Bayesian evidence synthesis for surrogate endpoints in the era of precision medicine

by

Anastasios Papanikos

Department of Health Sciences

University of Leicester

Thesis submitted for the degree of

Doctor *of* Philosophy

2020

Bayesian evidence synthesis for surrogate endpoints in the era of precision medicine

by

Anastasios Papanikos

This thesis considers a range of methodological challenges related to the trial-level validation of surrogate endpoints in disease areas where precision medicine have played an important role, and aims to addressed them by proposing novel statistical methodology.

Firstly, the thesis introduces two hierarchