) \equiv \frac{\sum_{q=1}^{Q} |\bar{d}_{k_q} - \hat{d}_{k_q}|}{Q},$$

where $\hat{d_k}$ is our model's estimate of the effect of treatment k relative to treatment A, and Q is the total number of simulations. Treatment A is treated as our reference treatment and hence set to 0 for the inference, therefore results are presented for treatments B and C only. We consider two populations of interest:

1. Full network population: x_{m_1m2} is computed across all studies in the network, weighted by sample size. Although this may not necessarily be the general population of patients in the target indication, we consider this to be the best possible estimate for the general population.

2. Target study population: x_{m_1m2} is computed using only the target study.

Uncertainty is measured as the posterior SD, as reported in the JAGS output. This is the same for both the full network population and the target population. For the MAIC it is obtained in the same way as the rest of the models, as the posterior SD in the JAGS output from the NMA using the reweighted trial data as the input. For the mixed AgD/IPD model, given that we have estimated a treatment effect and an effect for each covariate, we also use equation 5.1 to compute the estimate of the overall treatment effect for that population. Note that this cannot be done for the standard NMA models with the covariate, (Models 1b and 2b), as in this case we are assuming that the covariate is a prognostic variable. For the AgD/IPD mode the calculation is carried out in the JAGS model itself, in order to obtain the correct posterior SD and upper and lower CrI bounds from the trace.

5.2.4 NMA Models

The models used in this chapter are based on the models described in Section 2.8. We used the HR part of a model detailed in Woods et al. (2010) for the standard NMA HR model as well as the two HR models after the MAIC adjustment in the simulation study (Models 1a, 1c, and 1d):

$$\ln(H_{ik}) = \mu_i + d_k - d_b + \mathrm{re}_{ik} - \mathrm{re}_{ib},$$

where H_{ik} is the hazard ratio of treatment k versus the baseline treatment in study i, μ_i is the study effect of study i, d_k and d_b are the treatment effects for treatment k and the baseline treatment in each study, respectively, and $\mathrm{re}_{i,k}$ is the random effect deviation for arm k of study i. The prior distributions chosen are $\mu_i \sim N(0, 100^2)$, $d_k \sim N(0, 100^2)$, $\mathrm{re}_{ik} \sim N(0, \sigma^2)$, and $\sigma^2 \sim \mathrm{Unif}(0, 5)$. σ is the measure of between study heterogeneity in the HR model. All simulated and ndMM trials had exactly two arms so corrections for multiple arms did not need to be considered.

An extra model allowing for an extra term for each covariate was also considered (Model 1b):

$$\ln(H_{ik}) = \mu_i + d_k - d_b + \sum_{m=1}^{M} \beta_m (x_{m_b} - x_{m_k}) + \mathrm{re}_{ik} - \mathrm{re}_{ib},$$

where β_m is the effect of each covariate m, x_{m_b} is the proportion possessing the characteristic associated with the covariate m in the baseline arm, and x_{m_k} is the proportion possessing the characteristic associated with the covariate m in the treatment arm. As AgD models contain very little information about covariates, it was not possible to assign the vague priors, which were given the other parameters, to each β_m. Therefore the prior on each β_m follows $N(0, 3.16^2)$.

Although HR models are often preferred, sometimes limited information means that only medians are available. Therefore, we also consider models which use only medians and assume an exponential survival model. The first median model which we will present is the standard NMA median model, which includes NMAs that are implemented after a MAIC adjustment (Models 2a, 2c, and 2d). Let r_{ij}, the number of events in the j^{th} arm of the i^{th} trial. $r_{ij} \sim \text{binomial}(p_{ij}, n_{ij})$, where p_{ij} is the probability of an event in the j^{th} arm of the i^{th} trial and n_{ij} is the number of patients in the j^{th} arm of the i^{th} trial. As we are dealing with medians we obtain μ_i and δ_{ij} from:

$$S_{ij}(t^*) = \begin{cases} \exp(-t^* \exp(\mu_i)) & \text{if j=1} \\ \exp(-t^* \exp(\mu_i + \delta_{ij})) & \text{if j>1} \end{cases},$$

noting that for the median time p_{ij} is 50%. δ_{ij} is the effect of the treatment in the j^{th} arm of the i^{th} trial, μ_i is the baseline risk in study i and t^* is the median survival time. The treatment effects follow $\delta_{ij} \sim N(d_{t_{ij}} - d_{t_{i1}}, \sigma_\delta^2)$, where $d_{t_{ij}}$ denotes the effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment and σ_δ is the measure of between study heterogeneity in the HR model. The prior distributions chosen are $\mu_i \sim N(0, 1.83^2)$, $d_k \sim N(0, 1.83^2)$ and $\sigma_\delta \sim \text{Unif}(0, 2)$. The effect of treatment A (the reference treatment) is set to zero in the model with all other treatments being compared to treatment A.

An additional model which allows for covariate terms was also considered (Model 2b):

$$S_{ij}(t^*) = \begin{cases} \exp(-t^* \exp(\mu_i)) & \text{if j=1} \\ \exp(-t^* \exp(\mu_i + \delta_{ij} + \sum_{m=1}^{M} \beta_m x_{m_{ij}})) & \text{if j>1} \end{cases},$$

where $x_{m_{ij}}$ is the proportion of patients in the j^{th} arm of the i^{th} trial possessing the characteristic associated with the covariate, m. The prior on each β_m follows $N(0, 1.83^2)$. The parameters for μ, d, and β_m are chosen for the reasons discussed in Chapters 3 and 4.

We also considered a mixed AgD/IPD model to utilise IPD where possible and AgD otherwise. For the AgD part the model is:

$$S_{ij}(t^*) = \begin{cases} \exp(-t^* \exp(\mu_i)) & \text{if j=1} \\ \exp(-t^* \exp(\mu_i + \delta_{ij} + (\beta_{t_{ij}} - \beta_{t_{i1}})x_{ij})) & \text{if j>1} \end{cases},$$

and for the IPD part the rate λ is given by:

$$\log(\lambda) = \begin{cases} \mu_i + \beta_{0_i} x_{ijl} & \text{if j=1} \\ \mu_i + \beta_{0_i} x_{ijl} + \delta_{ij} + (\beta_{t_{ij}} - \beta_{t_{i1}})x_{ijl} & \text{if j>1} \end{cases}.$$

For the IPD dataset we can also include the trial-specific covariate effect β_{0_i}. In this case, $\beta_{t_{ij}}$ is the covariate effect for the treatment in the j^{th} arm of the i^{th} trial, i.e., the covariate is an effect modifier as it interacts with each treatment differently. This covariate effect can either be modelled as independent, (i.e., each $\beta_k \sim N(0, 1.83^2)$, where k represents the treatment), or as exchangeable, (i.e., the distribution for each β is $\beta_k \sim N(\mu_\beta, \sigma_\beta^2)$, with $\mu_\beta \sim N(0, 1.83^2)$ and $\sigma_\beta \sim \text{Unif}(0, 5)$). β_1 is set to zero, with all other covariate effects estimated relative to the covatiate effect of treatment 1, in order to aid convergence. This varies in parameterisation from Chapter 4, where the prior on β_1 was also from the same vague distribution. While the choice of parameterisation does not affect the absolute effect estimates, this method will improve convergence.

We also consider a model which assumes that each covariate is a prognostic variable:

$$S_{ij}(t^*) = \begin{cases} \exp(-t^* \exp(\mu_i)) & \text{if j=1} \\ \exp(-t^* \exp(\mu_i + \delta_{ij} + \beta x_{ij})) & \text{if j>1} \end{cases},$$

and for the IPD part the rate λ is given by:

$$\log(\lambda) = \begin{cases} \mu_i + \beta_{0_i} x_{ijl} & \text{if j=1} \\ \mu_i + \beta_0 x_{ijl} + \delta_{ij} + \beta x_{ijl} & \text{if j>1} \end{cases}.$$

For the ndMM example we compare the mixed AgD/IPD model that makes the assumption of effect modifier to the mixed AgD/IPD model that makes the assumption of prognostic variable for each covariate to identify the nature of each covariate. We also extend the mixed AgD/IPD model to using two or three covariates where applicable in both the simulation study and the ndMM example.

In this case we assume an independent treatment effect for each covariate.

All above models can be simplified to an FE model, by removing the variance on δ in the median models, and removing the re term in the HR models. When identifying the nature of the covariates using the mixed AgD/IPD models we used the FE models in order to reduce variance and detect true interactions. Given that the goal of MAIC is to reduce differences between studies, FE models could also be considered appropriate. However, while MAIC may reduce some heterogeneity there is no guarantee that it has removed all heterogeneity. For example, researchers may not have collected all relevant covariates when carrying out the study. In the ndMM example we know that we have not matched on cytogenetics and we have lost some information due to the fact that age was dichotomised. Therefore, we have assessed the suitability of both FE models and RE models.

5.3 Results

5.3.1 Simulation Study

In this section we compare the models in terms of MAE in both the full network population (i.e. all studies in the network) and the target population (i.e. just the B vs C studies in this case). The coverage probabilities associated with these estimates are also presented. We also consider the posterior SD, which applies to both populations. The dotted lines around each MAE and posterior SD estimate represent the MC Error on each side. Finally, we provide the average measure of heterogeneity for each data point.

As the computation time required for the mixed AgD/IPD model is substantially longer than the computation time for all other models, we present the results of this model for scenario 2 (varying proportion possessing the characteristic associated with each covariate in the BC-AgD trial) and scenario 3 (varying proportion possessing the characteristic associated with each covariate in the AB-IPD trials) only. However, in a preliminary analysis we examined the effect of the covariate-treatment interaction on this model, and we found it behaved similarly to the other models.

Covariate-Treatment Interaction

Figures 5.3 and 5.4 both show how an increase in the SD of the covariate-treatment interaction affects the MAE and posterior SD. The MAE increases for all models as the covariate-treatment interaction increases, but the HR model is particularly

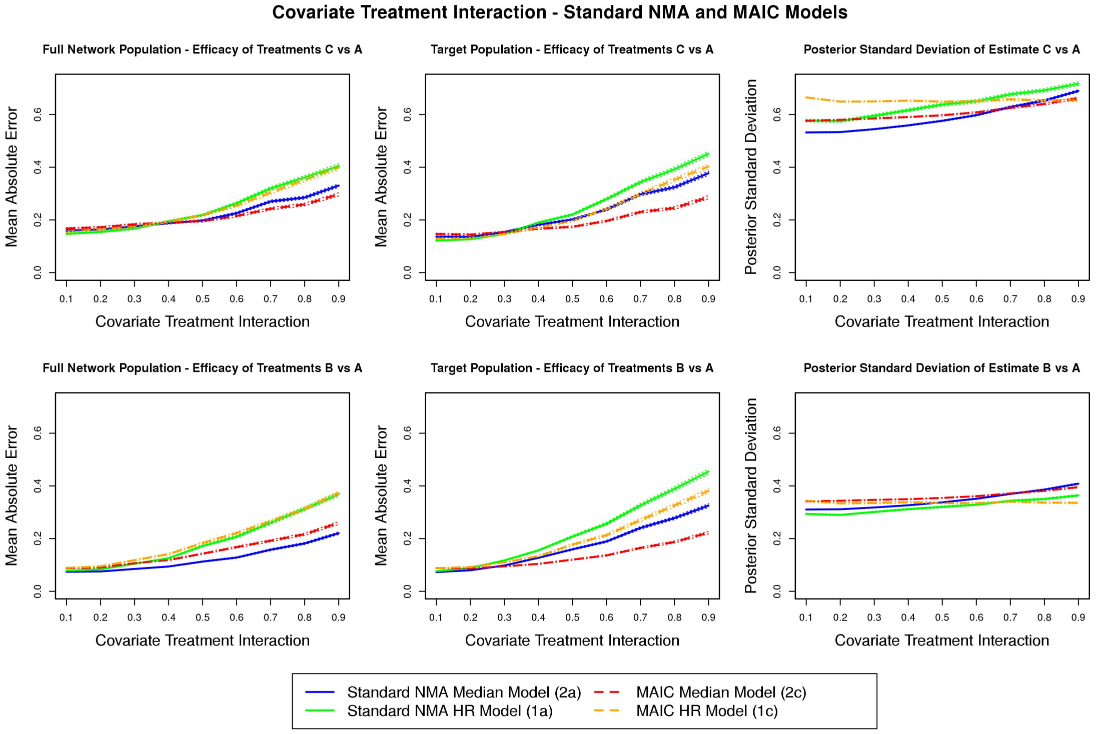


Figure 5.3: Examining mean absolute error (MAE) and posterior standard deviation (SD) while increasing the SD of the covariate-treatment interaction. The dotted lines around each MAE estimate represent the MC Error on each side. At the left most point of the x-axis the covariate is a prognostic variable and at all other points the covariate is an effect modifier. When the interaction is zero or relatively small there is little difference between models. HR models are worse than median models when the interaction is large. MAIC is better than a standard NMA when looking at the direct B versus A comparison in the target population, especially for large interactions. For the HR model in the full network population there is little difference between the standard NMA model and the MAIC model.

impacted. The posterior SD increases as the covariate-treatment interaction increases for the NMA models, but stays relatively constant for the MAIC models.

The top row of Figure 5.3 shows the impact on the indirect estimate of C versus A. We see that the MAIC adjustment gives better results than a standard NMA when the interaction is large, especially for the target population. At the left-hand side of the x-axis, when the interaction is zero, the covariate is a prognostic variable rather than an effect modifier. In this case all models are quite similar. However, we should bear in mind that it can be difficult to distinguish between prognostic variables and effect modifiers, and note that, in the case of an anchored comparison, there is a larger benefit to reweighting for an effect modifier than there is for not adjusting for a prognostic variable.

In the bottom row we examine the impact of the covariate-treatment interaction on the direct estimate of B versus A. Although it is not necessary to do an MAIC

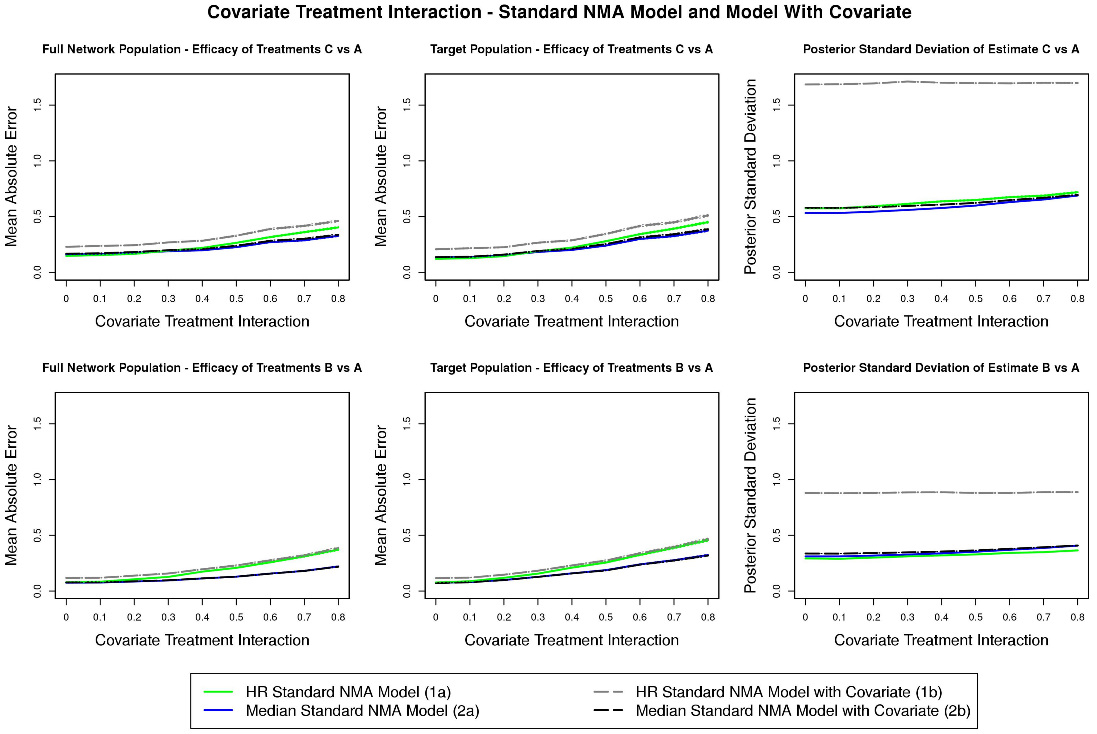


Figure 5.4: Examining mean absolute error (MAE) and posterior standard deviation (SD) while increasing the SD of the covariate-treatment interaction. The dotted lines around each MAE estimate represent the MC Error on each side. At the left most point of the x-axis the covariate is a prognostic variable and at all other points the covariate is an effect modifier. The HR standard NMA model with covariate produces a much larger posterior SD than the other models. This model also has the highest MAE for the C vs A estimate and the joint highest (with HR standard NMA model) MAE for the B vs A estimate.

for this estimate, it is important to understand how this estimate is affected by MAIC, as NMAs often report the relative efficacy of all treatments versus the reference treatment. For both the HR and the median models, MAIC produces a lower MAE than the standard NMA model in the target population. However, MAIC does not have a large impact on the MAE of the HR model in the full network population as both models give very similar MAEs. MAIC negatively impacts the MAE for the median model for the full network population. This highlights that the RCT evidence is of better quality than the MAIC evidence when we are interested in the direct comparison.

Finally, the graphs on the right hand side show the posterior SD. We see a small increase in posterior SD due to an increase in the covariate-treatment interaction for the standard NMA models, but the increase is not as large as the increase for the MAE. The posterior SD of the MAIC models are generally unaffected by an increase in the covariate-treatment interaction. For both the median and the

HR models the MAIC produces a larger posterior SD than the standard NMA when the covariate-treatment interaction is small, but as the covariate-treatment interaction increases, the posterior SDs of the standard NMA model increase to the point where they are larger than the posterior SD for the corresponding MAIC model.

Figure 5.4 shows the same simulation scenario, but this time the standard NMA Model with Covariate (1b and 2b) is compared to the standard NMA model. We explore this model as it may be a feasible alternative to MAIC, if a researcher does not have access to IPD. For the HR model (1b), this is clearly inferior as it produces much larger posterior SDs and slightly larger MAEs for the indirect C versus A estimate. We note that Figure 5.3 and Figure 5.4 have different scales due to these higher posterior SDs. Here we are including three extra terms in a model with only four studies, and hence we do not have enough data-points to include the extra terms. We should note that this model is already using a much less vague prior for the covariate effect, in order to reduce the uncertainty in the posterior SD, than it is for any other parameter. For the median model (2b), we get very similar results with and without the covariates. However, as this is a more complex model, there does not appear to be any benefit to including the extra terms given the limited data in this scenario. We therefore do not explore this model in further scenarios, due to a preference for model simplicity.

These simulations use the same prevalence of the covariates that are in the ndMM network, so it can plausibly give us an indication that MAIC may be of benefit to estimating the Len-Thal comparison, especially for the median model. However, we will now analyse other scenarios to ascertain the benefit of MAIC over more generic networks.

Varying Proportion of Patients Possessing Each Covariate

The remaining figures in this section all assume that the standard deviation of the covariate-treatment interaction is 0.8, which is the largest that we looked at in the previous graphs. Figures 5.5 and 5.6 show that the proportion of patients possessing the characteristics associated with the covariates in each trial greatly influences the effect of MAIC. In both figures we see that for the target population (centre graphs), MAIC produces an MAE which is lower than or the same as the corresponding standard NMA model. However, using a mixed AgD/IPD model also produces MAEs as low as the MAIC model, and when the proportion of patients possessing the characteristic associated with the covariate in the IPD trials is particularly low or high (Figure 5.6, middle column, edge of graphs), the mixed AgD/IPD model produces a lower MAE than the MAIC models. For the

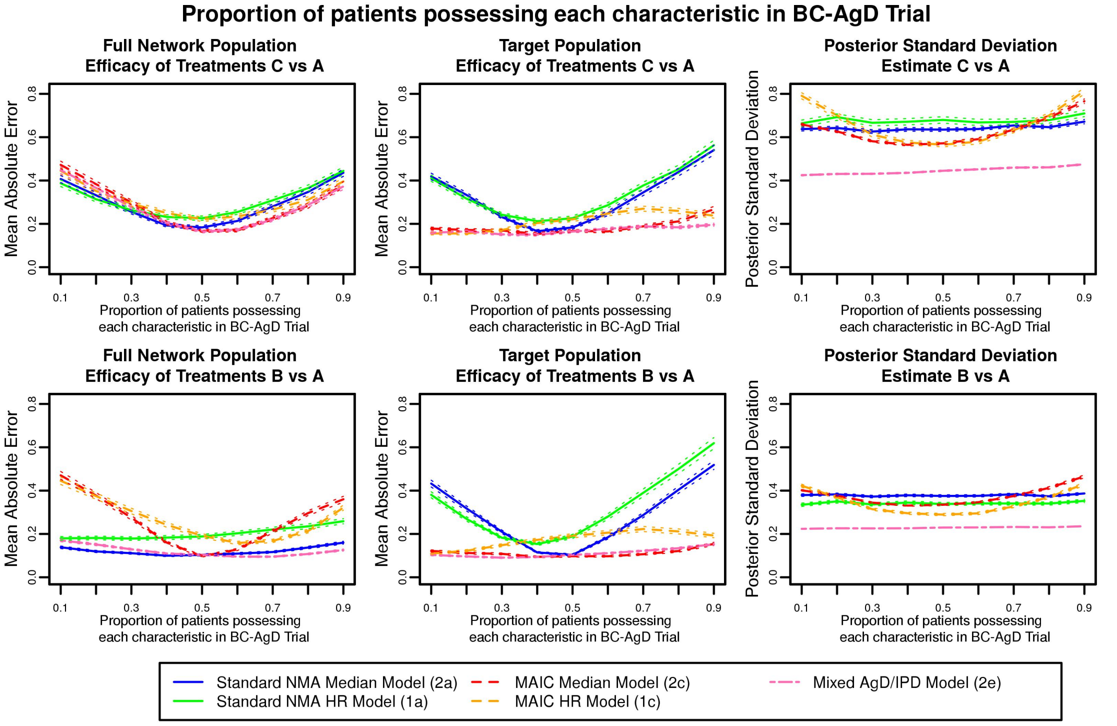


Figure 5.5: Examining mean absolute error (MAE) and posterior standard deviation (SD) while varying the proportion of patients possessing each characteristic in the BC-AgD study. The dotted lines around each MAE estimate represent the MC Error on each side. The proportion of patients possessing each characteristic in the IPD study is 40%, 50% and 60% for the three IPD-AB studies, respectively. There is little difference between the models in the full network population for the indirect C versus A estimate, while both the mixed AgD/IPD model and the standard NMA model produce lower MAEs than the MAIC model in the full network population for the direct B versus A estimate. For the target population the mixed AgD/IPD model and the MAIC median model produce the lowest MAE.

target population, when the distribution of covariates in the AB-IPD trials is very different to the BC-AgD trial (i.e. the extremes of the graphs) the standard NMA models give worse estimates than they do when the covariate make-up is similar between the studies. The median MAIC model and the mixed AgD/IPD models, however, are not as affected by a difference in covariates, as the MAE is quite flat, with the exception of the extreme points in Figures 5.5 5.6. The MAIC HR model also has a flat line for the target population in Figure 5.6, but has a slight concave downwards slope for the target population in Figure 5.5.

Looking at the effect of MAIC on the full network population (left hand side graphs) and on the posterior SD (right hand side graphs) we can see that there are downsides to running an MAIC. In Figure 5.6 the MAIC gives a posterior SD that is at least as big, if not greater than, the standard NMA model for both the direct AB estimate and the indirect AC estimate. This means we are increasing

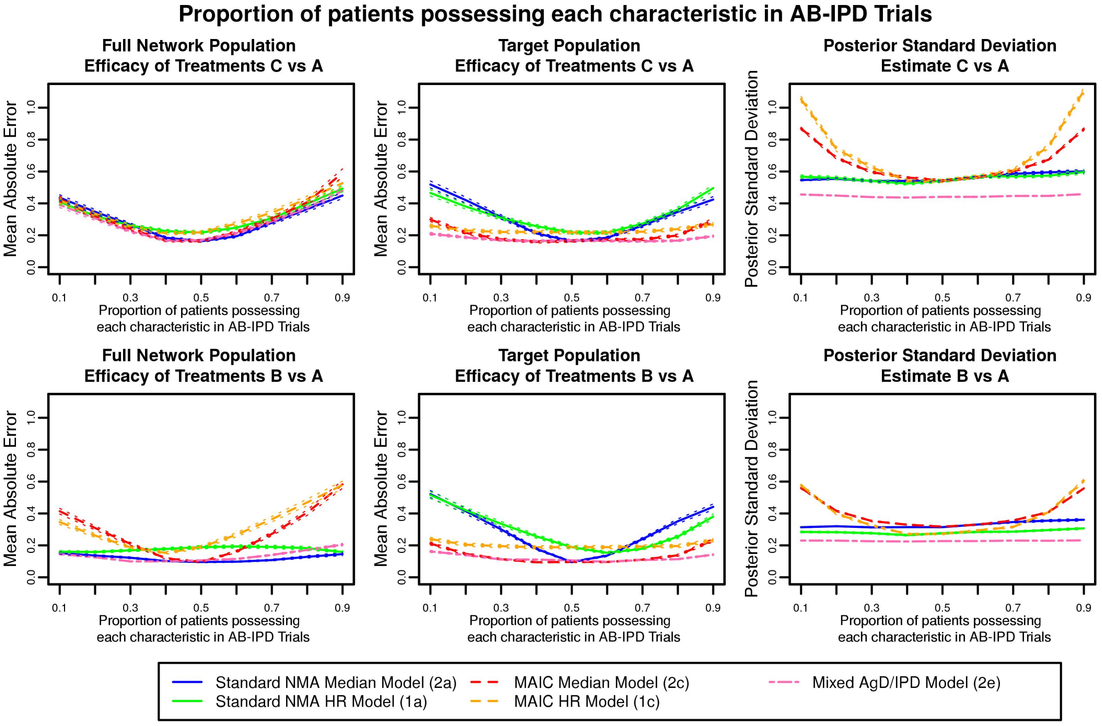


Figure 5.6: Examining mean absolute error (MAE) and posterior standard deviation (SD) while varying the proportion of patients possessing each characteristic in the AB-IPD study. The dotted lines around each MAE estimate represent the MC Error on each side. The proportion of patients possessing each characteristic in the BC-AgD study is 50%. There is little difference between the models in the full network population for the indirect C versus A estimate, while both the mixed AgD/IPD model and the standard NMA model produce lower MAEs than the MAIC model in the full network population for the direct B versus A estimate. For the target population both the mixed AgD/IPD model and the MAIC median model produce lower MAEs than the standard NMA model, with the mixed AgD/IPD model performing particularly well at the edge of the graph.

our uncertainty in our estimate by running an MAIC. In particular, we can see that the C versus A posterior SD is much larger than the corresponding MAE for the MAIC. We also see that the posterior SD increases when more reweighting is required in Figure 5.5. The MAE of the C vs A estimate for the full network population is quite similar for the standard NMA model and the MAIC model. However, the direct B vs A estimate from the MAIC in the full network population is usually worse, or at best similar to the the standard NMA, hence we do not recommend using MAIC for this estimate. The mixed AgD/IPD model shows no benefit in terms of MAE over the other models in terms of MAE in the full network population for the indirect C versus A estimate, and produces similar MAEs to the standard NMA model for the direct B versus A estimate. However, we also see that there is decreased uncertainty with this model.

There is a noticeable lack of symmetry in some models in Figures 5.5 and 5.6. This is due to the fact that the model is computed on the log scale, and therefore the distances between points on the axis are not equal on this scale.

As mentioned earlier, some of the HR models show a concave downwards slope. For example direct B versus A estimate in the target population graph in Figure 5.5. In Figure 5.5 the MAIC reweights the data such that the proportion possessing each characteristic is equal to the number on the x-axis. The patients in the trial are most heterogeneous when 50% of the patients in the trial possess each characteristic. This is approximately where the HR model produces the worst MAE (slight difference due to some calculations being on the log scale). This is consistent with Figure 2 of Shrier & Pang (2015), where the estimated odds ratio is impacted by the value of the covariate.

Figures 5.5 and 5.6 both assume that for any given point on the x-axis there is the same proportion of patients who possess the characteristic associated with each covariate in each trial, so it is not possible to differentiate between the two methods of reweighting the IPD. We therefore consider an additional scenario in where there is a difference between the proportion possessing the characteristic associated with each covariate in each IPD trial. Both the HR models and the median models give similar results. For clarity in the graphs we have kept the two models separate. As the median model has shown more promising results in general, we will present the results for the median model in the main text and the HR model in the appendix. Given that the purpose of this scenario is to compare the two MAIC methods, we exclude the AgD/IPD model due to the large computing power required for this model.

Figure 5.7 keeps the overall proportion of patients who possess each characteristic constant at 45%, while varying the IPD proportion between the three studies. For example, on the left most point of the x-axis all three AB-IPD studies have 45% of patients who possess each characteristic, however, on the right hand side Study 1 has 45% of patients who possess each characteristic, Study 2 has 90% of patients who possess each characteristic, and Study 3 has 10% of patients who possess each characteristic. We can see overall that when there is an increase in heterogeneity between the studies the posterior SD increases, as does the MAE in nearly all cases, barring the direct estimate of B versus A in the full network population. When the studies are similar there is not much of a difference between the two reweighting methods. However, the MAIC pooled trials method is a less stringent requirement, since the AB-IPD studies need only match the BC-AgD study between the three of them, whereas for MAIC separate trials each AB-IPD study needs to match the BC-AgD study. Therefore, less reweighting is carried

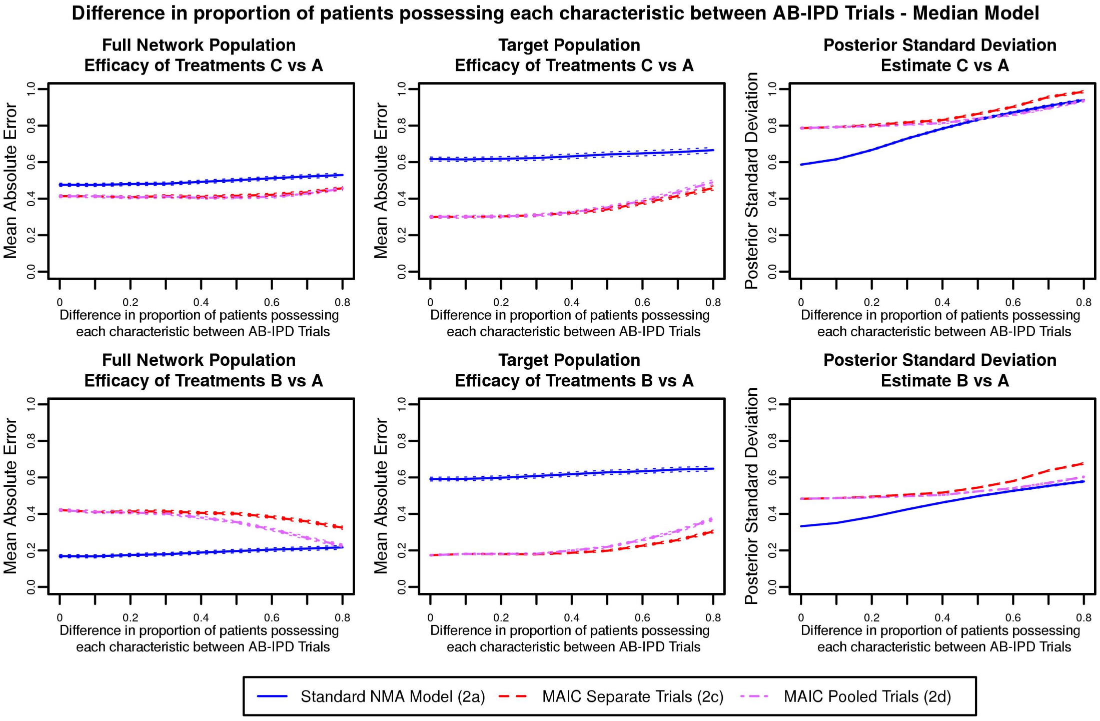


Figure 5.7: Examining mean absolute error (MAE) and posterior standard deviation (SD) while varying the difference in proportion possessing the characteristic associated with each covariate between AB-IPD trials. The dotted lines around each MAE estimate represent the MC Error on each side. On the left most point of the x-axis all three AB-IPD studies have 45% of patients possessing the characteristic associated with each covariate, however, on the right hand side Study 1 has 45% of patients possessing the characteristic associated with each covariate, Study 2 has 90% of patients possessing the characteristic associated with each covariate, and Study 3 has 10% of patients possessing the characteristic associated with each covariate. The numbers on the x-axis represent the difference in the proportion possessing the characteristic associated with each covariate between Study 2 and Study 3. The AB-AgD study has a fixed proportion possessing the characteristic associated with each covariate of 90%. MAIC pooled trials is a less stringent requirement and so less reweighting needs to be done with this method.

out for the pooled trials method and the posterior SD is slightly smaller for the pooled trials method. In fact, the posterior SD is more similar to the posterior SD of the standard NMA than it is to the separate trials MAIC. For the target population, for both the direct B versus A comparison and the indirect C versus A comparison, the separate trials method produces a slightly lower MAE than the pooled trials model, as the covariate make-up of the IPD trials become more different from each other. However, as there is a greater amount of interference in the AB-IPD studies, the direct B versus A estimate are worse for the separate trials MAIC than the pooled trials MAIC in the full network population. In fact, when one of the AB-IPD trials has the same covariate make-up as the BC-AgD trial, the MAIC pooled trials method is approximately as good as the standard NMA model in terms of the MAE for the BA estimate in the full network population.

Coverage Probabilities and Fixed Effects Models

We compared the coverage for both RE and FE models in Figures 5.8, 5.9, 5.10, and 5.11. In general, the RE models had coverage which was closer to the nominal 95% CrI, and this is why we choose to present the RE model for the main results. However, there were some exceptions to this. For example, in Figure 5.9, for the target population, when a large amount of patients possess the characteristic associated with each covariate, the median MAIC FE model has coverage which is closer to the nominal CrI than the corresponding RE models. Therefore, we have included the results of FE models in the appendix. FE models gave similar results to the RE models for the MAE. The posterior SD, however, was much smaller for the FE models compared to the RE models.

The standard NMA HR model with the covariate (Model 1b) generally had coverage which was too high (close to 100%), this was also the case for the RE mixed AgD/IPD model (Model 2e). For the rest of the models however, the coverage was generally below the nominal 95%, and in many cases had quite low coverage. In the target population we note that the MAIC models often had coverage which was closer to the nominal 95% than the corresponding standard NMA model. This is especially true for the FE models, which implies that MAIC has reduced the differences between trials. However, we would stress that in a real-world scenario, there may be more effect modifiers which have not been accounted for.

The coverage probability is driven by a combination of the MAE and the posterior SD. For example, the coverage probabilities for the standard NMA FE models in Figure 5.9 are almost the inverse of the MAE in Figure C.4. In this case the MAE increases at the edges of the graphs, but the posterior SD stays the same.

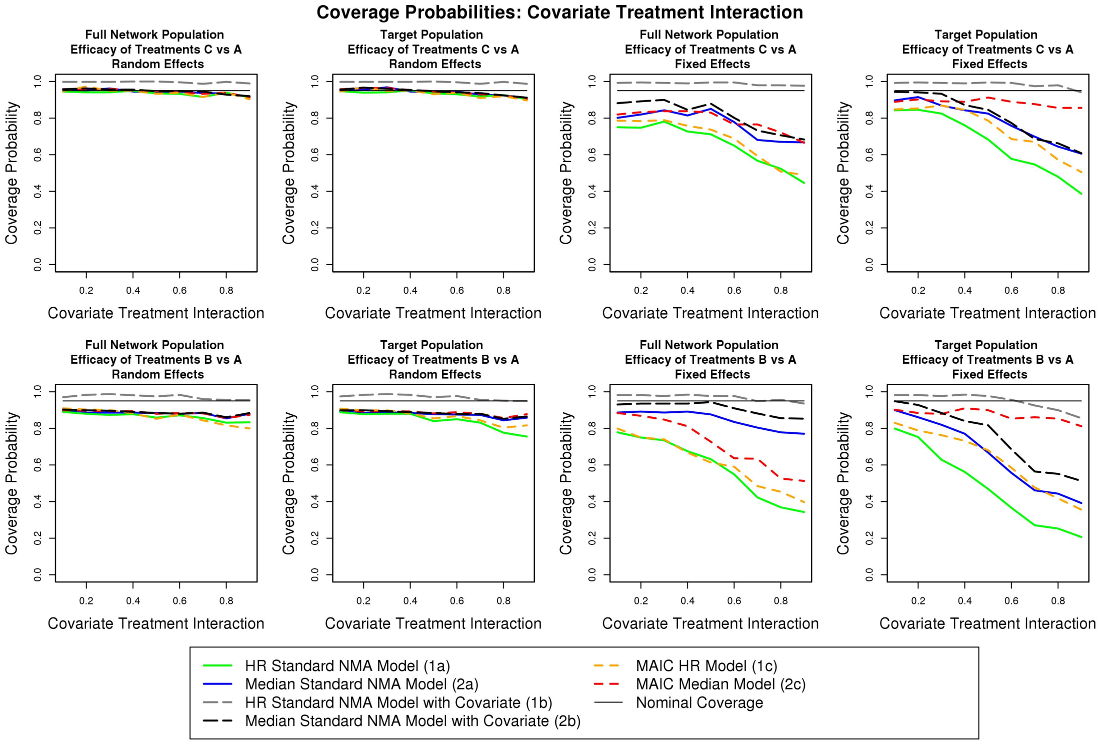


Figure 5.8: Examining coverage probability while increasing the standard deviation of the covariate-treatment interaction. The RE models are generally unaffected by an increase in the covariate-treatment interaction. Although the coverage of standard NMA models and the MAIC models are lower than the 95% nominal probability, in particular for the B versus A comparison. For the FE model, however, the coverage becomes much lower than the 95% nominal probability as the covariate-treatment interaction increases.

In other cases, for example in Figure 5.11, we see the coverage of the standard standard RE NMA models increase as the difference between the IPD studies, in term of the proportion of patients possessing the characteristic associated with the covariate, increases. This is almost entirely due to an increase in the posterior SD in Figures 5.7 and Figures C.1.

Effect on the Between Study Heterogeneity

Figure 5.12 shows how the between study heterogeneity is affected in the various scenarios explored previously. In many cases the models produce quite similar between study heterogeneity. On the top left panel we can see that the standard NMA HR model with the covariate (Model 1b) produces larger between study heterogeneity than the other models. This model often behaves differently from the other models, for example, the posterior SD of this models was also higher than the other models in Figure 5.4. In the bottom left panel we see that when the IPD trials have a relatively large or small number of patients possessing the

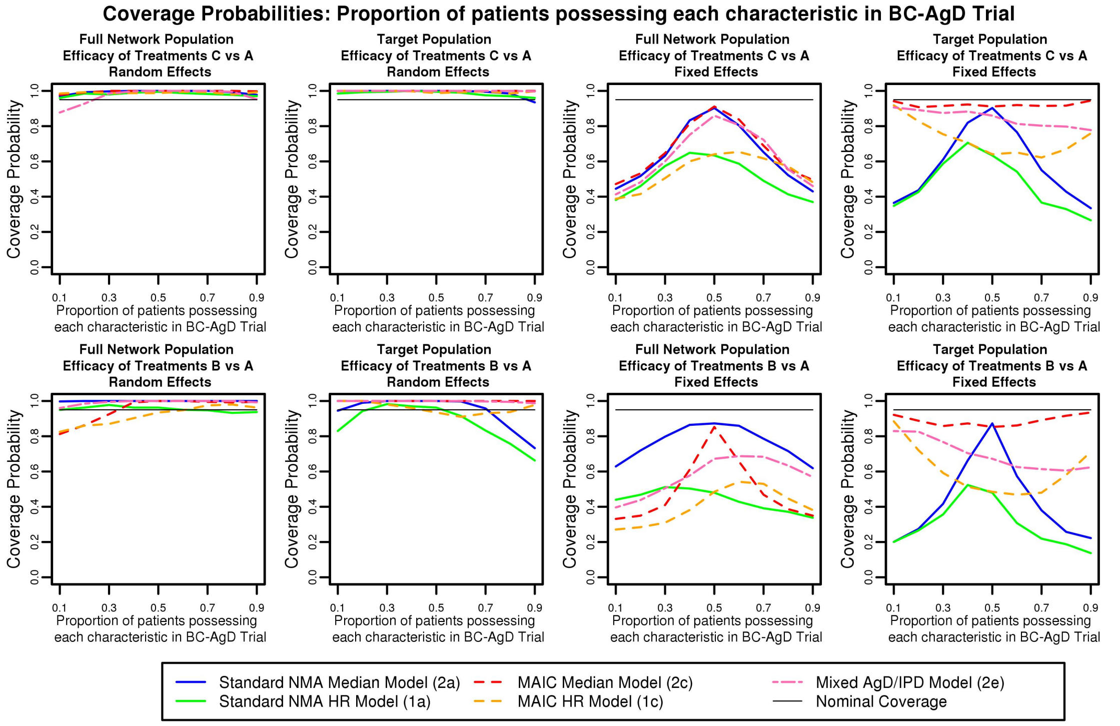


Figure 5.9: Examining coverage probability while varying the proportion of patients possessing the characteristic associated with each covariate in the BC-AgD study. The FE models are more sensitive to differing values of patients possessing the characteristic associated with each covariate than the RE models. For the FE models, in the target population the MAIC models generally have coverage closer to the nominal 95% than the standard NMA models, while in the full network population, for the B versus A estimate, the standard NMA models have coverage closer to the nominal 95% probability than the MAIC models.

characteristic associated with each covariate (i.e. at the edges of the graph), the MAIC model produces a larger measure of heterogeneity then the standard NMA models, most likely due to the large amount of reweighting required. In the bottom right panel we see a clear increasing trend as the difference in proportion of patients possessing the characteristic associated with each covariate increases between AB-IPD trials. Hence, as the difference in each patient population increases, more between trial heterogeneity is estimated, as the model does not explicitly take the effect modifiers into account. The standard NMA models are more affected by this difference. When the three IPD trials are similar the MAIC models again produce a larger estimate of between study heterogeneity than the standard NMA models. There is no noticeable difference between MAIC pooled trials and MAIC separate trials model.

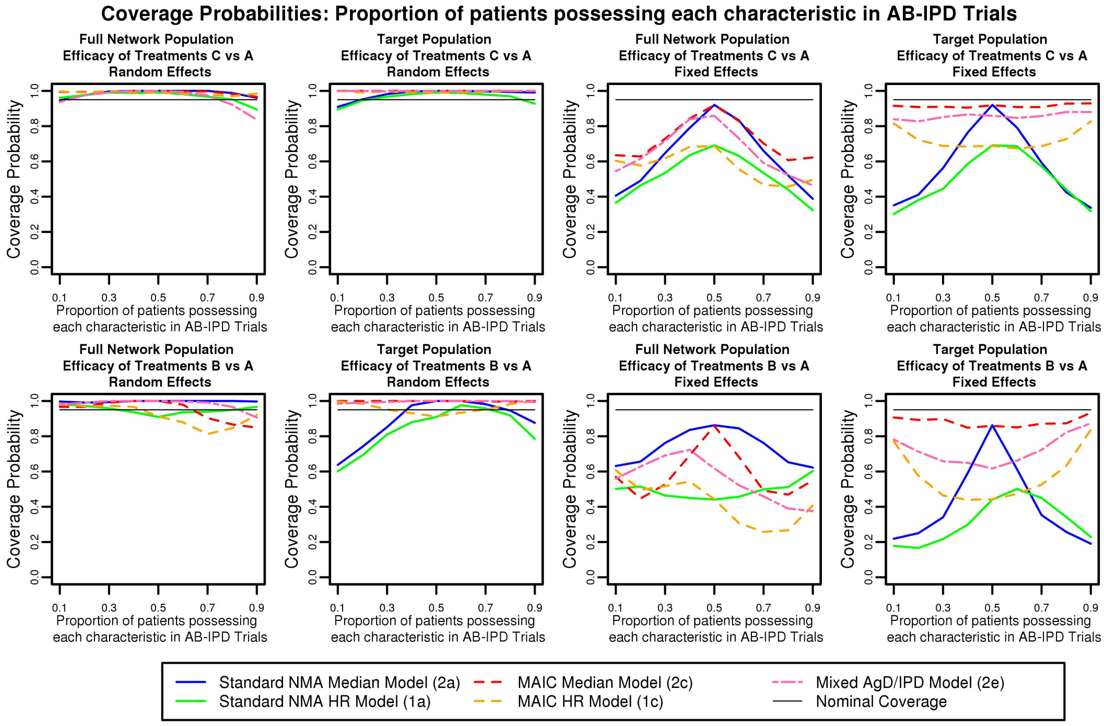


Figure 5.10: Examining coverage probability while varying the proportion of patients possessing the characteristic associated with each covariate in the AB-IPD study. The FE models are more sensitive to differing values of patients possessing the characteristic associated with each covariate than the RE models.

5.3.2 Applied Example in Newly Diagnosed Multiple Myeloma (ndMM)

We first investigate which covariates are appropriate to be used in the ndMM network. Table 5.3 shows the potential bias resulting from the differences between trials, assuming that there is an independent covariate-treatment interaction. A second scenario was considered assuming an exchangeable covariate-treatment interaction with similar results and we therefore just show the independent model. In both cases no covariate could be excluded as not interacting with the treatments by looking at the differences in DIC. We also found the potential bias in most cases to be relatively small. The only exception to this was Response which had the highest potential bias of 0.10. However, based on the DIC, Response was most likely to be a prognostic variable, although the differences in the DIC was below the threshold of 3. Clinical advice recommended that gender was unlikely to interact with treatment, so we excluded this as a covariate for matching, given that it also had one of the smallest potential biases. We considered using the other four covariates for matching. However, given that we did not have any information on cytogenetics in the McCarthy trial, we would not have been able to include this

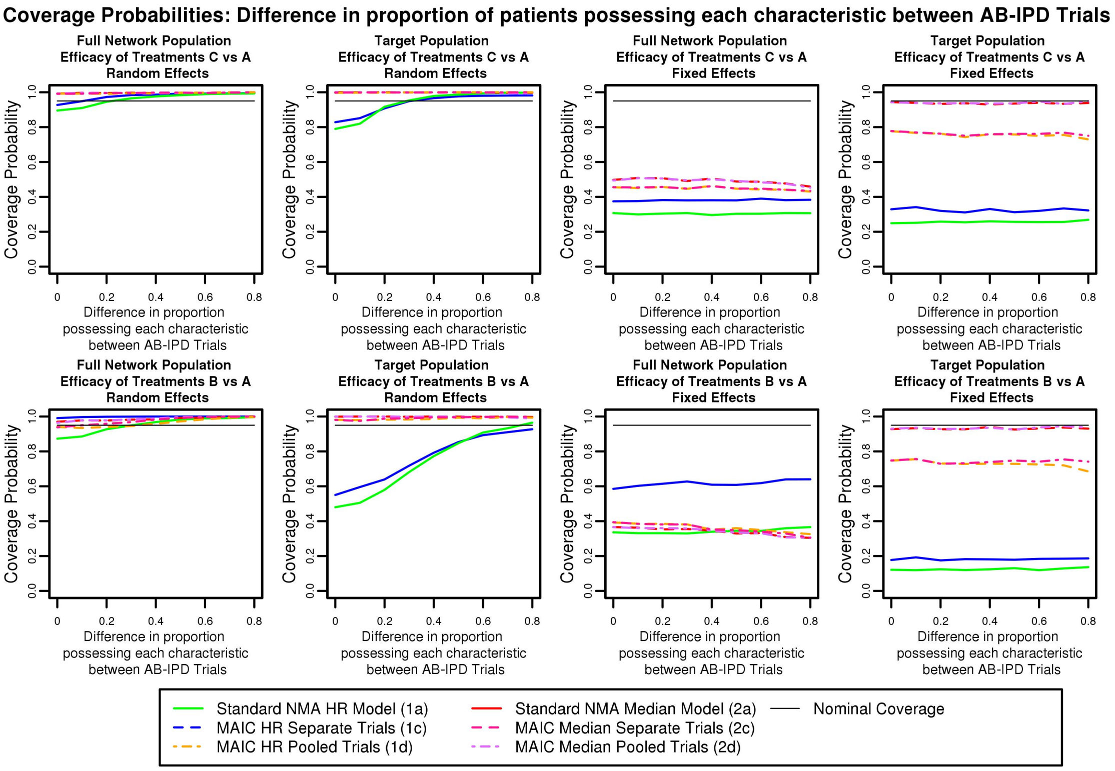


Figure 5.11: Examining coverage probability while varying the difference in proportion possessing the characteristic associated with each covariate between AB-IPD trials. In general RE models have coverage closer to the nominal 95% probability than FE models. There is no noticeable difference between the MAIC separate trials model and the MAIC pooled trials model.

Table 5.3: Potential bias if each covariate is an effect modifier. This is based on a fixed effects model using medians, using IPD where possible and assuming an independent covariate-treatment interaction. The potential bias is calculated by multiplying the absolute difference in interaction for Len versus Thal and the absolute difference between the proportion of patients possessing the characteristic associated with each covariate in Morgan trial versus the average across the Len-Placebo IPD trials.

		Age: <60	**ISS Stage: III**	**Response: CR/VGPR**	**Cytogenetics: Present**	**Gender: Male**
Absolute Difference in Interaction for Len versus Thal if Effect Modifier		0.24	0.26	0.31	0.01	0.31
Proportion possessing characteristic	**McCarthy (Len-Placebo)**	0.58	0.24	0.66	NA	0.54
	Attal (Len-Placebo)	0.64	0.20	0.59	0.13	0.57
	Palumbo (Len-Placebo)	0.69	0.13	0.39	0.36	0.54
	Average Len-Placebo	0.63	0.19	0.55	0.25	0.55
	Morgan (Thal-Placebo)	0.58	0.34	0.74	0.42	0.64
Potential Bias if Effect Modifier		0.01	0.04	0.06	0.002	0.03
DIC Difference (Positive favours Effect Modifier)		1	0	-2	0	1

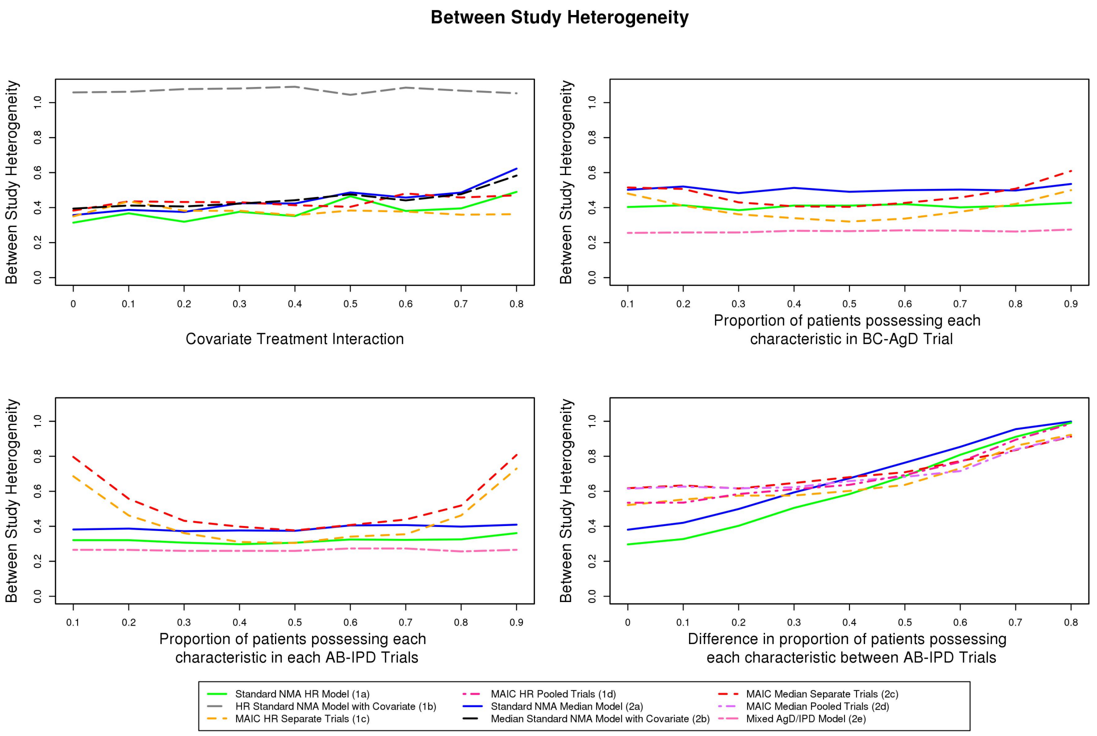


Figure 5.12: Effect on the between study heterogeneity. In the bottom left panel we see that when the IPD trials have a relatively large or small number of patients possessing the characteristic associated with each covariate (i.e. at the edges of the graph), the MAIC model produces a larger measure of heterogeneity then the standard NMA models, most likely due to the large amount of reweighting required. In the bottom right panel we see a clear increasing trend as the difference in proportion of patients possessing the characteristic associated each covariate increases between AB-IPD trials. When the three IPD trials are similar the MAIC model again produces a larger estimate of between study heterogeneity than the standard NMA models.

in a pooled trials model. Hence, to allow a equal comparison between pooled trials MAIC and separate trials MAIC we excluded cytogenetics as a covariate to be used in the MAIC. For a clinical analysis we could of course include cytogenetics, but as our goal was to analyse the differences between models we found it to be more appropriate to exclude it, given that it also had the smallest potential biases. In the case of the models that explicitly include a covariate (Models 1b, 2b, 2e), we included only the covariates for response and ISS stage, as we had no age information for the Jackson trial.

We considered both FE and RE models. However, when comparing the DIC between FE and RE for each of the nine models shown in Table 5.4, we found no differences greater than three, and therefore could not conclude that one model was more appropriate than the other. We therefore present the results of the RE model, as these, in general, had coverage which was closer to the nominal 95% CrI

Table 5.4: Estimate of between study heterogeneity for the random effects (RE) model, and the difference in DIC for the Fixed Effects (FE) model minus the RE model. Given that the difference in DIC is small, we conclude that there is not enough evidence to conclude that either the RE or FE model fits better than the other.

	Between Study Heterogeneity	**Difference in DIC (RE minus FE)**
HR Standard NMA (1a)	0.15	1.58
HR MAIC With Cov (1b)	0.41	0.63
HR MAIC Separate (1c)	0.22	1.02
HR MAIC Pooled (1d)	0.22	1.19
Median Standard NMA (2a)	0.15	1.78
Median MAIC With Cov (2b)	0.18	2.06
Median MAIC Separate (2c)	0.17	1.86
Median MAIC Pooled (2d)	0.17	1.83
Mixed AgD/IPD (2e)	0.31	-2.00

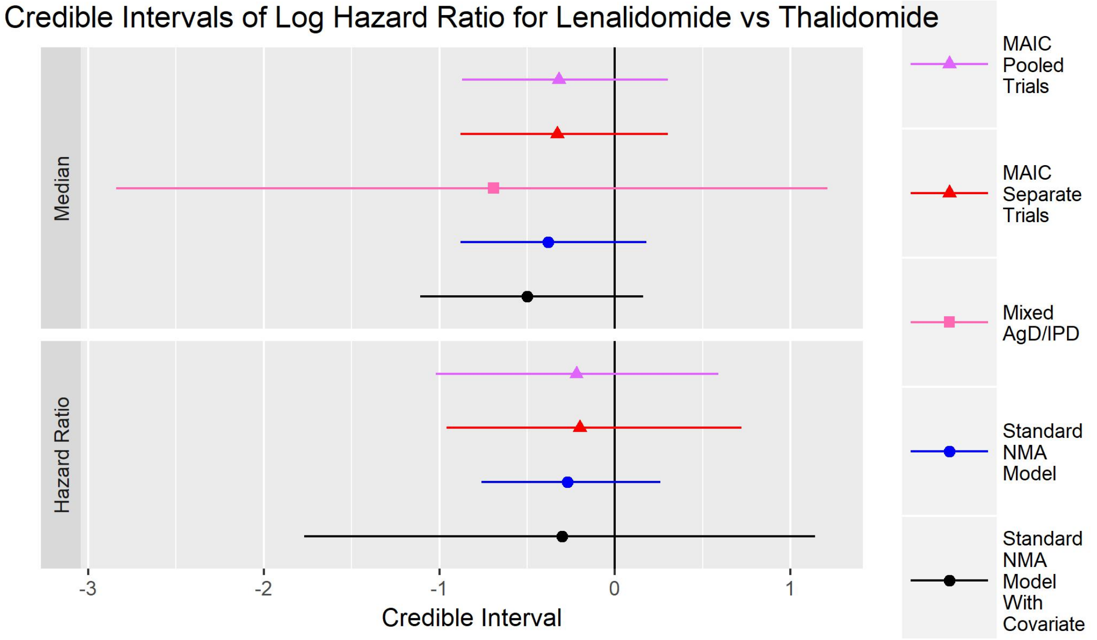


Figure 5.13: Credible intervals (CrI) for the log hazard ratios for ndMM network. The point estimates for lenalidomide are superior to thalidomide in all cases, however, the CrIs all cross the line of no effect at zero. MAIC does not have a large impact on the point estimate but slightly increases the width of the credible intervals.

in Section 5.3.1.

Figure 5.13 shows the results for the ndMM network. The point estimates in all cases indicate that lenalidomide is superior to thalidomide. However, the Credible Intervals (CrI) span zero. The relative size of the CrIs between models replicate what we saw in the simulation study. The mixed AgD/IPD has the widest CrI. The main reason for this is because we have no IPD available for a Thalidomide study. The standard NMA HR model with covariate also has quite a wide CrI. The CrI of the median model with covariate is slightly wider than the standard NMA median model also, but the difference between these two median models is not as big as the difference between the corresponding HR models. This again is similar to what has been observed in the simulation study. The MAIC models have larger CrIs than the corresponding standard NMA model for both the HR and median models. We also obtained CrIs for lenalidomide relative to placebo. In the case of MAIC relative to the unadjusted AgD estimate we found that the MAIC estimates were slightly shrunken towards zero, while once again becoming slightly more uncertain.

We can also look at the between study heterogeneity for the RE models, as shown in Table 5.4. For both the median and the HR models the between study heterogeneity is lowest for the standard NMA model, followed by the MAIC models, and is highest for the models which explicitly include the covariate. This may indicate that there are further unidentified covariates, which may be contributing to the between study heterogeneity.

5.4 Discussion

With MAIC generating more interest across the research and pharmaceutical community, it is important to evaluate the benefits and drawbacks compared to other models. While we saw some benefits to using MAIC over the standard NMA model, these same benefits can be obtained by using a mixed AgD/IPD NMA, which doesn't involve any post-hoc reweighting of data, and hence doesn't negatively impact the MAE in the full network population. Our results and recommendations are summarised in Table 5.5.

We also recommend that further analysis is carried out after conducting an MAIC, such as an indication of model fit and the effective sample size of the reweighted data. A further point to note is that the MAIC models generally produces posterior SDs that are larger than those produced from a standard NMA. However, in this simulation study we see that MAIC models generally have coverage below the 95% nominal CrI, and can be lower than the corresponding standard

Table 5.5: Summary of results and recommendations from simulation study

Comparison	Population of Interest	Result	Recommendation
C vs A	Full network	MAE is similar for all models.	Don't use MAIC.
C vs A	Target	Mixed AgD/IPD model and MAIC produce lowest MAE.	Use mixed AgD/IPD model as main analysis. Use MAIC only as a sensitivity analysis when interested in target population.
B vs A	Full network	Standard NMA and mixed AgD/IPD model produce lowest MAE.	Don't use MAIC.
B vs A	Target	AgD/IPD model and MAIC produce lowest MAE.	Use mixed AgD/IPD model as main analysis. Use MAIC only as a sensitivity analysis when interested in target population.

NMA model, which indicates that the increased posterior SD is not sufficiently accounting for the increased MAE. As mentioned in Chapter 4, Jansen (2012) also proposed IPD models for population adjustment, and found that the use of IPD does increase precision and reduce bias. Therefore, these models should also be considered as an alternative to MAIC.

A limitation of this study is that we have considered binary covariates only. The reason being is that the simulation study was motivated by the real world ndMM example, where only binary covariates were available, thus ensuring real world applicability. We note that it would have been preferable to include more detailed information on the covariates in the ndMM example as well, such as including the exact age of each patient rather than the binary covariate of greater or less than sixty. For instance, Schmitz et al. (2012) have noted the information lost when continuous outcomes are dichotomised. Unfortunately, however, due to the difficulty involved in sharing IPD, only binary covariates were available to us. An extension to the work could be to include continuous covariates also.

When carrying out the reweighting element of the MAIC we assumed that all covariates were independent of each other. However, in reality there could be interactions or correlations between the covariates. Therefore, it would be preferable to include an interaction term in the reweighting model, as suggested in Phillippo et al. (2016).

In the simulation study we have assumed that all effect modifiers are correctly identified and we have accounted for them in the MAIC. However, a further extension of this study could be to either have extra unidentified effect modifiers, or to assess the impact of incorrectly assuming that a prognostic variable is an effect modifier and using this in the MAIC as well. We touched on this when the standard deviation of the covariate-treatment was zero, as in this case the covariate is an prognostic variable, but we are treating it as if it is an effect modifier. This could be explored further in future work.

In our simulation study, the median models generally produced a lower MAE than the hazard ratio models, especially when there is a large covariate-treatment interaction. An extension of this work could be to investigate the performance of other survival distributions, and to investigate how well the exponential model performs when the simulated data does not come from an exponential distribution.

In the ndMM example, results are quite similar with and without the MAIC adjustment. However, by carrying out the MAIC it gives us increased confidence in the result, as it shows that results hold across populations. Were the results to have been quite contradictory, we should have more trust in the standard NMA results, as this is a higher form of evidence. We wish to stress once again that RCTs are the highest standard of evidence. It should be noted that post hoc adjustment of results cannot be a substitute for randomisation in clinical trials.

MAIC can be much more susceptible to publication bias than a clinical trial since at present there is no mechanism in place to ensure that an MAIC is registered before it is carried out. MAICs are relatively easy to undertake, provided one has the IPD, which is most likely owned by the manufacturer of the treatment. If an MAIC has negative implications for the owner of the IPD, then there is no obligation to publish the results. There is also a lot of flexibility for the analysts to choose which covariates to adjust for, which gives greater scope to choose covariates that may give better results for the owners of the IPD treatment. It is a limitation of our simulation study that we cannot explore such aspects here, though we do caution users of MAIC results on the possibility of these potential biases.

5.4.1 Conclusion

We have observed limited benefit to MAIC in the full network population. While MAIC can be beneficial as a sensitivity analysis to confirm results across patient populations, we advise that MAIC is used and interpreted with caution. We recommend that researchers use either a standard NMA model or a mixed AgD/IPD NMA model for their base case analysis.

Chapter 6

Conclusion

6.1 Summary

This thesis assesses a number of methods for making the most out of all available evidence, even when the evidence is of a lower quality than we would like. We investigate the benefits and drawbacks of each method employed, all the while being mindful that some of these methods may introduce bias into estimates of relative treatment effects.

In Chapter 3 we evaluated the benefits and drawbacks of including single-arm evidence in an NMA. We found that when the between-study variability was low, it is potentially beneficial to incorporate this type of evidence, provided that it is unbiased. However, when the variability between studies is high, aggregate level matching of single-arm trials can produce quite an inaccurate estimate. We suggest that researchers attempt to quantify the between-study variability, by examining the baseline effect across all studies. The estimates for the networks analysed in this chapter were close to the crossing point where including matched evidence produces more biased estimates, which indicates that this method may not be suitable for many networks where RCT data is also available. In addition, even when the between-study variability estimate is low, there can still be quite a lot of uncertainty associated with this estimate. Therefore, we also recommend carrying out sensitivity analyses, as suggested in Chapter 3. Of course, there are some situations where including matched evidence is the only option available to statisticians to connect the network. However, it is important to bear in mind that any method of including single-arm trials can never replace RCTs, both due to the inherent inferior data quality and the relative ease with which it can be manipulated. Therefore, it is crucial that companies still continue to run unbiased RCTs, whenever feasible.

In Chapter 4 we assessed the benefit of incorporating IPD into an NMA, whenever possible. While it is not necessary to obtain a full IPD network, obtaining a few IPD trials can be quite beneficial in an NMA. This is particularly true when trying to understand the nature of the covariate-treatment interaction through the DIC.

In Chapter 5 we assessed the benefit of an anchored MAIC with a time-to-event outcome. We found that while the MAIC can increase the accuracy when considering the population that we are matching to, we may lose generalisability with regards to the total network population. MAIC can also increase our uncertainty in the estimates due to a smaller effective sample size, but the increase in posterior SD does not always reflect the full increase in uncertainty. We therefore recommend to use MAIC as a sensitivity analysis only if we are specifically interested in the target population, otherwise using a standard NMA model or a mixed AgD/IPD model is preferable. We also investigated the best way to utilise multiple IPD studies of the same comparison in an MAIC. We found that pooling the IPD studies can lead to less uncertainty, but may not be as accurate with regards to the target AgD trial population as matching each IPD trial separately.

6.2 Discussion

In this thesis we explore a number of ways to deal with poor quality evidence. Overall, there can be benefits to making the most of the evidence available to us, but a consistent theme is that care is required to ensure correct use of methods and the awareness that is needed to avoid bias.

The most radical method of dealing with imperfect evidence explored in this thesis is aggregate level matching of single armed studies. This method may be necessary as pragmatism is needed at times in order to work with the available data. However, it is also important that we do not lose the emphasis on RCTs, as indicated in Grieve et al. (2016), and we do not want to disincentivise running RCTs, by lax acceptance of these methods, and by extension of poorer quality evidence. We need to ensure that pharmaceutical companies continue to run large, high quality RCTs to prove that their treatments work and that they are cost effective. Therefore we do not recommend making reimbursement decisions based on aggregate level matching alone. However, there are some situations in which we need information on the efficacy of treatment when RCTs and IPD are unavailable, and therefore, we attempt to present a framework and some guidance for making decisions in such an environment.

The more decisions that need to be made in any analysis the more susceptible it is to manipulation. We can look at the example of matching by covariates, through either aggregate level matching or MAIC. If a party with a vested interest is choosing the covariates to match on then they can choose the covariates which best suit their interests. This is clearly a much greater problem for an unanchored MAIC than an anchored MAIC. However, even with an anchored MAIC there is still the possibility for manipulation. The main danger here is that an investigator can check the results of many combinations of covariates with relative ease, before sharing results, and then provide a justification after the analysis. Although we found benefits to MAIC over a standard NMA in the target population of interest, these same benefits could be seen by using a mixed AgD/IPD model, which does not require a post hoc reweighting of data, and hence doesn't have the same negative impact on the full network population.

We have considered observational evidence as well in this thesis. Although it may be prone to more bias than RCTs, the TRIO and ICORN datasets in Chapter 4 demonstrate the wealth of information that is available from this form of evidence. We should also note that all three studies included in the IPD chapter came from observational evidence. It is clearly much easier to share this type of evidence, as opposed to proprietary evidence owned by pharmaceutical companies running RCTs. With the detailed IPD that is available from some observational studies, there is great potential to adjust for bias arising from differing levels of patient characteristics. We have seen a clear benefit of using IPD, in both identifying the nature of the covariates, and adjusting for differing levels of covariates across trials in a network.

Overall, novel statistical methods in general have the potential to improve analysis and decision making, but come with the caveat that care is required. This thesis has been no exception, and while these methods can add value, caution must be exercised when independently evaluating results from methods such as these.

6.3 Future Work

A number of extensions can be made to the work carried out so far:

- For the single-arm methodology these include:
    - Further investigation to find the best method for choosing a matched arm. While this is quite clear in circumstances where we have identified one covariate only, having a number of covariates leads to the additional challenge of how to weight each covariate. We have identified a number of methods for choosing the weights but it is yet unclear if there is one method that is more effective than the rest.

  - Instead of choosing the best match we could choose a weighted average of the top x number of matches. This would be an interesting way to extend this current work. Additionally, we could look at choosing one arm at a given probability, which is determined by the distance of each arm from the single arm of interest.
  - Compare the matching methods in this thesis to other available matching methods, such as random effects on baseline (Thom et al. (2015)). This has already begun in an ISPOR workshop (Thom et al. (2017)).

- For the IPD chapter the simulation study can be extended to:
  - Networks with more than one identified covariate.
  - Using an applied example with continuous covariates instead of binary covariates.
  - Using further models from Donegan et al. (2013), which consider within-trial and across-trial treatment-by-covariate interactions separately.

- Work on the MAIC chapter can be extended by:
  - Analysing the impact of IPD on a disconnected network, and assessing whether it is less biased than the aggregate level matching explored in this thesis.
  - Undertaking a similar analysis of MAIC using a binomial or continuous outcome or using continuous covariates.
  - We compared a median time-to-event model assuming an exponential survival rate to a HR model. The simulation study was set up such that the true survival function followed an exponential model also. However, it could be worth investigating how these two models compared when the true survival function differs from the assumption in the models.
  - Undertaking a similar analysis of the other major population adjusted indirect comparison method, Simulated Treatment Comparison (STC).
  - Comparing the strengths and limitations of MAIC versus the strengths and limitations of STC, in order to identify if one method is likely to perform better than the other in a given context.
  - Given the particularly large potential for publication bias in an MAIC it could be useful to analyse past NMAs which used MAIC to see if this has arisen. In particular, we could examine previous unanchored MAIC or STC analyses which have resulted in positive recommendations from HTA agencies to see if any later trials confirm the conclusions reached.

- Other future work includes:
    - We have not included any continuous outcome in this thesis. All chapters could be extended to consider continuous outcomes.
    - We saw the IPD chapter that the exchangeable model had a posterior SD that was much larger than its estimate error. We also saw that the hierarchical model has much larger posterior SDs than estimate errors. The hierarchical model assumes exchangeability on the study type level. Further investigation is necessary to investigate why exchangeable models produce such large posterior SDs.

6.4 Final Remarks

Under ideal circumstances, we could compare all treatments using standard NMA techniques. However, in this thesis we have seen that this is not always possible, either because of disconnected or single-arm evidence, or because of differing patient populations. In this thesis we have provided and critically assessed techniques to account for these difficulties. While these techniques can be crucial in assessing new treatments, we must always remember the importance of comparing these techniques to the ideal scenario of using the highest standard of evidence to ensure that bias is kept to a minimum. Although it is important to investigate these methods we must remember that even highly sophisticated statistical methods are unlikely to be able to replace an RCT.

Bibliography

Afdhal, N., Zeuzem, S., Kwo, P., Chojkier, M., Gitlin, N., Puoti, M., Romero-Gomez, M., Zarski, J.-P., Agarwal, K., Buggisch, P. et al. (2014), 'Ledipasvir and sofosbuvir for untreated HCV genotype 1 infection', *New England Journal of Medicine* **370**(20), 1889–1898.

Akobeng, A. (2005), 'Understanding randomised controlled trials', *Archives of disease in childhood* **90**(8), 840–844.

Ascierto, P. A., McArthur, G. A., Dréno, B., Atkinson, V., Liszkay, G., Di Giacomo, A. M., Mandalà, M., Demidov, L., Stroyakovskiy, D., Thomas, L. et al. (2016), 'Cobimetinib combined with vemurafenib in advanced brafv600-mutant melanoma (cobrim): updated efficacy results from a randomised, double-blind, phase 3 trial', *The Lancet Oncology* **17**(9), 1248–1260.

Ashby, D. (2006), 'Bayesian statistics in medicine: a 25 year review', *Statistics in medicine* **25**(21), 3589–3631.

Attal, M., Lauwers-Cances, V., Marit, G., Caillot, D., Moreau, P., Facon, T., Stoppa, A. M., Hulin, C., Benboubker, L., Garderet, L. et al. (2012), 'Lenalidomide maintenance after stem-cell transplantation for multiple myeloma', *New England Journal of Medicine* **366**(19), 1782–1791.

Austin, P. C. (2011), 'A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality', *Multivariate behavioral research* **46**(1), 119–151.

Bailey, A. L. (1950), 'Credibility procedures', *Proceedings of the Causalty Actuarial Society* pp. 7–23.

Balch, C. M., Gershenwald, J. E., Soong, S.-j., Thompson, J. F., Atkins, M. B., Byrd, D. R., Buzaid, A. C., Cochran, A. J., Coit, D. G., Ding, S. et al. (2009), 'Final version of 2009 AJCC melanoma staging and classification', *Journal of clinical oncology* **27**(36), 6199–6206.

Belger, M., Brnabic, A., Kadziola, Z., Petto, H. & Faries, D. (2015), 'Inclusion of multiple studies in matching adjusted indirect comparisons (MAIC)', *Value in Health* **18**(3), A33.

Bell, H., Wailoo, A. J., Hernandez, M., Grieve, R., Faria, R., Gibson, L. & Grimm, S. (2016), 'The use of real world data for the estimation of treatment effects in NICE decision making. Report by the Decision Support Unit, ScHARR, University of Sheffield'.

Berger, J. O. & Berry, D. A. (1988), 'Statistical analysis and the illusion of objectivity', *American Scientist* **76**(2), 159–165.

Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. (2010), 'A basic introduction to fixed-effect and random-effects models for meta-analysis', *Research synthesis methods* **1**(2), 97–111.

Borenstein, M., Hedges, L. V., Higgins, J. & Rothstein, H. R. (2009*a*), *Introduction to Meta-Analysis*, Wiley Online Library.

Borenstein, M., Hedges, L. V., Higgins, J. & Rothstein, H. R. (2009*b*), *Meta-Regression*, Wiley Online Library.

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

Caldwell, D. M., Ades, A. & Higgins, J. (2005), 'Simultaneous comparison of multiple treatments: combining direct and indirect evidence', *BMJ: British Medical Journal* **331**(7521), 897.

Cameron, C., Fireman, B., Hutton, B., Clifford, T., Coyle, D., Wells, G., Dormuth, C. R., Platt, R. & Toh, S. (2015), 'Network meta-analysis incorporating randomized controlled trials and non-randomized comparative cohort studies for assessing the safety and effectiveness of medical treatments: challenges and opportunities', *Systematic reviews* **4**(1), 147.

Chib, S. & Greenberg, E. (1995), 'Understanding the metropolis-hastings algorithm', *The american statistician* **49**(4), 327–335.

Cooper, N. J., Peters, J., Lai, M. C., Juni, P., Wandel, S., Palmer, S., Paulden, M., Conti, S., Welton, N. J., Abrams, K. R. et al. (2011), 'How valuable are multiple treatment comparison methods in evidence-based health-care evaluation?', *Value in Health* **14**(2), 371–380.

Cornfield, J. (1951), 'A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix', *Journal of the National Cancer Institute* **11**(6), 1269–1275.

Debray, T., Moons, K. G., Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. & Reitsma, J. B. (2015), 'Get real in individual participant data (IPD) meta-analysis: a review of the methodology', *Research synthesis methods* **6**(4), 293–309.

Debray, T. P., Schuit, E., Efthimiou, O., Reitsma, J. B., Ioannidis, J. P., Salanti, G., Moons, K. G. & Workpackage, G. (2016), 'An overview of methods for network meta-analysis using individual participant data: when do benefits arise?', *Statistical methods in medical research* pp. 1351–1364.

Dias, S. & Ades, A. (2016), 'Absolute or relative effects? Arm-based synthesis of trial data', *Research synthesis methods* **7**(1), 23–28.

Dias, S., Ades, A., Welton, N. J., Jansen, J. P. & Sutton, A. J. (2018), *Network meta-analysis for decision-making*, John Wiley & Sons.

Dias, S., Sutton, A. J., Ades, A. & Welton, N. J. (2013), 'Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials', *Medical Decision Making* **33**(5), 607–617.

Dias, S., Sutton, A. J., Welton, N. J. & Ades, A. (2013), 'Evidence synthesis for decision making 3: heterogeneitysubgroups, meta-regression, bias, and bias-adjustment', *Medical Decision Making* **33**(5), 618–640.

Dias, S., Welton, N. J., Sutton, A. J., Caldwell, D. M., Lu, G. & Ades, A. (2013), 'Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials', *Medical Decision Making* **33**(5), 641–656.

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

Donegan, S., Williamson, P., D'Alessandro, U., Garner, P. & Smith, C. T. (2013), 'Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials', *Statistics in medicine* **32**(6), 914–930.

Dore, G. J., Conway, B., Luo, Y., Janczewska, E., Knysz, B., Liu, Y., Streinu-Cercel, A., Caruntu, F. A., Curescu, M., Skoien, R. et al. (2016), 'Efficacy and safety of ombitasvir/paritaprevir/r and dasabuvir compared to IFN-containing regimens in genotype 1 HCV patients: The MALACHITE-I/II trials', *Journal of hepatology* **64**(1), 19–28.

DuMouchel, W. H. & Harris, J. E. (1983), 'Bayes methods for combining the results of cancer studies in humans and other species', *Journal of the American Statistical Association* **78**(382), 293–308.

Eddy, D. M., Hasselblad, V., Shachter, R. et al. (1992), *Meta-analysis by the confidence profile method*, Academic Press London.

Efthimiou, O., Mavridis, D., Debray, T., Samara, M., Belger, M., Siontis, G., Leucht, S. & Salanti, G. (2017), 'Combining randomized and non-randomized evidence in network meta-analysis', *Statistics in medicine* **36**(8), 1210–1226.

Faltinsen, E. G., Storebø, O. J., Jakobsen, J. C., Boesen, K., Lange, T. & Gluud, C. (2018), 'Network meta-analysis: the highest level of medical evidence?', *BMJ Evidence-Based Medicine* **23**(2), 56–59.

Faraoni, D. & Schaefer, S. T. (2016), 'Randomized controlled trials vs. observational studies: why not just live together?', *BMC anesthesiology* **16**(1), 102.

Feld, J. J., Kowdley, K. V., Coakley, E., Sigal, S., Nelson, D. R., Crawford, D., Weiland, O., Aguilar, H., Xiong, J., Pilot-Matias, T. et al. (2014), 'Treatment of HCV with ABT-450/r–ombitasvir and dasabuvir with ribavirin', *New England Journal of Medicine* **370**(17), 1594–1603.

Flaherty, K. T., Infante, J. R., Daud, A., Gonzalez, R., Kefford, R. F., Sosman, J., Hamid, O., Schuchter, L., Cebon, J., Ibrahim, N. et al. (2012), 'Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations', *New England Journal of Medicine* **367**(18), 1694–1703.

Flaherty, K. T., Robert, C., Hersey, P., Nathan, P., Garbe, C., Milhem, M., Demidov, L. V., Hassel, J. C., Rutkowski, P., Mohr, P. et al. (2012), 'Improved survival with mek inhibition in braf-mutated melanoma', *New England Journal of Medicine* **367**(2), 107–114.

Flamm, S., Bacon, B., Curry, M., Dieterich, D., Kowdley, K., Milligan, S., Tsai, N., Younossi, Z. & Afdhal, N. (2017), Real-world treatment utilization and results in the renaissance of HCV care: analyses of treatment for 8,955 patients from the TRIO network, *in* 'The International Liver Congress', Amsterdam, The Netherlands.

Flury, B. K. & Riedwyl, H. (1986), 'Standard distance in univariate and multivariate analysis', *The American Statistician* **40**(3), 249–251.

Fried, M. W., Buti, M., Dore, G. J., Flisiak, R., Ferenci, P., Jacobson, I., Marcellin, P., Manns, M., Nikitin, I., Poordad, F. et al. (2013), 'Once-daily simeprevir (TMC435) with pegylated interferon and ribavirin in treatment-naïve genotype 1 hepatitis c: The randomized PILLAR study', *Hepatology* **58**(6), 1918–1929.

Gane, E. J., Stedman, C. A., Hyland, R. H., Ding, X., Svarovskaia, E., Subramanian, G. M., Symonds, W. T., McHutchison, J. G. & Pang, P. S. (2014), 'Efficacy of nucleotide polymerase inhibitor sofosbuvir plus the NS5A inhibitor ledipasvir or the NS5B non-nucleoside inhibitor GS-9669 against HCV genotype 1 infection', *Gastroenterology* **146**(3), 736–743.

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

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* **7**, 457–472.

Gelman, A., Shirley, K. et al. (2011), 'Inference from simulations and monitoring convergence', *Handbook of markov chain monte carlo* **6**, 163–174.

Good, I. J. (1950), *Probability and the Weighing of Evidence*, Hafner Publishing Company.

Goring, S., Gustafson, P., Liu, Y., Saab, S., Cline, S. & Platt, R. (2016), 'Disconnected by design: analytic approach in treatment networks having no common comparator', *Research Synthesis Methods* .

Gray, E., Leahy, J., O'Leary, A., Norris, S. & Walsh, C. (2015), 'Estimation of the relative efficacy of licensed regimens for genotype-1 hcv infection using a mixed treatment comparison', *Value in Health* **18**(7), A577.

Gray, E., O'Leary, A., Bergin, C. & Norris, S. (2017), 'Effectiveness of interferon-free therapy for the treatment of hcv-patients with compensated cirrhosis treated through the Irish early access program', *Expert Review of Gastroenterology & Hepatology* pp. 1–9.

Green, S. & Higgins, J. (2005), 'Cochrane handbook for systematic reviews of interventions'.

Grieve, R., Abrams, K., Claxton, K., Goldacre, B., James, N., Nicholl, J., Parmar, M., Parker, C., Sekhon, J. S., Smeeth, L. et al. (2016), 'Cancer drugs fund requires further reform', *Bmj* **354**.

Griffiths, E. A., Macaulay, R., Vadlamudi, N. K., Uddin, J. & Samuels, E. R. (2017), 'The role of noncomparative evidence in health technology assessment decisions', *Value in Health* .

Guyot, P., Ades, A., Ouwens, M. J. & Welton, N. J. (2012), 'Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves', *BMC medical research methodology* **12**(1), 1.

Haidich, A.-B. (2010), 'Meta-analysis in medical research', *Hippokratia* **14**(Suppl 1), 29.

Hamid, O., Puzanov, I., Dummer, R., Schachter, J., Daud, A., Schadendorf, D., Blank, C., Cranmer, L., Robert, C., Pavlick, A. et al. (2016), 'Final overall survival for keynote-002: pembrolizumab (pembro) versus investigator-choice chemotherapy (chemo) for ipilimumab (ipi)-refractory melanoma', *Annals of Oncology* **27**(suppl_6).

Hauschild, A., Grob, J. J., Demidov, L. V., Jouary, T., Gutzmer, R., Millward, M., Rutkowski, P., Blank, C. U., Miller, W. H., Kaempgen, E. et al. (2013), 'An update on break-3, a phase iii, randomized trial: Dabrafenib (dab) versus dacarbazine (dtic) in patients with braf v600e-positive mutation metastatic melanoma (mm).'.

Hersh, E. M., ODay, S. J., Powderly, J., Khan, K. D., Pavlick, A. C., Cranmer, L. D., Samlowski, W. E., Nichol, G. M., Yellin, M. J. & Weber, J. S. (2011), 'A phase ii multicenter study of ipilimumab with or without dacarbazine in chemotherapy-naive patients with advanced melanoma', *Investigational new drugs* **29**(3), 489–498.

Hess, K. R. (1995), 'Graphical methods for assessing violations of the proportional hazards assumption in cox regression', *Statistics in medicine* **14**(15), 1707–1723.

Hézode, C., Forestier, N., Dusheiko, G., Ferenci, P., Pol, S., Goeser, T., Bronowicki, J.-P., Bourlière, M., Gharakhanian, S., Bengtsson, L. et al. (2009), 'Telaprevir and peginterferon with or without ribavirin for chronic HCV infection', *New England Journal of Medicine* **360**(18), 1839–1850.

Hickman, J. C. & Heacox, L. (1999), 'Credibility theory: the cornerstone of actuarial science', *North American Actuarial Journal* **3**(2), 1–8.

Higgins, J. P. & Thompson, S. G. (2002), 'Quantifying heterogeneity in a meta-analysis', *Statistics in medicine* **21**(11), 1539–1558.

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

Higgins, J. P. & Whitehead, A. (1996), 'Borrowing strength from external trials in a meta-analysis', *Statistics in medicine* **15**(24), 2733–2749.

Hodi, F. S., O'day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C. et al. (2010), 'Improved survival with ipilimumab in patients with metastatic melanoma', *New England Journal of Medicine* **363**(8), 711–723.

Hong, H., Chu, H., Zhang, J. & Carlin, B. P. (2016*a*), 'A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons', *Research synthesis methods* **7**(1), 6–22.

Hong, H., Chu, H., Zhang, J. & Carlin, B. P. (2016*b*), 'Rejoinder to the discussion of "A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons" by s. dias and ae ades', *Research synthesis methods* **7**(1), 29–33.

Hong, H., Fu, H. & Carlin, B. P. (2018), 'Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* pp. 1047–1069.

Hunter, J. E. & Schmidt, F. L. (2000), 'Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge', *International Journal of Selection and Assessment* **8**(4), 275–292.

Ibrahim, J. G., Chen, M.-H. et al. (2000), 'Power prior distributions for regression models', *Statistical Science* **15**(1), 46–60.

Ikle, F. C., Aronson, G. J. & Madansky, A. (1958), 'On the risk of an accidental or unauthorized nuclear detonation'.

Ioannidis, J. P., Haidich, A.-B., Pappa, M., Pantazis, N., Kokori, S. I., Tektonidou, M. G., Contopoulos-Ioannidis, D. G. & Lau, J. (2001), 'Comparison of evidence of treatment effects in randomized and nonrandomized studies', *Journal of the American Medical Association* **286**(7), 821–830.

Jackson, G. H., Davies, F. E., Pawlyn, C., Cairns, D. A., Striha, A., Collett, C., Waterhouse, A., Jones, J. R., Kishore, B., Garg, M. et al. (2016), 'Response adapted induction treatment improves outcomes for myeloma patients; results of the phase III myeloma xi study'.

Jacobson, I. M., Dore, G. J., Foster, G. R., Fried, M. W., Radu, M., Rafalsky, V. V., Moroz, L., Craxi, A., Peeters, M., Lenz, O. et al. (2014), 'Simeprevir with pegylated interferon alfa 2a plus ribavirin in treatment-naive patients with chronic hepatitis C virus genotype 1 infection (QUEST-1): a phase 3, randomised, double-blind, placebo-controlled trial', *The Lancet* **384**(9941), 403–413.

Jacobson, I. M., McHutchison, J. G., Dusheiko, G., Di Bisceglie, A. M., Reddy, K. R., Bzowej, N. H., Marcellin, P., Muir, A. J., Ferenci, P., Flisiak, R. et al. (2011), 'Telaprevir for previously untreated chronic hepatitis C virus infection', *New England Journal of Medicine* **364**(25), 2405–2416.

Jaff, M. R., Nelson, T., Ferko, N., Martinson, M., Anderson, L. H. & Hollmann, S. (2017), 'Endovascular interventions for femoropopliteal peripheral artery disease: A network meta-analysis of current technologies', *Journal of Vascular and Interventional Radiology* pp. 1617–1627.

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

Jansen, J. P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., Lee, K., Boersma, C., Annemans, L. & Cappelleri, J. C. (2011), 'Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1', *Value in Health* **14**(4), 417–428.

Jansen, J. P. & Naci, H. (2013), 'Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers', *BMC medicine* **11**(1), 1.

Kanters, S., Ford, N., Druyts, E., Thorlund, K., Mills, E. J. & Bansback, N. (2016), 'Use of network meta-analysis in clinical guidelines', *Bulletin of the World Health Organization* **94**(10), 782.

Kass, R. E. & Wasserman, L. (1996), 'The selection of prior distributions by formal rules', *Journal of the American Statistical Association* **91**(435), 1343–1370.

Khan, K. S., Ter Riet, G., Glanville, J., Sowden, A. J., Kleijnen, J. et al. (2001), *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*, number 4 (2n), NHS Centre for Reviews and Dissemination.

Korn, E. L., Liu, P.-Y., Lee, S. J., Chapman, J.-A. W., Niedzwiecki, D., Suman, V. J., Moon, J., Sondak, V. K., Atkins, M. B., Eisenhauer, E. A. et al. (2008), 'Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials', *Journal of Clinical Oncology* **26**(4), 527–534.

Krauss, A. (2018), 'Why all randomised controlled trials produce biased results', *Annals of medicine* **50**(4), 312–322.

Kühnast, S., Schiffner-Rohe, J., Rahnenfuehrer, J., Leverkus, F. et al. (2017), 'Evaluation of adjusted and unadjusted indirect comparison methods in benefit assessment', *Methods of information in medicine* **56**(3), 261–267.

Kumada, H., Toyota, J., Okanoue, T., Chayama, K., Tsubouchi, H. & Hayashi, N. (2012), 'Telaprevir with peginterferon and ribavirin for treatment-naive patients chronically infected with HCV of genotype 1 in Japan', *Journal of hepatology* **56**(1), 78–84.

Kunz, R. & Oxman, A. D. (1998), 'The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials', *Bmj* **317**(7167), 1185–1190.

Kwo, P., Gitlin, N., Nahass, R., Bernstein, D., Rojter, S., Schiff, E., Davis, M., Ruane, P. J., Younes, Z., Kalmeijer, R. et al. (2015), 'A phase-3, randomised, open-label study to evaluate the efficacy and safety of 8 and 12 weeks of Simeprevir (SMV) plus Sofosbuvir (SOF) in treatment-naive and-experienced patients with chronic HCV genotype 1 infection without cirrhosis: Optimist-1', *J Hepatol* **62**(Suppl 2), S270.

Kwo, P. Y., Lawitz, E. J., McCone, J., Schiff, E. R., Vierling, J. M., Pound, D., Davis, M. N., Galati, J. S., Gordon, S. C., Ravendhran, N. et al. (2010), 'Efficacy of boceprevir, an NS3 protease inhibitor, in combination with peginterferon alfa-2b and ribavirin in treatment-naive patients with genotype 1 hepatitis C infection (SPRINT-1): an open-label, randomised, multicentre phase 2 trial', *The Lancet* **376**(9742), 705–716.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. & Jones, D. R. (2005), 'How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBugs', *Statistics in medicine* **24**(15), 2401–2428.

Larkin, J., Minor, D., D'Angelo, S., Neyns, B., Smylie, M., Miller, W., Gutzmer, R., Linette, G., Chmielowski, B., Lao, C. D. et al. (2017), 'Overall survival in patients with advanced melanoma who received nivolumab versus investigator's choice chemotherapy in checkmate 037: a randomized, controlled, open-label phase iii trial.', *Journal of Clinical Oncology* .

Lawitz, E., Lalezari, J. P., Hassanein, T., Kowdley, K. V., Poordad, F. F., Sheikh, A. M., Afdhal, N. H., Bernstein, D. E., DeJesus, E., Freilich, B. et al. (2013), 'Sofosbuvir in combination with peginterferon alfa-2a and ribavirin for non-cirrhotic, treatment-naive patients with genotypes 1, 2, and 3 hepatitis C infection: a randomised, double-blind, phase 2 trial', *The Lancet infectious diseases* **13**(5), 401–408.

Lawitz, E., Mangia, A., Wyles, D., Rodriguez-Torres, M., Hassanein, T., Gordon, S. C., Schultz, M., Davis, M. N., Kayali, Z., Reddy, K. R. et al. (2013), 'Sofosbuvir for previously untreated chronic hepatitis C infection', *New England Journal of Medicine* **368**(20), 1878–1887.

Leahy, J., OLeary, A., Afdhal, N., Gray, E., Milligan, S., Wehmeyer, M. H. & Walsh, C. (2018), 'The impact of individual patient data in a network meta analysis: An investigation into parameter estimation and model selection', *Research Synthesis Methods* . doi:10.1002/jrsm/term.1305.

Long, G. V., Stroyakovskiy, D., Gogas, H., Levchenko, E., De Braud, F., Larkin, J., Garbe, C., Jouary, T., Hauschild, A., Grob, J.-J. et al. (2015), 'Dabrafenib and trametinib versus dabrafenib and placebo for val600 braf-mutant melanoma: a multicentre, double-blind, phase 3 randomised controlled trial', *The Lancet* **386**(9992), 444–451.

Lu, G. & Ades, A. (2004), 'Combination of direct and indirect evidence in mixed treatment comparisons', *Statistics in medicine* **23**(20), 3105–3124.

Lumley, T. (2002), 'Network meta-analysis for indirect treatment comparisons', *Statistics in medicine* **21**(16), 2313–2324.

Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2012), *The BUGS book: A practical introduction to Bayesian analysis*, CRC press.

Manns, M., Marcellin, P., Poordad, F., de Araujo, E. S. A., Buti, M., Horsmans, Y., Janczewska, E., Villamil, F., Scott, J., Peeters, M. et al. (2014), 'Simeprevir with pegylated interferon alfa 2a or 2b plus ribavirin in treatment-naive patients with chronic hepatitis C virus genotype 1 infection (QUEST-2): a randomised, double-blind, placebo-controlled phase 3 trial', *The Lancet* **384**(9941), 414–426.

McArthur, G. A., Chapman, P. B., Robert, C., Larkin, J., Haanen, J. B., Dummer, R., Ribas, A., Hogg, D., Hamid, O., Ascierto, P. A. et al. (2014), 'Safety and efficacy of vemurafenib in brafv600e and brafv600k mutation-positive melanoma (brim-3): extended follow-up of a phase 3, randomised, open-label study', *The lancet oncology* **15**(3), 323–332.

McCarthy, P. L., Owzar, K., Hofmeister, C. C., Hurd, D. D., Hassoun, H., Richardson, P. G., Giralt, S., Stadtmauer, E. A., Weisdorf, D. J., Vij, R. et al. (2012), 'Lenalidomide after stem-cell transplantation for multiple myeloma', *New England Journal of Medicine* **366**(19), 1770–1781.

McHutchison, J. G., Everson, G. T., Gordon, S. C., Jacobson, I. M., Sulkowski, M., Kauffman, R., McNair, L., Alam, J. & Muir, A. J. (2009), 'Telaprevir with peginterferon and ribavirin for chronic HCV genotype 1 infection', *New England Journal of Medicine* **360**(18), 1827–1838.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. (2009), 'Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement', *Annals of internal medicine* **151**(4), 264–269.

Morgan, G. J., Gregory, W. M., Davies, F. E., Bell, S. E., Szubert, A. J., Brown, J. M., Coy, N. N., Cook, G., Russell, N. H., Rudin, C. et al. (2012), 'The role of maintenance thalidomide therapy in multiple myeloma: Mrc myeloma ix results and meta-analysis', *Blood* **119**(1), 7–15.

O'Hagan, A. & Forster, J. J. (2004), *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*, Vol. 2, Arnold.

O'Hagan, A. & Luce, B. (2003), 'A primer on bayesian statistics in health economics and outcomes research', *Sheffield: Centre for Bayesian Statistics in Health Economics* .

O'Rourke, K. (2007), 'An historical perspective on meta-analysis: dealing quantitatively with varying study results', *Journal of the Royal Society of Medicine* **100**(12), 579–582.

Osinusi, A., Meissner, E. G., Lee, Y.-J., Bon, D., Heytens, L., Nelson, A., Sneller, M., Kohli, A., Barrett, L., Proschan, M. et al. (2013), 'Sofosbuvir and ribavirin for hepatitis C genotype 1 in patients with unfavorable treatment characteristics: a randomized clinical trial', *Journal of the American Medical Association* **310**(8), 804–811.

Owen, R. K., Tincello, D. G. & Keith, R. A. (2015), 'Network meta-analysis: development of a three-level hierarchical modeling approach incorporating dose-related constraints', *Value in Health* **18**(1), 116–126.

Palumbo, A., Cavallo, F., Gay, F., Di Raimondo, F., Ben Yehuda, D., Petrucci, M. T., Pezzatti, S., Caravita, T., Cerrato, C., Ribakovsky, E. et al. (2014), 'Autologous transplantation and maintenance therapy in multiple myeloma', *New England Journal of Medicine* **371**(10), 895–905.

Phillippo, D., Ades, T., Dias, S., Palmer, S., Abrams, K. & Welton, N. (2016), '(2016). NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. (Technical Support Documents). Decision Support Unit, ScHARR, University of Sheffield: NICE Decision Support Unit.'.

Plummer, M. (2012), 'Jags version 3.3. 0 user manual', *International Agency for Research on Cancer, Lyon, France* .

Pol, S., Ghalib, R. H., Rustgi, V. K., Martorell, C., Everson, G. T., Tatum, H. A., Hézode, C., Lim, J. K., Bronowicki, J.-P., Abrams, G. A. et al. (2012), 'Daclatasvir for previously untreated chronic hepatitis C genotype-1 infection: a randomised, parallel-group, double-blind, placebo-controlled, dose-finding, phase 2a trial', *The Lancet infectious diseases* **12**(9), 671–677.

Poordad, F., McCone Jr, J., Bacon, B. R., Bruno, S., Manns, M. P., Sulkowski, M. S., Jacobson, I. M., Reddy, K. R., Goodman, Z. D., Boparai, N. et al. (2011), 'Boceprevir for untreated chronic HCV genotype 1 infection', *New England Journal of Medicine* **364**(13), 1195–1206.

Poppe, K. K., Doughty, R. N., Yu, C.-M., Quintana, M., Møller, J. E., Klein, A. L., Gamble, G. D., Dini, F. L., Whalley, G. A. et al. (2011), 'Understanding differences in results from literature-based and individual patient meta-analyses:

An example from meta-analyses of observational data', *International journal of cardiology* **148**(2), 209–213.

Prevost, T. C., Abrams, K. R. & Jones, D. R. (2000), 'Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening', *Statistics in medicine* **19**(24), 3359–3376.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Robert, C., Karaszewska, B., Schachter, J., Rutkowski, P., Mackiewicz, A., Stroiakovski, D., Lichinitser, M., Dummer, R., Grange, F., Mortier, L. et al. (2015), 'Improved overall survival in melanoma with combined dabrafenib and trametinib', *New England Journal of Medicine* **372**(1), 30–39.

Robert, C., Ribas, A., Wolchok, J. D., Hodi, F. S., Hamid, O., Kefford, R., Weber, J. S., Joshua, A. M., Hwu, W.-J., Gangadhar, T. C. et al. (2014), 'Anti-programmed-death-receptor-1 treatment with pembrolizumab in ipilimumab-refractory advanced melanoma: a randomised dose-comparison cohort of a phase 1 trial', *The Lancet* **384**(9948), 1109–1117.

Rothwell, P. M. (2005), 'External validity of randomised controlled trials:to whom do the results of this trial apply?', *The Lancet* **365**(9453), 82–93.

Rouse, B., Chaimani, A. & Li, T. (2017), 'Network meta-analysis: an introduction for clinicians', *Internal and emergency medicine* **12**(1), 103–111.

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

Salanti, G., Higgins, J. P., Ades, A. & Ioannidis, J. P. (2008), 'Evaluation of networks of randomized trials', *Statistical methods in medical research* **17**(3), 279–301.

Saramago, P., Sutton, A. J., Cooper, N. J. & Manca, A. (2012), 'Mixed treatment comparisons using aggregate and individual participant level data', *Statistics in medicine* **31**(28), 3516–3536.

Schmitz, S., Adams, R. & Walsh, C. (2012), 'The use of continuous data versus binary data in mtc models: a case study in rheumatoid arthritis', *BMC medical research methodology* **12**(1), 167.

Schmitz, S., Adams, R. & Walsh, C. (2013), 'Incorporating data from various trial designs into a mixed treatment comparison model', *Statistics in medicine* **32**(17), 2935–2949.

Schmitz, S., Maguire, Á., Morris, J., Ruggeri, K., Haller, E., Kuhn, I., Leahy, J., Homer, N., Khan, A., Bowden, J. et al. (2018), 'The use of single armed observational data to closing the gap in otherwise disconnected evidence networks: a network meta-analysis in multiple myeloma', *BMC medical research methodology* **18**(1), 66.

Schoenfeld, D. (1982), 'Partial residuals for the proportional hazards regression model', *Biometrika* **69**(1), 239–241.

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

Sharpe, W. F. (1994), 'The sharpe ratio', *The journal of portfolio management* **21**(1), 49–58.

Sherman, K. E., Flamm, S. L., Afdhal, N. H., Nelson, D. R., Sulkowski, M. S., Everson, G. T., Fried, M. W., Adler, M., Reesink, H. W., Martin, M. et al. (2011), 'Response-guided telaprevir combination treatment for hepatitis C virus infection', *New England Journal of Medicine* **365**(11), 1014–1024.

Shrier, I., Boivin, J.-F., Steele, R. J., Platt, R. W., Furlan, A., Kakuma, R., Brophy, J. & Rossignol, M. (2007), 'Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? a critical examination of underlying principles', *American journal of epidemiology* **166**(10), 1203–1209.

Shrier, I. & Pang, M. (2015), 'Confounding, effect modification, and the odds ratio: common misinterpretations', *Journal of clinical epidemiology* **68**(4), 470–474.

Signorovitch, J. E., Sikirica, V., Erder, M. H., Xie, J., Lu, M., Hodgkins, P. S., Betts, K. A. & Wu, E. Q. (2012), 'Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research', *Value in Health* **15**(6), 940–947.

Signorovitch, J. E., Wu, E. Q., Andrew, P. Y., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R. & Mulani, P. M. (2010), 'Comparative effectiveness without head-to-head trials', *Pharmacoeconomics* **28**(10), 935–945.

Simpson, E. (2010), 'Edward simpson: Bayes at bletchley park', *Significance* **7**(2), 76–80.

Simpson, R. & Pearson, K. (1904), 'Report on certain enteric fever inoculation statistics', *The British Medical Journal* pp. 1243–1246.

Song, F., Clark, A., Bachmann, M. O. & Maas, J. (2012), 'Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons', *BMC Medical Research Methodology* **12**(1), 1–14.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004), *Bayesian approaches to clinical trials and health-care evaluation*, Vol. 13, John Wiley & Sons.

Spiegelhalter, D. J. & Best, N. G. (2003), 'Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling', *Statistics in medicine* **22**(23), 3687–3709.

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

Spiegelhalter, D. J. et al. (2004), 'Incorporating bayesian ideas into health-care evaluation', *Statistical Science* **19**(1), 156–174.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2014), 'OpenBUGS User Manual'.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I. et al. (2016), 'ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions', *British Medical Journal* **355**, i4919.

Stewart, L. A. & Tierney, J. F. (2002), 'To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data', *Evaluation & the health professions* **25**(1), 76–97.

Sud, S. & Douketis, J. (2009), 'The devil is in the details or not? a primer on individual patient data meta-analysis', *Evidence-based medicine* **14**(4), 100–101.

Sulkowski, M. S., Gardiner, D. F., Rodriguez-Torres, M., Reddy, K. R., Hassanein, T., Jacobson, I., Lawitz, E., Lok, A. S., Hinestrosa, F., Thuluvath, P. J. et al. (2014), 'Daclatasvir plus sofosbuvir for previously treated or untreated chronic HCV infection', *New England Journal of Medicine* **370**(3), 211–221.

Sutton, A. J. & Abrams, K. R. (2001), 'Bayesian methods in meta-analysis and evidence synthesis', *Statistical methods in medical research* **10**(4), 277–303.

Thom, H. H., Capkun, G., Cerulli, A., Nixon, R. M. & Howard, L. S. (2015), 'Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension', *BMC medical research methodology* **12**(1), 15–34.

Thom, H., Leahy, J. & Jansen, J. P. (2017), 'Disconnected or limited evidence in a network meta-analysis: What can be done?'. Workshop at the International Society For Pharmacoeconomics and Outcomes Research Conference, Glasgow, 2017.

Thorlund, K., Thabane, L. & Mills, E. J. (2013), 'Modelling heterogeneity variances in multiple treatment comparison meta-analysis. Are informative priors the better solution?', *BMC medical research methodology* **13**(1), 1.

Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M. & Rovers, M. (2015), 'Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use', *PLoS medicine* **12**(7), 1001855.

Tremblay, G., Holbrook, T., Milligan, G., Pelletier, C. & Rietschel, P. (2016), 'Matching-adjusted indirect treatment comparison in patients with radioiodine-refractory differentiated thyroid cancer', *Comparative Effectiveness Research* **2016**(1), 13–21.

Trentino, K., Farmer, S., Gross, I., Shander, A. & Isbister, J. (2016), 'Observational studies-should we simply ignore them in assessing transfusion outcomes?', *BMC anesthesiology* **16**(1), 96.

Tudur Smith, C., Marcucci, M., Nolan, S. J., Iorio, A., Sudell, M., Riley, R., Rovers, M. M. & Williamson, P. R. (2016), 'Individual participant data meta-analyses compared with meta-analyses based on aggregate data', *The Cochrane Library* .

Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G. & Higgins, J. P. (2012), 'Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews', *International journal of epidemiology* **41**(3), 818–827.

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. & Thompson, S. G. (2009), 'Bias modelling in evidence synthesis', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(1), 21–47.

Twombly, R. (2006), ‘Criticism of tumor response criteria raises trial design questions’, *Journal of the National Cancer Institute* **98**(4), 232–234.

Valentine, J. C. & Thompson, S. G. (2013), ‘Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions’, *Research synthesis methods* **4**(1), 26–35.

Van Sanden, S., Baculea, S., Diels, J. & Cote, S. (2017), ‘Comparative efficacy of ibrutinib versus obinutuzumab+ chlorambucil in first-line treatment of chronic lymphocytic leukemia: A matching-adjusted indirect comparison’, *Advances in Therapy* pp. 1–12.

Van Walraven, C. (2010), ‘Individual patient meta-analysis - rewards and challenges’, *Journal of clinical epidemiology* **63**(3), 235–237.

Veroniki, A. A., Straus, S. E., Soobiah, C., Elliott, M. J. & Tricco, A. C. (2016), ‘A scoping review of indirect comparison methods and applications using individual patient data’, *BMC medical research methodology* **16**(1), 47.

Wehmeyer, M. H., Eißing, F., Jordan, S., Röder, C., Hennigs, A., Degen, O., Hüfner, A., Hertling, S., Schmiedel, S., Sterneck, M. et al. (2014), ‘Safety and efficacy of protease inhibitor based combination therapy in a single-center real-life cohort of 110 patients with chronic hepatitis C genotype 1 infection’, *BMC gastroenterology* **14**(1), 1.

Welton, N., Ades, A., Carlin, J., Altman, D. & Sterne, J. (2009), ‘Models for potentially biased evidence in meta-analysis using empirically based priors’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(1), 119–136.

Welton, N. J., Caldwell, D., Adamopoulos, E. & Vedhara, K. (2009), ‘Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease’, *American Journal of Epidemiology* **169**(9), 1158–1165.

Welton, N. J., Sutton, A. J., , Cooper, N., Abrams, K. R. & Ades, A. (2012), *Evidence synthesis for decision making in healthcare*, Vol. 132, John Wiley & Sons.

Whitehead, A. (2002), *Meta-analysis of controlled clinical trials*, Vol. 7, John Wiley & Sons.

Wolchok, J. D., Neyns, B., Linette, G., Negrier, S., Lutzky, J., Thomas, L., Waterfield, W., Schadendorf, D., Smylie, M., Guthrie Jr, T. et al. (2010), ‘Ipilimumab monotherapy in patients with pretreated advanced melanoma: a randomised, double-blind, multicentre, phase 2, dose-ranging study', *The lancet oncology* **11**(2), 155–164.

Wolpert, R. L., Mengersen, K. L. et al. (2004), 'Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke', *Statistical Science* **19**(3), 450–471.

Woods, B. S., Hawkins, N. & Scott, D. A. (2010), 'Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial', *BMC medical research methodology* **10**(1), 1.

Zimmer, L., Eigentler, T. K., Kiecker, F., Simon, J., Utikal, J., Mohr, P., Berking, C., Kämpgen, E., Dippel, E., Stadler, R. et al. (2015), 'Open-label, multicenter, single-arm phase ii decog-study of ipilimumab in pretreated patients with different subtypes of metastatic melanoma', *Journal of translational medicine* **13**(1), 351.

Appendix A

For Chapter 3 (Single Arm Evidence)

In Chapter 3 we assessed the impact of incorporating single-arm evidence in an NMA using aggregate level matching on one covariate. This appendix extends the work in Chapter 3 to a scenario with 3 covariates. We also provide additional material such as the original simulations and the MC error associated with them, assumptions surrounding the choice of priors, extra details of the studies in the applied HCV infection example, and the WinBUGS code used.

A.1 Scenario With Three Covariates

In the main text of Chapter 3 we carried out a simulation study in which matching was based on one covariate. We now extend this to matching based on three covariates.

A.1.1 Methods

The goal of matching by covariates is to minimise the difference between matched arms. When we have information about multiple covariates, there is a question of how much weight we should put on each covariate. We may want to give a bigger weight to variables that more strongly influence outcome. We define M to be the number of covariates considered. The difference in each arm is computed by: $\Delta_{ij,k} = \sum_{m=1}^{M} |w_m(x_{m_{ij}} - x_{m_k})|$, where x_{m_k} is the proportion of patients with characteristic associated with the covariate m in the single agent trial with treatment k. w is an additional parameter in this equation compared to the main text, which denotes the estimated weight of each covariate.

We considered a number of different methods to reflect the weight when we have multiple identified covariates:

1. Equal Weights: Matching by equal weights. In this case each covariate is given the same weight.

2. Covariate effect: Weighting by our best estimate of the effect of each covariate. We run a meta-regression on the RCT evidence alone to obtain an estimate for the covariates of interest. We use the estimate of the mean of each covariate ($\hat{\beta}$) as the weights, and use the results of this equation to choose the matched arm.

3. Covariate effect with SD: Again, we run a meta-regression on the RCT evidence to estimate the covariates. We obtain the weight by dividing the estimate of the mean by the posterior SD, $\frac{\hat{\beta}}{\sigma_\beta}$. In this way we are penalising uncertainty in our estimate. If two different covariates both have an estimate of 2, for example, but covariate A has a standard deviation of 1 and covariate B has a standard deviation of 10, then we would put more weight on covariate A, as we can be more confident that the estimate is not due to noise. In this example covariate B would need an estimate of 20 to be given the same weight as covariate A. This is analogous to a Sharpe ratio or information ratio in finance where expectancy is divided by the standard deviation Sharpe (1994).

While methods 2 and 3 are the most informed method they require extra analysis to be carried out. In addition, the meta-regression is based on the RCT network alone, so while we would expect results to be indicative of the matched network, they may not hold exactly for the single agent trials.

For the simulation study, in addition to the methods set out above we compare the results of the matching methods to only using RCT evidence, randomly matching, and methods that require perfect information.

1. RCT Only: Including RCTs only.

2. Random Matching: Randomly selecting a matching arm.

3. True Covariate Effects: This assumes that the covariate effects β are known and used for the weights $\mathbf{w}$. Therefore if we knew that the effect of covariate A was 2 and the effect of covariate B was 0.5, then covariate A would be given four times as much weight as covariate B.

4. Equal Weights: Matching using equal weights as described in above.

5. Largest Covariate: Matching only by the covariate with the largest effect (as defined by the simulation study) while ignoring all other covariates.

6. Estimated Covariate Effects: Using the mean of β from the meta-regression as described above.

7. Estimated Covariate Effects with SD: Using the mean and posterior SD of β from the meta-regression as described above.

Methods 3 and 5 use information which we do not have in reality, but they can give us an indication of how well our model works. Therefore we include them as part of the simulation study. We have excluded the plug-in estimator model in this scenario to limit the number of lines on the graphs.

A.1.2 Results

We can see from figure A.1 that when including three covariates the study effect influences our estimate error in the same way that it did with one covariate. The effect of the three covariates are set to be $\beta = (-0.52, -1.04, -2.08)$. Random matching stands out from the other matching types as producing the worst MAE and largest posterior SD. It is difficult to tell the difference between the other matching methods, so any attempt at matching by covariates is likely to be beneficial. For this reason that we decided to use equal weights in the HCV infection application.

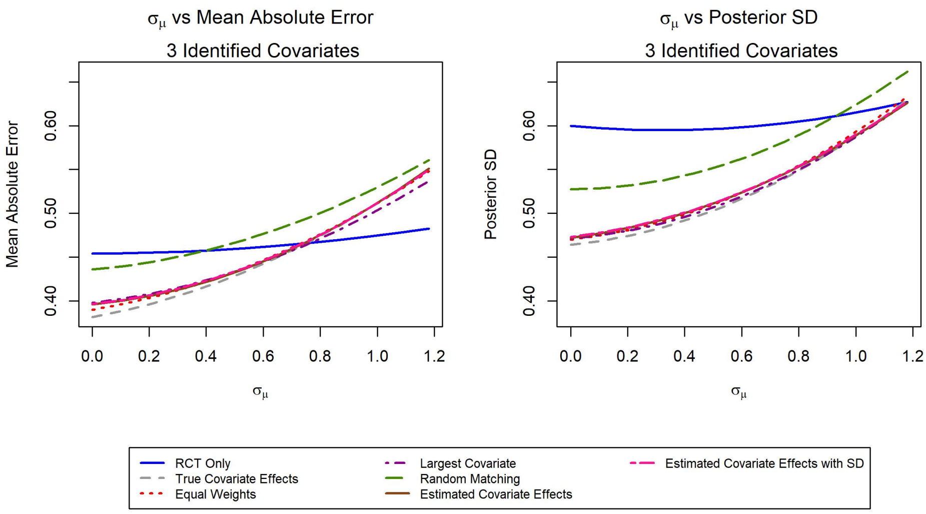


Figure A.1: Effect of the between-study variability, σ_μ, on the estimate error and posterior standard deviation. In this case three covariates have been identified, the effect, $\beta = (-0.52, -1.04, -2.08)$. The extreme left point on the graph shows the scenario where the study effect is set to zero for every study. The variability between the studies increases with the horizontal axis.

A.2 Graphs and Monte Carlo Error of Original Simulations

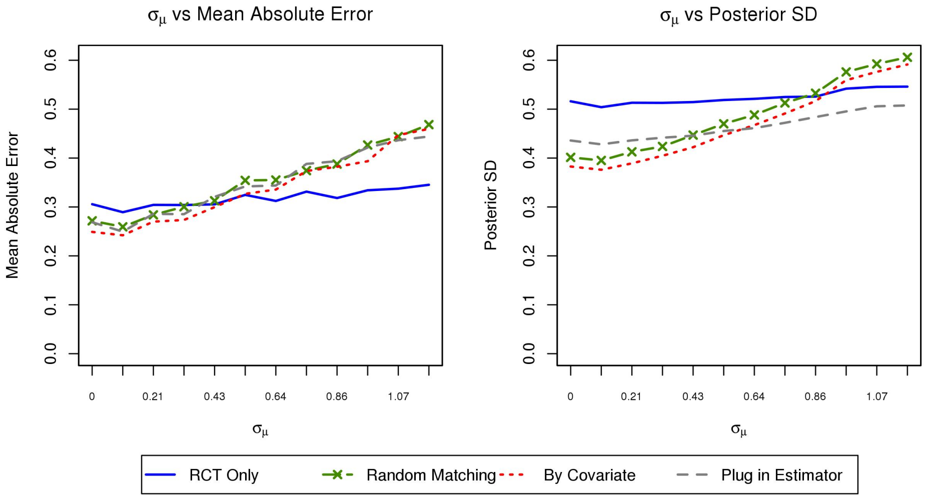


Figure A.2: Effect of the between-study variability, σ_μ, on the estimate error and posterior SD (original data). Covariate effect, β=-1.04.

Table A.1: MC Error for Figure A.2

	Mean Absolute Error				Posterior SD			
σ_μ	RCT Only	Random	By Covariate	Plug in Estimator	RCT Only	Random	By Covariate	Plug in Estimator
0	0.009	0.007	0.006	0.007	0.006	0.004	0.005	0.005
0.11	0.008	0.007	0.006	0.006	0.006	0.004	0.004	0.004
0.21	0.009	0.007	0.007	0.007	0.006	0.004	0.005	0.005
0.32	0.009	0.008	0.007	0.008	0.006	0.004	0.005	0.005
0.43	0.008	0.008	0.007	0.008	0.006	0.004	0.005	0.005
0.54	0.009	0.009	0.008	0.009	0.006	0.004	0.005	0.005
0.64	0.008	0.009	0.008	0.009	0.005	0.004	0.006	0.006
0.75	0.009	0.009	0.009	0.009	0.005	0.004	0.006	0.006
0.86	0.009	0.01	0.009	0.01	0.005	0.004	0.006	0.006
0.97	0.009	0.01	0.009	0.011	0.005	0.004	0.006	0.006
1.07	0.009	0.011	0.011	0.011	0.005	0.004	0.006	0.007
1.18	0.009	0.011	0.011	0.011	0.005	0.004	0.007	0.007

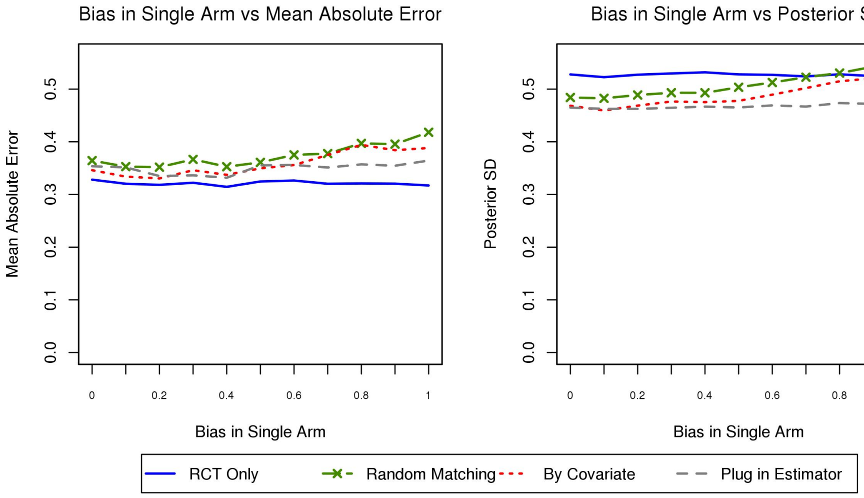


Figure A.3: Effect of bias in the single-arm trials on the estimate error and posterior SD (original data). Covariate effect, β=-1.04.

Table A.2: MC Error for Figure A.3

	Mean Absolute Error				Posterior SD			
σ_μ	RCT Only	Random	By Covariate	Plug in Estimator	RCT Only	Random	By Covariate	Plug in Estimator
0	0.009	0.01	0.009	0.009	0.004	0.006	0.005	0.004
0.1	0.009	0.009	0.008	0.009	0.004	0.006	0.005	0.004
0.2	0.009	0.008	0.008	0.008	0.004	0.006	0.005	0.004
0.3	0.009	0.009	0.009	0.008	0.004	0.006	0.006	0.004
0.4	0.008	0.008	0.008	0.008	0.004	0.006	0.006	0.004
0.5	0.008	0.008	0.008	0.008	0.004	0.006	0.005	0.004
0.6	0.009	0.008	0.009	0.008	0.004	0.006	0.006	0.004
0.7	0.008	0.009	0.009	0.008	0.004	0.006	0.006	0.004
0.8	0.008	0.009	0.01	0.009	0.004	0.006	0.006	0.004
0.9	0.008	0.009	0.009	0.008	0.004	0.006	0.006	0.004
1	0.008	0.01	0.009	0.008	0.004	0.006	0.006	0.004

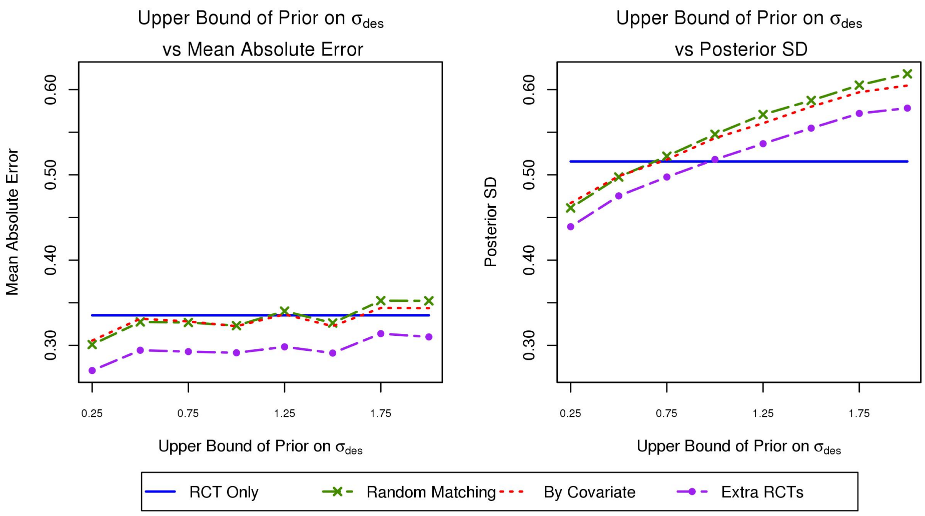


Figure A.4: Effect of the prior on the between-study design effect (σ_{des}) on the estimate error and posterior SD (original data). Between-study variability, σ_μ=0.59, covariate effect, β=-1.04.

Table A.3: MC Error for Figure A.4

	Mean Absolute Error			Posterior SD		
σ_{des}	**Random**	**By Covariate**	**Extra RCTs**	**Random**	**By Covariate**	**Extra RCTs**
0.25	0.008	0.008	0.007	0.004	0.004	0.004
0.5	0.008	0.009	0.008	0.004	0.004	0.004
0.75	0.008	0.008	0.008	0.004	0.004	0.004
1	0.008	0.008	0.008	0.004	0.005	0.004
1.25	0.008	0.008	0.007	0.004	0.004	0.004
1.5	0.007	0.007	0.007	0.004	0.004	0.004
1.75	0.008	0.008	0.007	0.005	0.005	0.005
2	0.008	0.008	0.007	0.005	0.005	0.005

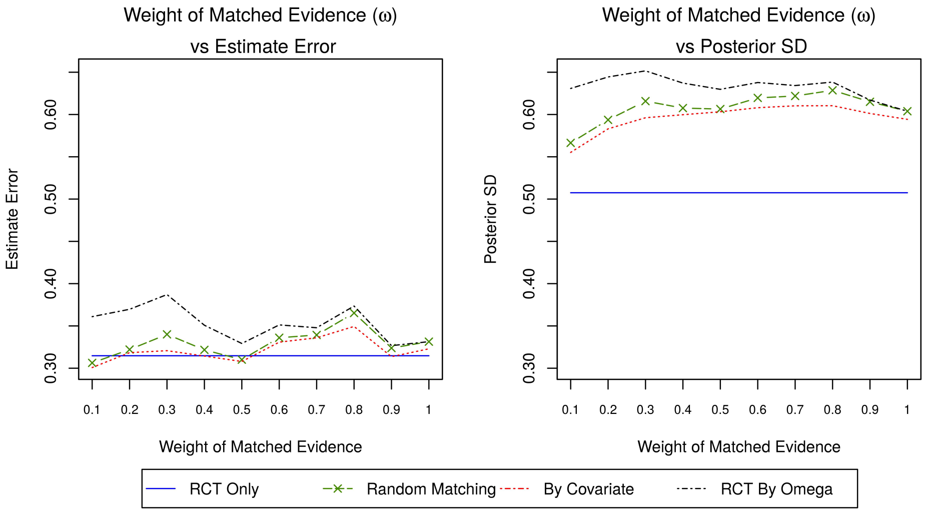


Figure A.5: Effect of ω on the estimate error and posterior SD (original data). Between study variability, σ_μ=0.59, covariate effect, β=-1.04, prior on between study design effect, $\sigma_{des} \sim$ unif(0, 2).

Table A.4: *MC Error for Figure A.5*

	Mean Absolute Error			**Posterior SD**		
ω	**Random**	**By Covariate**	**RCT By Omega**	**Random**	**By Covariate**	**RCT By Omega**
0.1	0.013	0.013	0.014	0.009	0.009	0.01
0.2	0.013	0.012	0.013	0.009	0.009	0.01
0.3	0.017	0.016	0.018	0.011	0.011	0.011
0.4	0.013	0.012	0.013	0.008	0.008	0.009
0.5	0.012	0.012	0.012	0.008	0.009	0.009
0.6	0.013	0.014	0.014	0.008	0.008	0.009
0.7	0.014	0.014	0.014	0.009	0.009	0.009
0.8	0.012	0.012	0.013	0.008	0.008	0.008
0.9	0.015	0.014	0.014	0.010	0.010	0.010
1	0.012	0.011	0.012	0.007	0.007	0.007

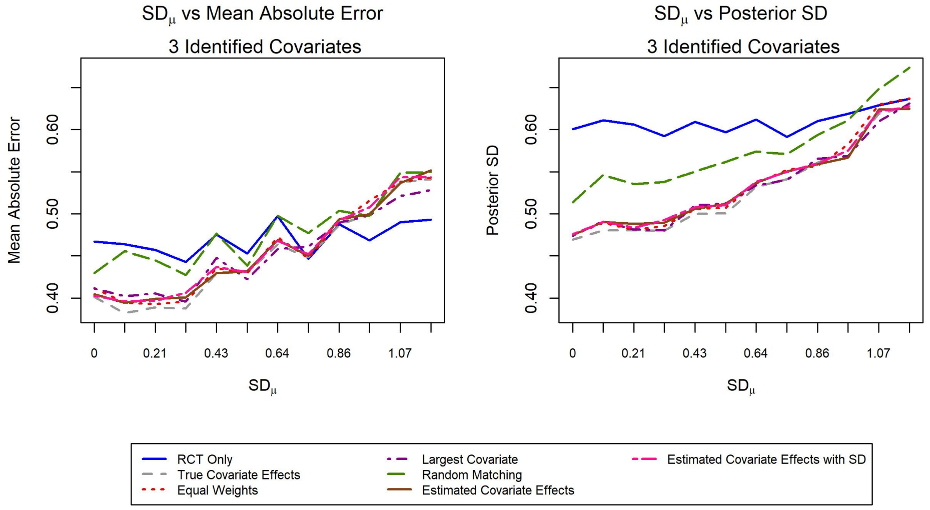


Figure A.6: Effect of the between-study variability, σ_μ, on the estimate error and posterior SD (original data). In this case three covariates have been identified, $\beta = (-0.52, -1.04, -2.08)$.

Table A.5: MC Error for Figure A.6

	Mean Absolute Error							Posterior SD						
σ_μ σ_μ	**RCT Only**	**Random**	**True Covariate**	**Equal Weights**	**Biggest**	**Beta**	**Beta With SD**	**RCT Only**	**Random**	**True Covariate**	**Equal Weights**	**Biggest**	**Beta**	**Beta With SD**
0	0.022	0.018	0.017	0.016	0.017	0.017	0.017	0.012	0.010	0.010	0.009	0.009	0.010	0.009
0.11	0.020	0.018	0.016	0.016	0.016	0.017	0.017	0.011	0.011	0.010	0.009	0.010	0.010	0.010
0.21	0.021	0.018	0.016	0.016	0.016	0.016	0.015	0.011	0.010	0.010	0.010	0.010	0.010	0.010
0.32	0.020	0.017	0.015	0.015	0.016	0.016	0.016	0.011	0.011	0.010	0.010	0.009	0.010	0.010
0.43	0.023	0.019	0.019	0.019	0.018	0.018	0.019	0.011	0.011	0.010	0.010	0.010	0.010	0.010
0.54	0.019	0.017	0.017	0.017	0.016	0.016	0.016	0.011	0.011	0.010	0.009	0.010	0.009	0.009
0.64	0.020	0.018	0.016	0.017	0.017	0.017	0.016	0.011	0.011	0.010	0.010	0.010	0.010	0.010
0.75	0.020	0.019	0.017	0.017	0.017	0.017	0.017	0.011	0.012	0.011	0.011	0.011	0.011	0.011
0.86	0.023	0.021	0.020	0.018	0.018	0.019	0.019	0.011	0.012	0.011	0.010	0.011	0.011	0.011
0.97	0.020	0.020	0.020	0.019	0.020	0.020	0.020	0.011	0.011	0.011	0.010	0.010	0.011	0.011
1.07	0.022	0.020	0.022	0.022	0.020	0.023	0.022	0.011	0.012	0.011	0.010	0.010	0.011	0.011
1.18	0.021	0.020	0.020	0.020	0.019	0.020	0.020	0.011	0.012	0.011	0.011	0.011	0.012	0.012

A.3 Comparing Priors

Table A.6: Comparing Log Odds Ratio (LOR) for each treatment versus PR for the HCV infection RCT Only using a commonly used WinBugs prior (mean, precision) versus the prior purposed in this chapter. The are sizable differences in some of the LORs, but the order of treatment ranking does not change. Smaller numbers were also tested for the precision of the standard prior but this lead to numerical problems in WinBugs.

	N(0, τ=0.001)		**N(0, τ=0.298)**	
	mean	**sd**	**mean**	**sd**
DCV/PR	3.01	1.18	2.15	0.90
BOC/PR	1.13	0.23	1.10	0.23
SIM/PR	1.18	0.22	1.17	0.22
TEL/PR	1.11	0.20	1.06	0.21
SOF/PR	1.90	0.71	1.64	0.65
PrOD±RBV	3.62	0.66	3.24	0.61

A.4 Details of HCV infection Studies

Table A.7: Details of HCV Infection Studies

	Study	**Treatment**	**N**	**SVR**	**G1a**	**Cirrhotic**	**Viral Load >800,000 IU/ml**
RCTs	ADVANCE	PR	361	158	58%	6%	77%
	ADVANCE	TEL/PR	363	271	59%	6%	77%
	PROVE1	PR	75	31	67%	0%	92%
	PROVE1	TEL/PR	79	48	67%	0%	84%
	PROVE1	TEL/PR	79	53	61%	0%	86%
	PROVE2	PR	82	38	43%	0%	83%
	PROVE2	TEL/PR	81	56	38%	0%	90%
	Kumada	PR	63	31	0%	0%	29%
	Kumada	TEL/PR	126	92	2%	0%	21%
	SPRINT1 PART 1	PR	104	39	51%	8%	90%
	SPRINT1 PART 1	Boc/PR	103	58	51%	7%	87%
	SPRINT1 PART 1	Boc	103	77	58%	6%	90%
	Sprint-2	PR	363	137	63%	4%	85%
	Sprint-2	BOC/PR	368	233	64%	4%	91%
	Sprint-2	BOC/PR	366	242	65%	7%	93%
	PILLAR	PR	77	50	38%	0%	82%
	PILLAR	SIM/PR	78	63	46%	0%	82%
	PILLAR	SIM/PR	75	53	45%	0%	84%
	PILLAR	SIM/PR	77	60	48%	0%	90%
	PILLAR	SIM/PR	79	67	47%	0%	91%
	QUEST-1	PR	130	65	57%	13%	74%
	QUEST-1	SIM/PR	264	210	56%	12%	83%
	QUEST-2	PR	134	67	40%	11%	73%
	QUEST-2	SIM/PR	257	209	41%	7%	77%
	PROTON	PR	26	15	77%	0%	Unknown
	PROTON	SOF/PR	47	42	74%	0%	Unknown
	NCT00874770	PR	12	3	58%	0%	Unknown
	NCT00874770	DCV/PR	12	10	75%	0%	Unknown
	MALACHITE-I/II	TEL/PR	75	60	45%	0%	Unknown
	MALACHITE-I/II	PrOD±RBV	236	231	29%	0%	Unknown
	ELECTRON	SOF/RBV	25	21	Unknown	0%	Unknown
	ELECTRON	SOF/LDV±RBV	25	25	80%	0%	Unknown
Single-Arm Trials	ILLUMINATE	TEL/PR	278	216	72%	9%	84%
	NEUTRINO	SOF/PR	292	261	77%	100%	91%
	SAPPHIRE-I	PrOD±RBV	473	455	68%	0%	78%
	ION-1	SOF/LDV±RBV	217	215	66%	17%	79%
	COMMAND-1	DCV/SOF±RBV	15	15	73%	0%	Unknown
	OPTIMIST-1	SIM/SOF±RBV	115	112	Unknown	0%	Unknown
	SPARE	SOF/RBV	25	17	80%	4%	64%

A.5 Winbugs Code

A.5.1 RCT and Pooled Model

```
model{
for(i in 1:ns){
   w[i,1] <- 0
   delta[i,1] <- 0
   mu[i] ~ dnorm(0,0.298)
#chosen to get an rather flat distribution on the probability scale
   for (k in 1:na[i]) {
      r[i,k] ~ dbin(p[i,k],n[i,k])
      logit(p[i,k]) <- mu[i] + delta[i,k]
}

  for (k in 2:na[i]) {
      delta[i,k] ~ dnorm(md[i,k],taud[i,k])
       md[i,k] <- d[t[i,k]] - d[t[i,1]]+ sw[i,k]
  taud[i,k] <- tau*2*(k-1)/k
      w[i,k] <- (delta[i,k] - d[t[i,k]] + d[t[i,1]])
      sw[i,k] <- sum(w[i,1:k-1])/(k-1)
   }
 }
d[1]<-0

for (k in 2:nt){ d[k] ~ dnorm(0,0.298)}
sd ~ dunif(0,2)
tau <- pow(sd,-2)

for (c in 1:(nt-1)) {
for (k in (c+1):nt) {
or[k,c] <- exp(d[k] - d[c])
lor[k,c] <- (d[k]-d[c])
}
}

for (k in 1:nt) {
rk[k] <- nt+1-rank(d[],k)
for (j in 1:nt){
best[j,k] <- equals(rk[k],j)}
}
} #end
```

The following is the data for the RCT only dataset

```
list(
ns=12,
nt=7,
na=c(2, 3, 2, 2, 3, 3, 5, 2, 2, 2, 2, 2),
n=structure(.Data= c(361, 363, NA, NA, NA,
75, 79, 79, NA, NA,
82, 81, NA, NA, NA,
63, 126, NA, NA, NA,
104, 103,  103, NA, NA,
363, 368, 368, NA, NA,
77, 78, 75, 77, 79,
130, 264, NA, NA, NA,
134, 257, NA, NA, NA,
26, 47, NA, NA, NA,
12, 12, NA, NA, NA,
75, 236, NA, NA, NA
), .Dim=c(12, 5)),
r=structure(.Data= c(158,271, NA, NA, NA,
31, 48, 53, NA, NA,
38, 56, NA, NA, NA,
31,92, NA, NA, NA,
39, 58, 77, NA, NA,
137, 233, 242, NA, NA,
50, 63, 53, 60, 67,
65, 210,  NA, NA, NA,
67, 209,  NA, NA, NA,
15, 42, NA, NA, NA,
3, 10, NA, NA, NA,
60, 231, NA, NA, NA
), .Dim=c(12, 5)),
t=structure(.Data= c(1, 5, NA, NA, NA,
1, 5, 5, NA, NA,
1, 5, NA, NA, NA,
1, 5, NA, NA, NA,
1, 3,  3, NA, NA,
1, 3, 3, NA, NA,
1, 4, 4, 4, 4,
1, 4, NA, NA, NA,
1, 4, NA, NA, NA,
1, 6, NA, NA, NA,
1, 2, NA, NA, NA,
5, 7, NA, NA, NA
), .Dim=c(12, 5))
)
```

The following includes the pooled dataset. The final 7 lines of each matrix refer to the matched arms. The two study types are indistinguishable in the model. Note that there is a 13th RCT which is not included in the RCT only dataset. This is because this RCT is not connected to the main network unless the matched evidence is also included.

```
list(
ns=20,
nt=11,
na=c(2, 3, 2, 2, 3, 3, 5, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2),
n=structure(.Data= c(
361, 363, NA, NA, NA,
75, 79, 79, NA, NA,
82, 81, NA, NA, NA,
63, 126, NA, NA, NA,
104, 103,  103, NA, NA,
363, 368, 368, NA, NA,
77, 78, 75, 77, 79,
130, 264, NA, NA, NA,
134, 257, NA, NA, NA,
26, 47, NA, NA, NA,
12, 12, NA, NA, NA,
75, 236, NA, NA, NA,
25, 25, NA, NA, NA,
278, 363, NA, NA, NA, #Matched (second column) starts here
292, 278, NA, NA, NA,
473, 79, NA, NA, NA,
217, 130, NA, NA, NA,
15, 47, NA, NA, NA,
115, 75, NA, NA, NA,
25, 473, NA, NA, NA
), .Dim=c(20, 5)),
r=structure(.Data= c(
158,271, NA, NA, NA,
31, 48, 53, NA,NA,
38, 56, NA, NA,NA,
31, 92, NA, NA,NA,
39, 58, 77, NA,NA,
137, 233, 242, NA, NA,
50, 63, 53, 60, 67,
65, 210, NA, NA,NA,
67, 209, NA, NA, NA,
15, 42, NA, NA, NA,
3, 10, NA, NA, NA,
60, 231, NA, NA, NA,
21, 25, NA, NA, NA,
216, 137, NA, NA, NA, #Matched (second column) starts here
261, 216, NA, NA, NA,
455, 48, NA, NA, NA,
215, 65, NA, NA, NA,
15, 42, NA, NA, NA,
112, 60, NA, NA, NA,
17, 455, NA, NA, NA
), .Dim=c(20, 5)))
t=structure(.Data= c(
1, 5, NA, NA, NA,
1, 5, 5, NA, NA,
1, 5, NA, NA, NA,
1, 5, NA, NA, NA,
1, 3,  3, NA, NA,
1, 3, 3, NA, NA,
1, 4, 4, 4, 4,
1, 4, NA, NA, NA,
1, 4, NA, NA, NA,
1, 6, NA, NA, NA,
1, 2, NA, NA, NA,
5, 7, NA, NA, NA,
11, 8, NA, NA, NA,
5, 1, NA, NA, NA, #Matched (second column) starts here
6, 5, NA, NA, NA,
7, 5, NA, NA, NA,
8, 1, NA, NA, NA,
9, 6, NA, NA, NA,
10, 5, NA, NA, NA,
11, 7, NA, NA, NA
), .Dim=c(20, 5)),
```

A.5.2 Hierarchical Model

```
model{
### Overall Level ###
d[1]<-0
RCTd[1]<-0
MATCHEDd[1]<-0

for (k in 2:nt){
RCTd[k]~dnorm(d[k],COtauRCT)
MATCHEDd[k]~dnorm(d[k],COtauMATCHED)
d[k]~dnorm(0,0.298)}

for (k in 1:nt){
```

```
d_branch[k, 3]<-RCTd[k]
d_branch[k, 1]<-MATCHEDd[k]
d_branch[k, 2]<-d[k]
Used_d[k]<-d_branch[k, RCTorMatched[k]] }

COsd~dunif(0,SDUpper)
COvar<-pow(COsd,2)
COtauMATCHED<-Omega/COvar
COtauRCT<-1/COvar

for (c in 1:(nt-1)) {
for (k in (c+1):nt) {
or[k,c] <- exp(Used_d[k] - Used_d[c])
lor[k,c] <- (Used_d[k]-Used_d[c])
}
}

### RCTs###
for(i in 1:RCTns){
   RCTw[i,1] <- 0
   RCTdelta[i,1] <- 0
   RCTmu[i] ~ dnorm(0,0.298)
   for (k in 1:RCTna[i]) {
      RCTr[i,k] ~ dbin(RCTp[i,k],RCTn[i,k])
      logit(RCTp[i,k]) <- RCTmu[i] + RCTdelta[i,k]
}
  for (k in 2:RCTna[i]) {
      RCTdelta[i,k] ~ dnorm(RCTmd[i,k],RCTtaud[i,k])
      RCTmd[i,k] <- RCTd[RCTt[i,k]] - RCTd[RCTt[i,1]]+ RCTsw[i,k]
RCTtaud[i,k] <- RCTtau*2*(k-1)/k
      RCTw[i,k] <- (RCTdelta[i,k] - RCTd[RCTt[i,k]] + RCTd[RCTt[i,1]])
      RCTsw[i,k] <- sum(RCTw[i,1:k-1])/(k-1)
   }
 }
RCTsd ~ dunif(0,2)
RCTtau <- pow(RCTsd,-2)

### MATCHED###
for(i in 1:MATCHEDns){
   MATCHEDw[i,1] <- 0
   MATCHEDdelta[i,1] <- 0
   MATCHEDmu[i] ~ dnorm(0,0.298)
   for (k in 1:MATCHEDna[i]) {
      MATCHEDr[i,k] ~ dbin(MATCHEDp[i,k],MATCHEDn[i,k])
      logit(MATCHEDp[i,k]) <- MATCHEDmu[i] + MATCHEDdelta[i,k]
}

  for (k in 2:MATCHEDna[i]) {
      MATCHEDdelta[i,k] ~ dnorm(MATCHEDmd[i,k],MATCHEDtaud[i,k])
      MATCHEDmd[i,k] <- MATCHEDd[MATCHEDt[i,k]] - MATCHEDd[MATCHEDt[i,1]]+ MATCHEDsw[i,k]
      MATCHEDtaud[i,k] <- MATCHEDtau*2*(k-1)/k
      MATCHEDw[i,k] <- (MATCHEDdelta[i,k] - MATCHEDd[MATCHEDt[i,k]] + MATCHEDd[MATCHEDt[i,1]])
      MATCHEDsw[i,k] <- sum(MATCHEDw[i,1:k-1])/(k-1)
   }
 }
MATCHEDsd ~ dunif(0,2)
MATCHEDtau <- pow(MATCHEDsd,-2)
} #end
```

The following shows the dataset for the hierarchical model. In this case the RCT and matched evidence are treated as different study types.

```
list(
RCTns=12,
RCTna=c(2, 3, 2, 2, 3, 3, 5, 2, 2, 2, 2, 2),
RCTn=structure(.Data= c(
361, 363, NA, NA, NA,
75, 79, 79, NA, NA,
82, 81, NA, NA, NA,
63, 126, NA, NA, NA,
104, 103,  103, NA, NA,
363, 368, 368, NA, NA,
77, 78, 75, 77, 79,
130, 264, NA, NA, NA,
134, 257, NA, NA, NA,
26, 47, NA, NA, NA,
12, 12, NA, NA, NA,
75, 236, NA, NA, NA
), .Dim=c(12, 5)),
RCTt=structure(.Data= c(
1, 5, NA, NA, NA,
1, 5, 5, NA, NA,
1, 5, NA, NA, NA,
1, 5, NA, NA, NA,
1, 3,  3, NA, NA,
1, 3, 3, NA, NA,
1, 4, 4, 4, 4,
1, 4, NA, NA, NA,
1, 4, NA, NA, NA,
1, 6, NA, NA, NA,
1, 2, NA, NA, NA,
5, 7, NA, NA, NA
), .Dim=c(12, 5)),
RCTr=structure(.Data= c(
158,271, NA, NA, NA,
31, 48, 53, NA, NA,
38, 56, NA, NA, NA,
31,92, NA, NA, NA,
39, 58, 77, NA, NA,
137, 233, 242, NA, NA,
50, 63, 53, 60, 67,
65, 210,  NA, NA, NA,
67, 209,  NA, NA, NA,
15, 42, NA, NA, NA,
3, 10, NA, NA, NA,
60, 231, NA, NA, NA
), .Dim=c(12, 5)),
MATCHEDns=7,
nt=11,
MATCHEDna=c(2, 2, 2, 2, 2, 2, 2),
MATCHEDn=structure(.Data= c(
278, 363,
292, 278,
473, 79,
217, 130,
15, 47,
115, 75,
25, 473
), .Dim=c(7, 2)),
MATCHEDt=structure(.Data= c(
5, 1,
6, 5,
7, 5,
8, 1,
9, 6,
10, 5,
11, 7
), .Dim=c(7, 2)),
MATCHEDr=structure(.Data= c(
216, 137,
261, 216, ,
455, 48,
215, 65,
15, 42,
112, 60,
17, 455
), .Dim=c(7, 2)),
RCTorMatched=c(2, 3, 3, 3, 2, 2, 2, 1, 1, 1, 1),
Omega=0.5,
SDUpper=2)
```

A.5.3 Plug in Estimator Model

```
model{
# Model for RCTs #
for(i in 1:ns){
w[i,1] <- 0
mu[i] ~ dnorm(0,0.298)
for (k in 1:na[i]) {
r[i,k] ~ dbin(p[i,k],n[i,k])
logit(p[i,k]) <- mu[i] + delta[i,k]
}
delta[i,1] <- d[t[i,1]]

for (k in 2:na[i]) {
delta[i,k] ~ dnorm(md[i,k],taud[i,k])
md[i,k] <- d[t[i,k]] + sw[i,k]
taud[i,k] <- tau *2*(k-1)/k
w[i,k] <- (delta[i,k] - d[t[i,k]])
sw[i,k] <- sum(w[i,1:k-1])/(k-1)
}
}

# Model for Matched Studies #
for(i in 1:MATns){
       MATr[i] ~ dbin(MATp[i],MATn[i])
       logit(MATp[i]) <- mu.cut[i] + MATdelta[i]
mu.cut[i] <- cut(mu[ChosenRCT[i]])
MATdelta[i]~dnorm(d[MATt[i]],tau.cut)
}

# Specify remaining priors
d[1]<-0
for (k in 2:nt){ d[k] ~ dnorm(0,0.298) }
sd ~ dunif(0,2)
tau <- pow(sd,-2)
tau.cut<-cut(tau)

# pairwise ORs and LORs for all possible pair-wise comparisons, if nt>2
for (c in 1:(nt-1)) {
for (k in (c+1):nt) {
or[k,c] <- exp(d[k] - d[c])
lor[k,c] <- (d[k]-d[c])
}
}

  # ranking on relative scale
  for (k in 1:nt) {
    rk[k] <- nt+1-rank(d[],k)
    for (j in 1:nt){
    best[j,k] <- equals(rk[k],j)}
  }

}
```

The following shows the dataset for the Plug in estimator model.

```
list(
ns=12,
nt=11,
na=c(2, 3, 2, 2, 3, 3, 5, 2, 2, 2, 2, 2),
n=structure(.Data= c(
361, 363, NA, NA, NA,
75, 79, 79, NA, NA,
82, 81, NA, NA, NA,
63, 126, NA, NA, NA,
104, 103,  103, NA, NA,
363, 368, 368, NA, NA,
77, 78, 75, 77, 79,
130, 264, NA, NA, NA,
134, 257, NA, NA, NA,
26, 47, NA, NA, NA,
12, 12, NA, NA, NA,
75, 236, NA, NA, NA
), .Dim=c(12, 5)),
t=structure(.Data= c(
1, 5, NA, NA, NA,
1, 5, 5, NA, NA,
1, 5, NA, NA, NA,
1, 5, NA, NA, NA,
1, 3,  3, NA, NA,
1, 3, 3, NA, NA,
1, 4, 4, 4, 4,
1, 4, NA, NA, NA,
1, 4, NA, NA, NA,
1, 6, NA, NA, NA,
1, 2, NA, NA, NA,
5, 7, NA, NA, NA
), .Dim=c(12, 5)),
r=structure(.Data= c(
158,271, NA, NA, NA,
31, 48, 53, NA, NA,
38, 56, NA, NA, NA,
31,92, NA, NA, NA,
39, 58, 77, NA, NA,
137, 233, 242, NA, NA,
50, 63, 53, 60, 67,
65, 210,  NA, NA, NA,
67, 209,  NA, NA, NA,
15, 42, NA, NA, NA,
3, 10, NA, NA, NA,
60, 231, NA, NA, NA
), .Dim=c(12, 5)),
MATCHEDns=7,
MATCHEDn=c(278, 292, 473, 217, 15, 115, 25),
MATCHEDt=c(5, 6, 7, 8, 9, 10, 11),
MATCHEDr=c(216, 261, 455, 215, 15, 112, 17),
ChosenRCT=c(10, 6, 11, 8, 10, 7, 2))
```

A.5.4 Model For Estmaiting σ_μ

In order to estimate σ_μ we use the RCT part of the plug-in estimator model. This is because it is neccessary to have a common reference treatment so that the study effects, μ_i are comparable across studies.

```
model{
for(i in 1:ns){
w[i,1] <- 0
mu[i] ~ dnorm(0,0.298)
for (k in 1:na[i]) {
r[i,k] ~ dbin(p[i,k],n[i,k])
logit(p[i,k]) <- mu[i] + delta[i,k]
}
delta[i,1] <- d[t[i,1]]

for (k in 2:na[i]) {
delta[i,k] ~ dnorm(md[i,k],taud[i,k])
md[i,k] <- d[t[i,k]] + sw[i,k]
taud[i,k] <- tau *2*(k-1)/k
w[i,k] <- (delta[i,k] - d[t[i,k]])
sw[i,k] <- sum(w[i,1:k-1])/(k-1)
}
}

d[1]<-0
for (k in 2:nt){ d[k] ~ dnorm(0,0.298) }
sd ~ dunif(0,2)
tau <- pow(sd,-2)
tau.cut<-cut(tau)

}
```

This is simply the RCT only dataset.

```
list(
ns=12,
nt=11,
na=c(2, 3, 2, 2, 3, 3, 5, 2, 2, 2, 2, 2),
n=structure(.Data= c(
361, 363, NA, NA, NA,
75, 79, 79, NA, NA,
82, 81, NA, NA, NA,
63, 126, NA, NA, NA,
104, 103,  103, NA, NA,
363, 368, 368, NA, NA,
77, 78, 75, 77, 79,
130, 264, NA, NA, NA,
134, 257, NA, NA, NA,
26, 47, NA, NA, NA,
12, 12, NA, NA, NA,
75, 236, NA, NA, NA
), .Dim=c(12, 5)),
t=structure(.Data= c(
1, 5, NA, NA, NA,
1, 5, 5, NA, NA,
1, 5, NA, NA, NA,
1, 5, NA, NA, NA,
1, 3,  3, NA, NA,
1, 3, 3, NA, NA,
1, 4, 4, 4, 4,
1, 4, NA, NA, NA,
1, 4, NA, NA, NA,
1, 6, NA, NA, NA,
1, 2, NA, NA, NA,
5, 7, NA, NA, NA
), .Dim=c(12, 5)),
r=structure(.Data= c(
158,271, NA, NA, NA,
31, 48, 53, NA, NA,
38, 56, NA, NA, NA,
31,92, NA, NA, NA,
39, 58, 77, NA, NA,
137, 233, 242, NA, NA,
50, 63, 53, 60, 67,
65, 210,  NA, NA, NA,
67, 209,  NA, NA, NA,
15, 42, NA, NA, NA,
3, 10, NA, NA, NA,
60, 231, NA, NA, NA
), .Dim=c(12, 5)))
```

Appendix B

For Chapter 4 (Individual Patient Data)

Table B.1: Number of simulations for each scenario. The differences in the number of runs is due to the computing power available and the relative length of time required for each scenario.

True Identical	**True Exchangeable**
RCT Binary Covariate	
248	314
Observational Binary Covariate	
241	301
RCT Continuous Covariate	
138	153
Observational Continuous Covariate	
122	128

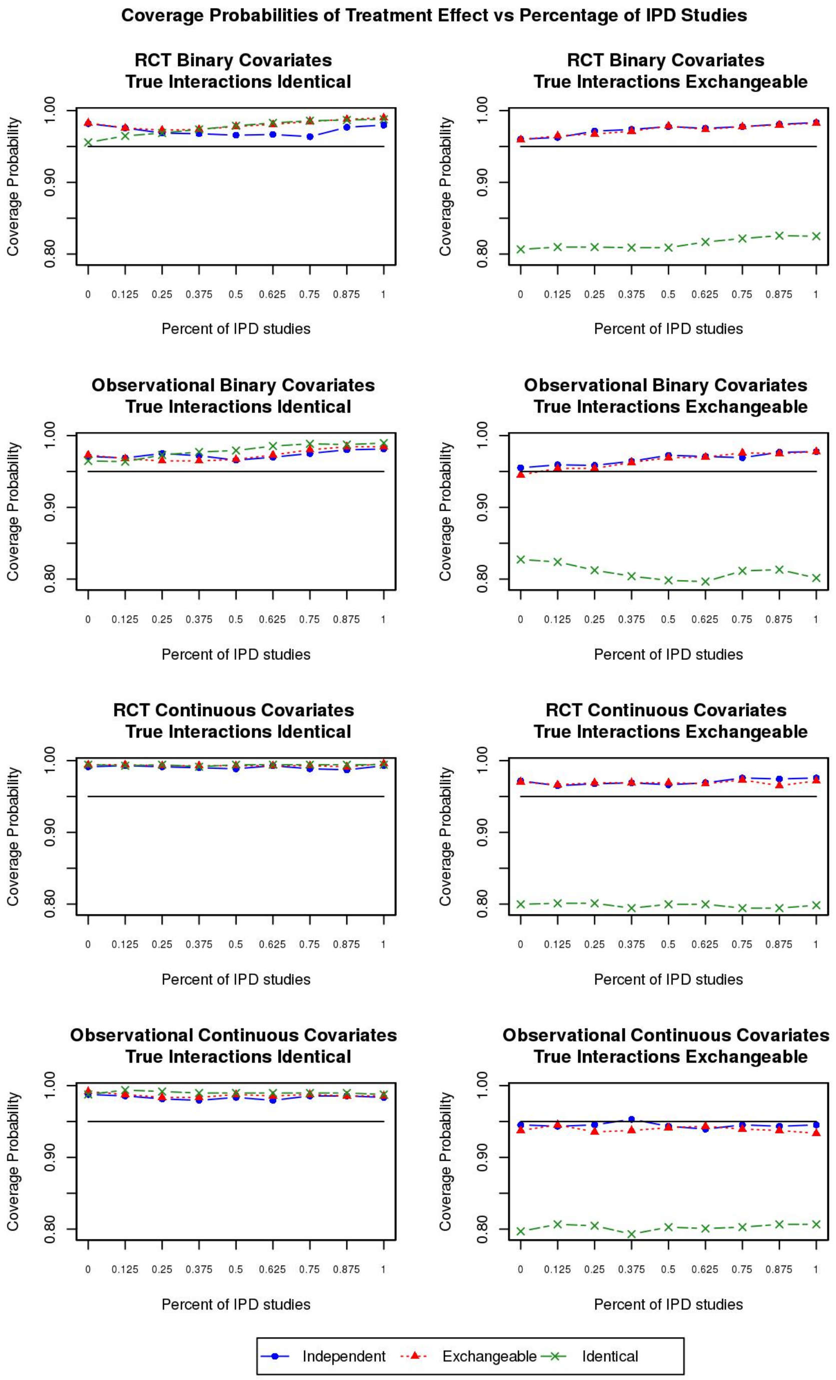


Figure B.1: Coverage probabilities of the 95% credible interval for treatment Effect vs percentage of individual patient data (IPD) studies. Coverage above the solid black 95% line often indicates that posterior standard deviations are too conservative, while coverage below this line often indicates that posterior standard deviations are too precise.

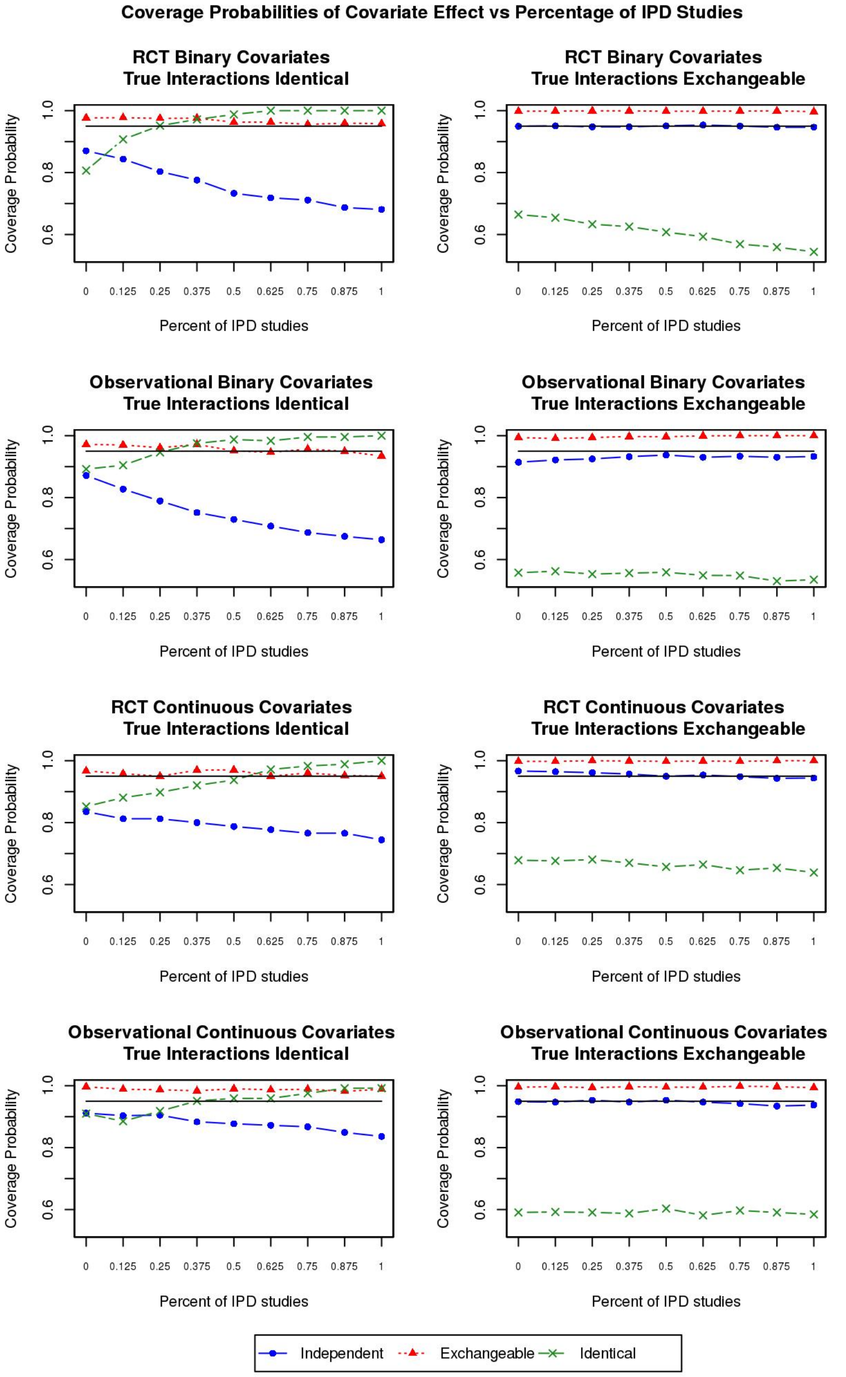


Figure B.2: Coverage probabilities of the 95% credible interval for covariate effect vs percentage of individual patient data (IPD) studies. Coverage above the solid black 95% line often indicates that posterior standard deviations are too conservative, while coverage below this line often indicates that posterior standard deviations are too precise. IPD can cause over confidence when the incorrect model is chosen.

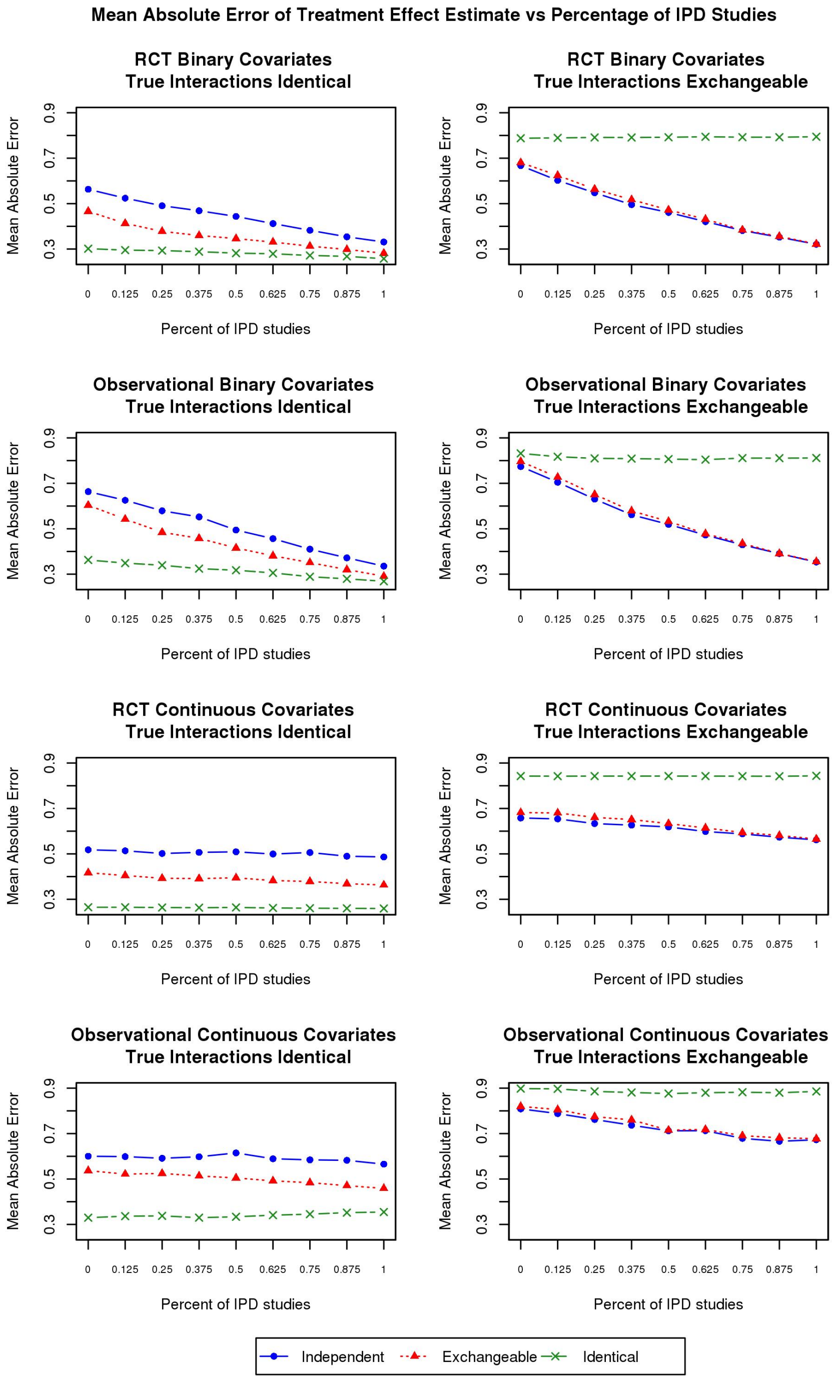


Figure B.3: Mean absolute error for the estimate of the treatment effect vs percentage of individual patient data (IPD) studies. IPD has a particularly large effect in the scenarios with the binary covariate.

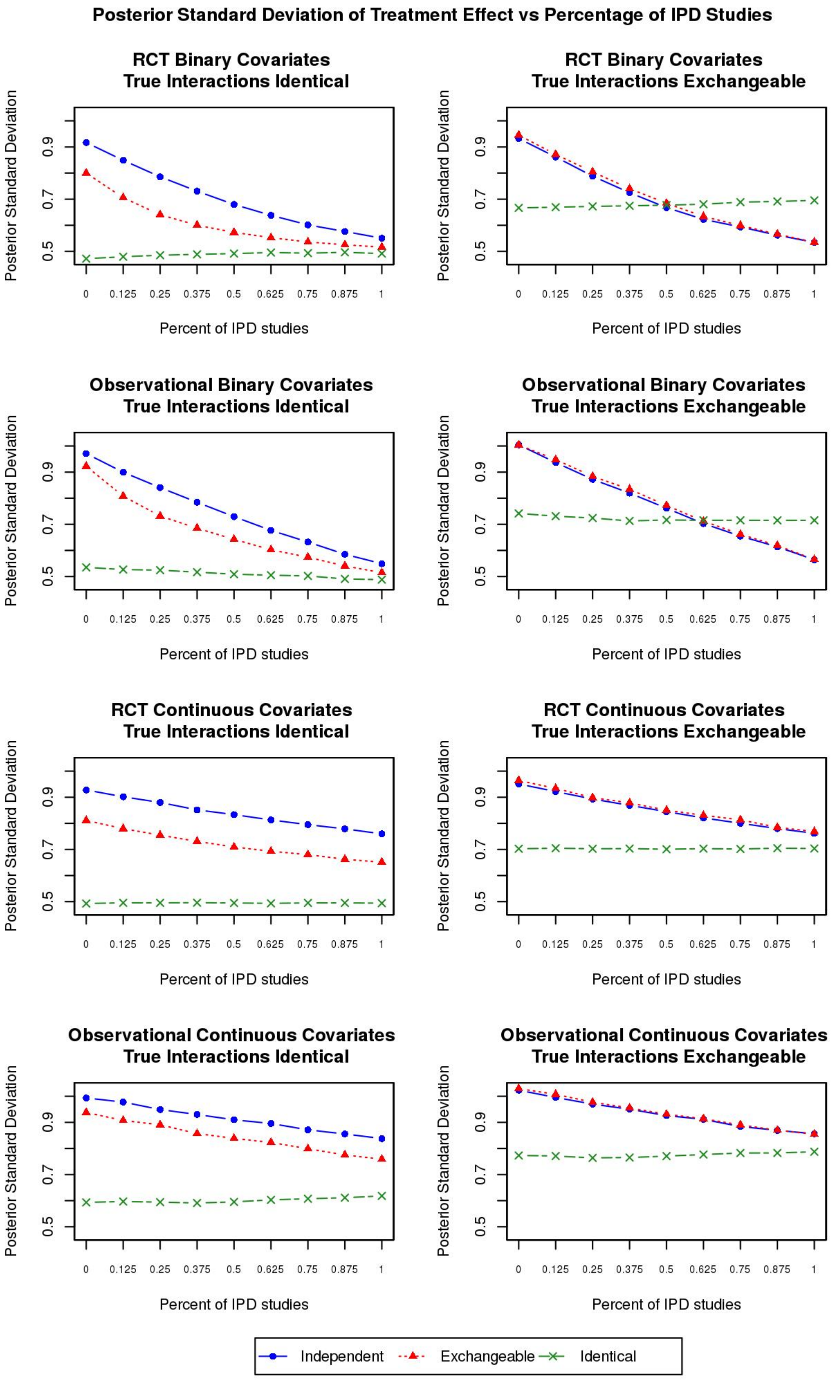


Figure B.4: Posterior standard deviation (SD) for treatment effect vs percentage of individual patient data (IPD) studies. IPD decreases the posterior SD for the independent and exchangeable (effect modifiers) models.

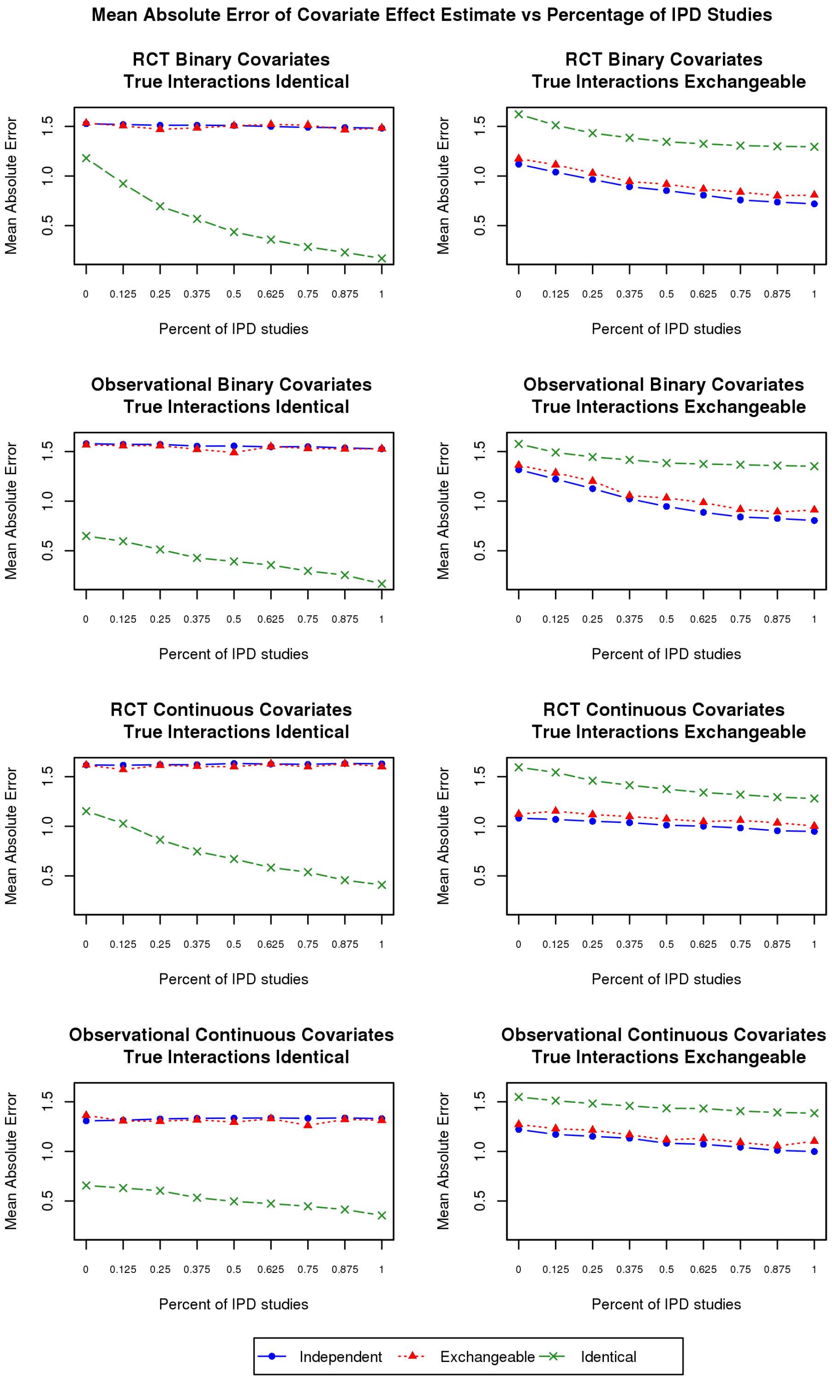


Figure B.5: Mean absolute error (MAE) for the estimate of the covariate effect vs percentage of individual patient data (IPD) studies. As the amount of IPD increases, the MAE of the estimate decreases for a number of models.

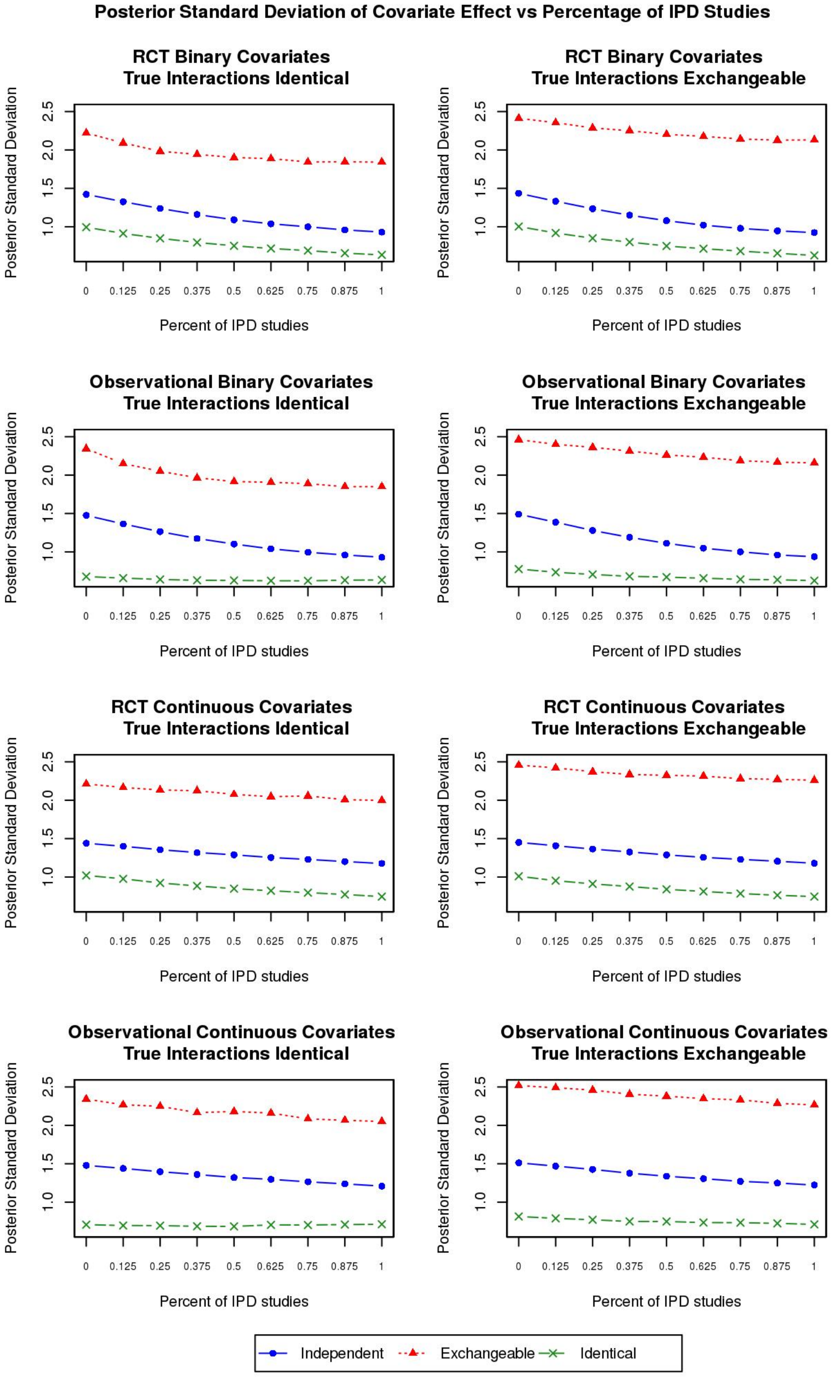


Figure B.6: Posterior standard deviation for covariate effect vs percentage of individual patient data (IPD) studies. The effect of IPD is not as noticeable in this case but can still be seen in some scenarios.

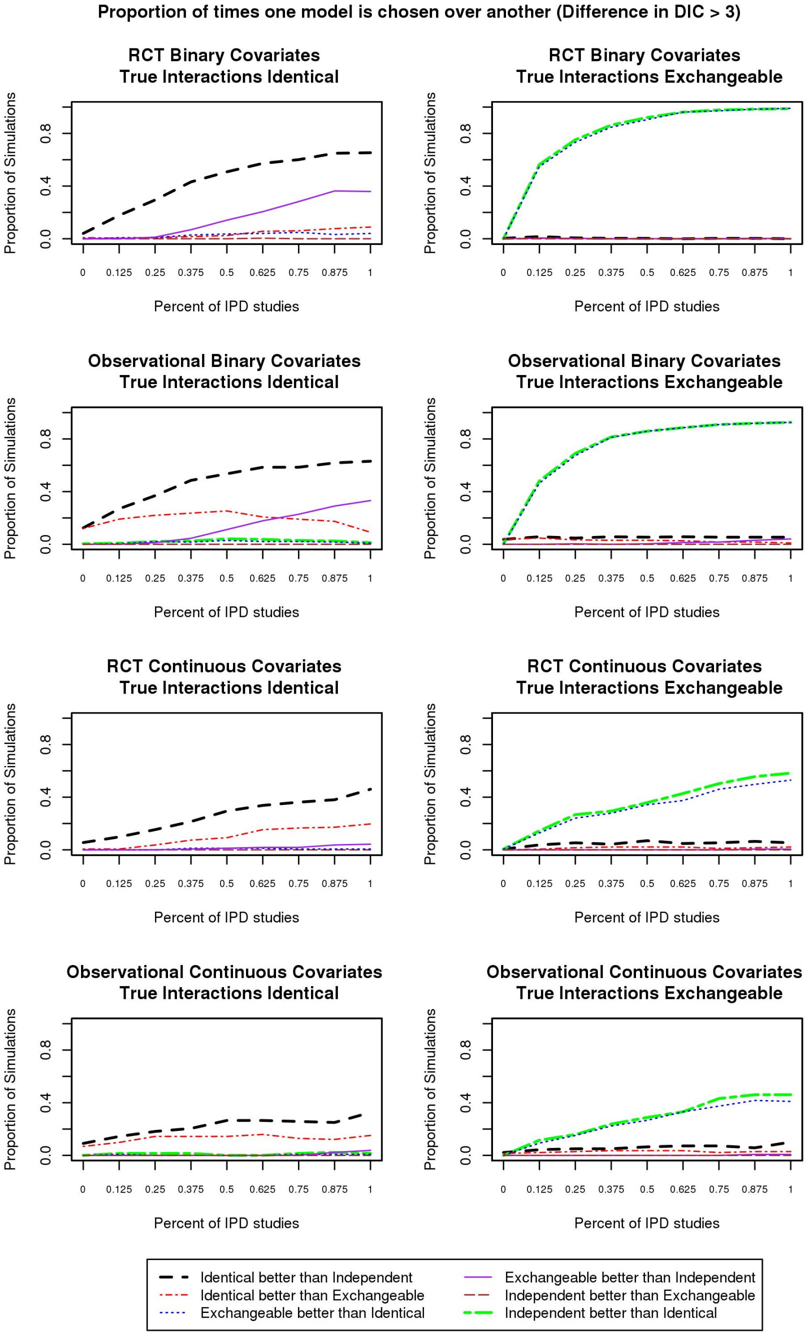


Figure B.7: Proportion of Deviance Information Criterion (DIC) differences greater than 3 vs percentage of individual patient data (IPD) studies. The lines track the number of iterations when there is a meaningful difference between two models. There are seldom differences between the models with a full aggregate dataset. However, as the amount of IPD increases the correct model is identified more often (up to 100% of the time in the case of an exchangeable model). The increase due to IPD is much larger in the scenarios with the binary covariate.

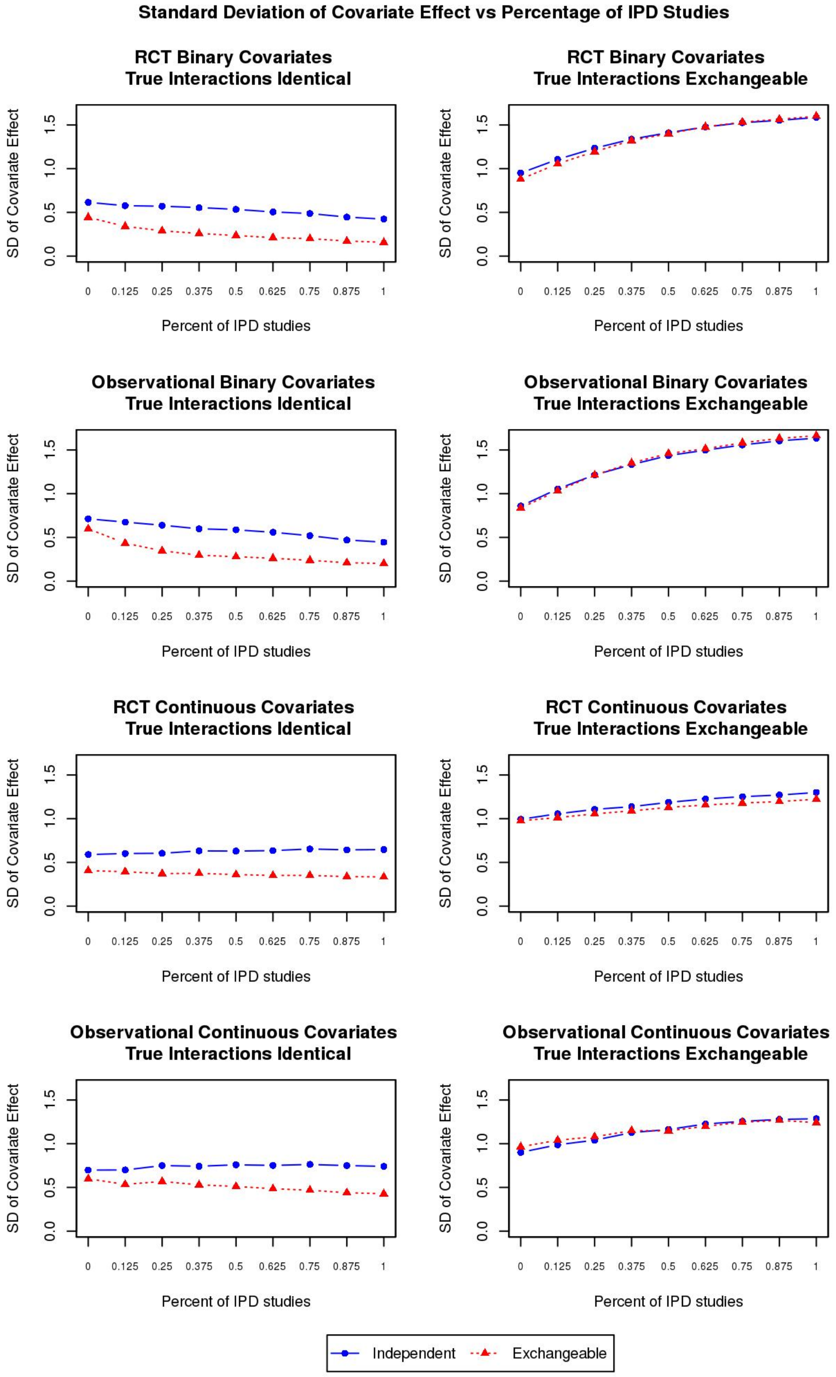


Figure B.8: Standard deviation of the interaction of the covariate with the five treatments for the independent and exchangeable models vs percentage of individual patient data (IPD) studies. Exchangeable and independent models will estimate the covariate effects to be more different from each other when they are actually exchangeable as opposed to truly identical.

Appendix C

For Chapter 5 (Matching Adjusted Indirect Comparison)

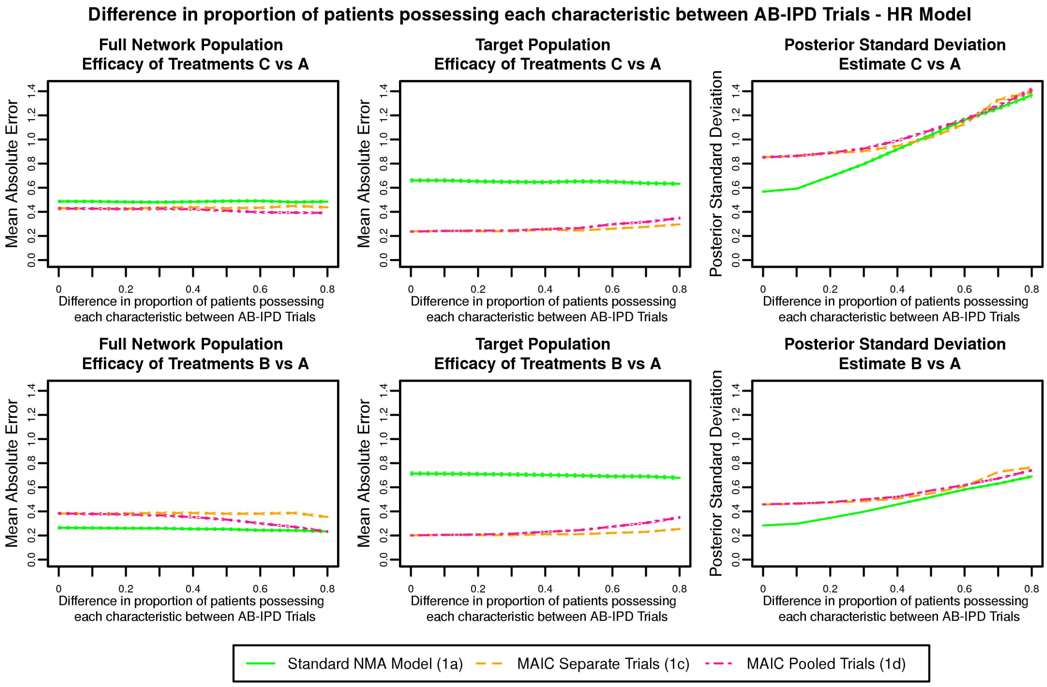


Figure C.1: Examining MAE and posterior SD while varying the difference in proportion possessing the characteristic associated with each covariate between AB-IPD trials. The dotted lines around each MAE estimate represent the MC Error on each side. On the left most point of the x-axis all three AB-IPD studies have 45% of patients possessing the characteristic associated with each covariate, however, on the right hand side Study 1 has 45% of patients possessing the characteristic associated with each covariate, Study 2 has 90% of patients possessing the characteristic associated with each covariate, and Study 3 has 10% of patients possessing the characteristic associated with each covariate. The numbers on the x-axis represent the difference in the proportion possessing the characteristic associated with each covariate between Study 2 and Study 3. The AB-AgD study has a fixed proportion possessing the characteristic associated with each covariate of 90%.

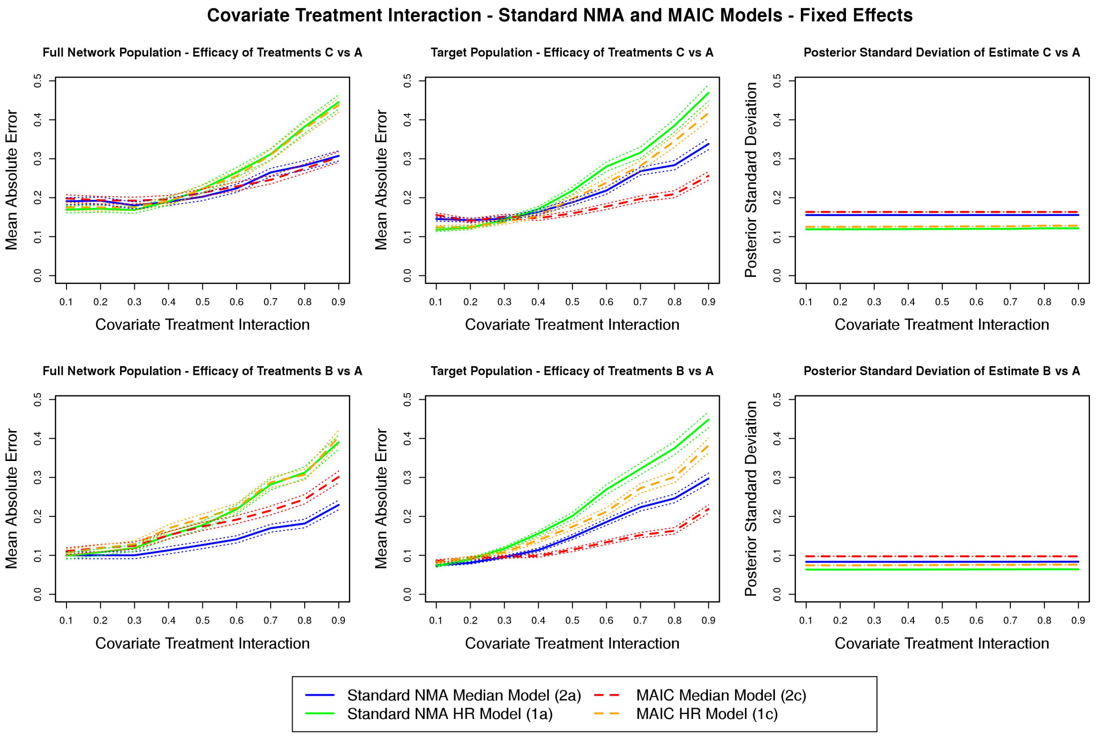


Figure C.2: Fixed effects models: Examining MAE and posterior SD while increasing the standard deviation of the covariate-treatment interaction: Standard NMA and standard NMA with covariate models

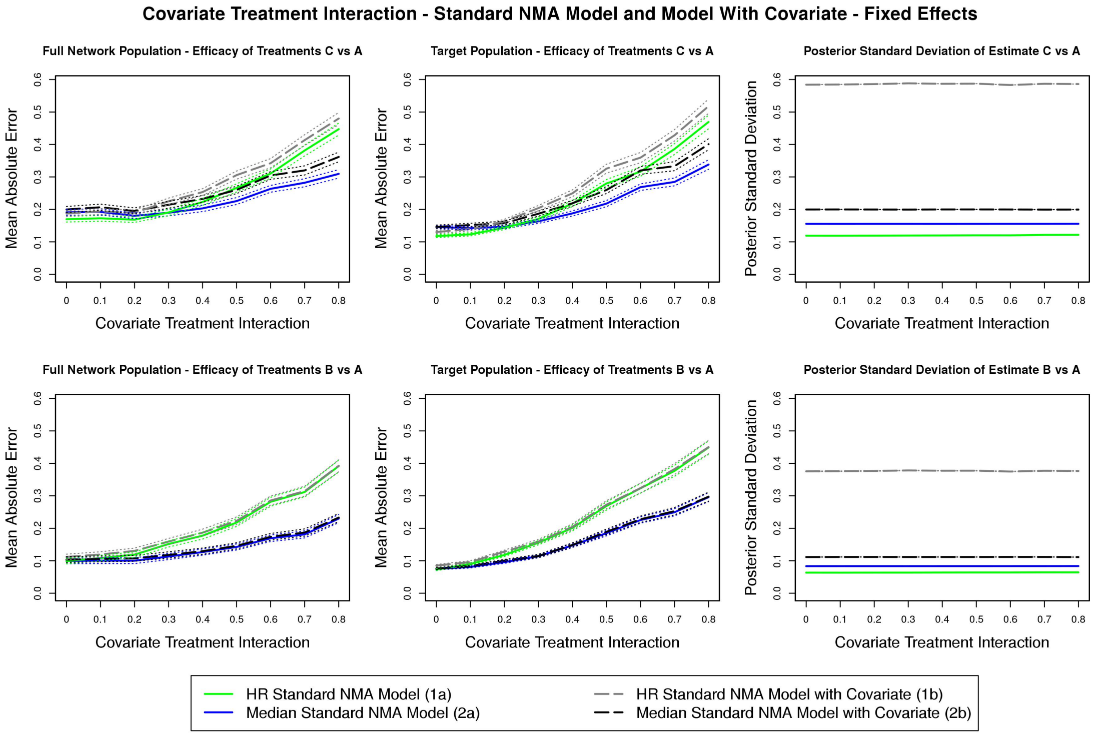


Figure C.3: Fixed effects models: Examining MAE and posterior SD while increasing the standard deviation of the covariate-treatment interaction: Standard NMA, MAIC, and mixed AgD/IPD models

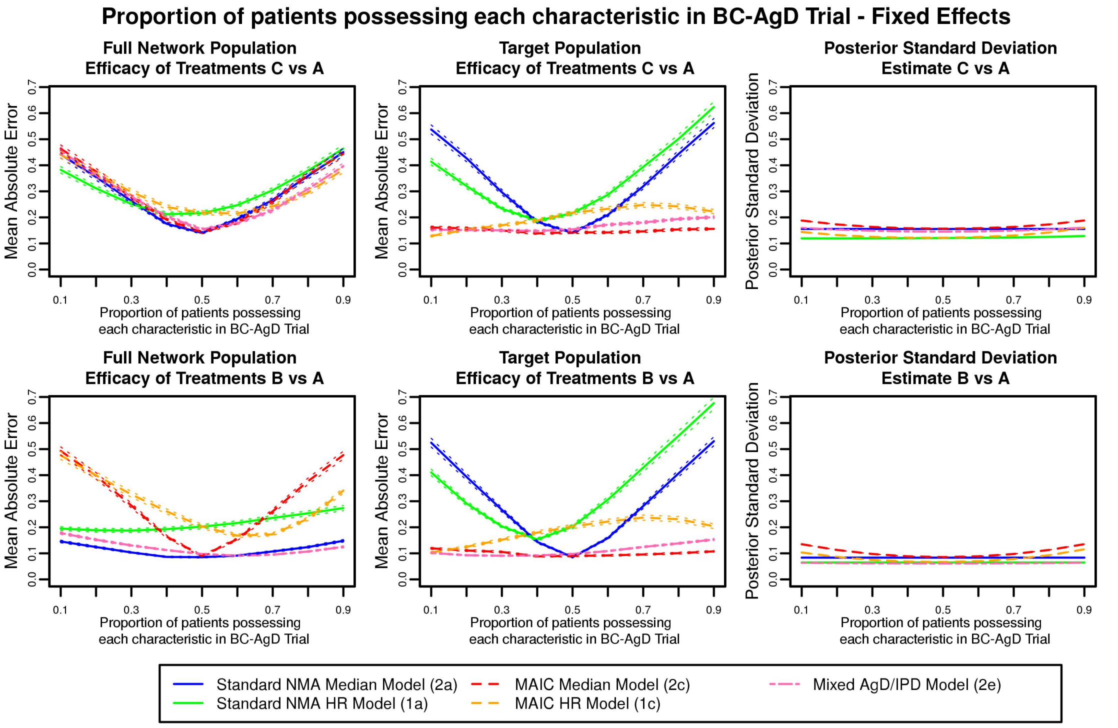


Figure C.4: Fixed effects models: Examining MAE and posterior SD while varying the proportion of patients possessing each characteristic in the BC-AgD study

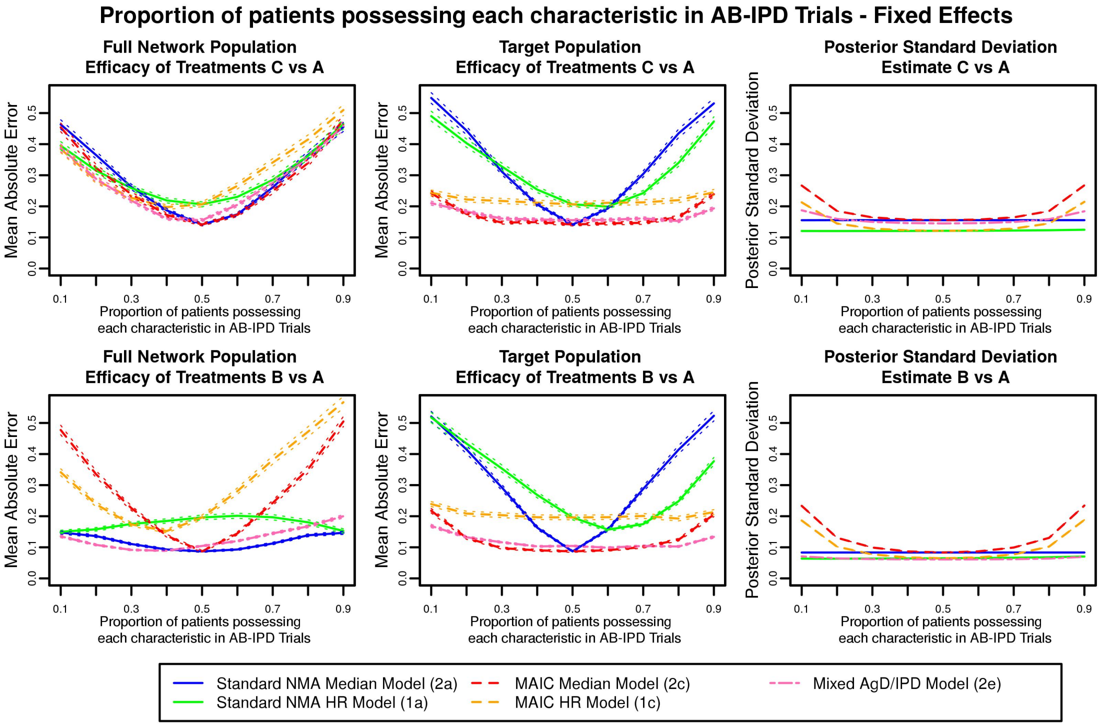


Figure C.5: Fixed effects models: Examining MAE and posterior SD while varying the proportion of patients possessing each characteristic in the AB-IPD study

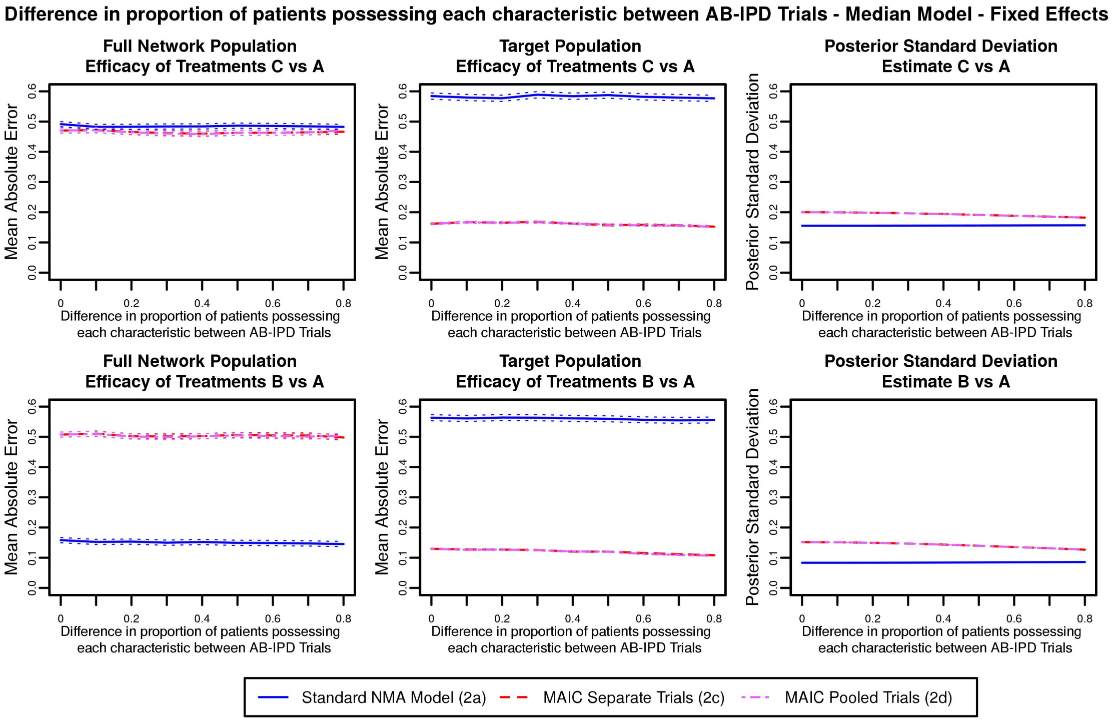


Figure C.6: Fixed effects models: Examining MAE and posterior SD while varying the difference in proportion possessing the characteristic associated with each covariate between AB-IPD trials

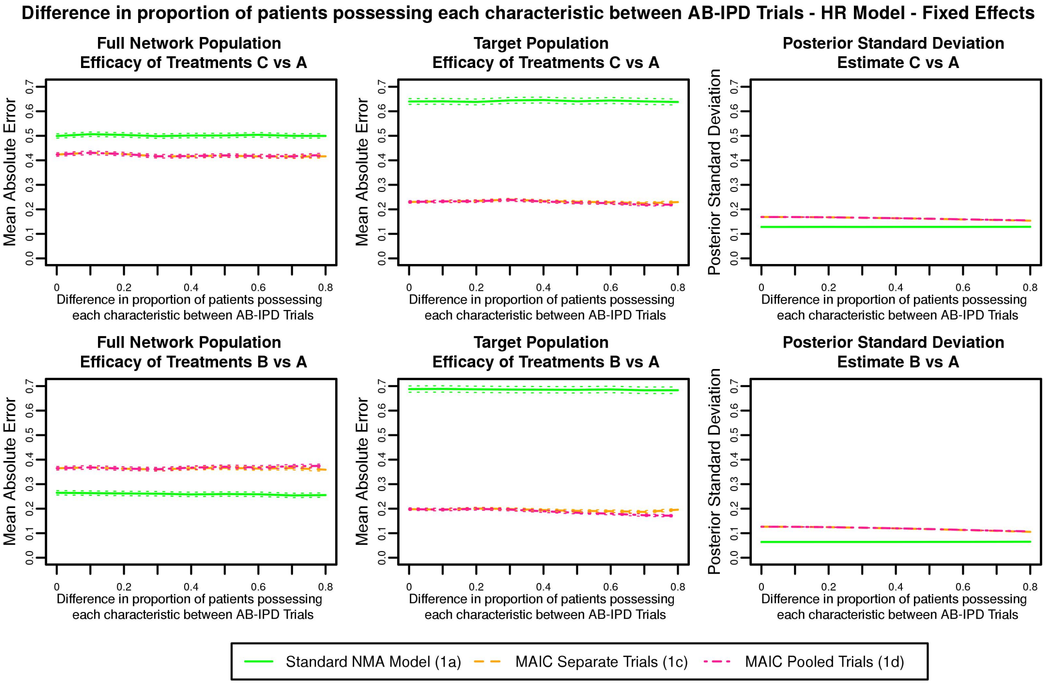


Figure C.7: Fixed effects models: Examining MAE and posterior SD while varying the difference in proportion possessing the characteristic associated with each covariate between AB-IPD trials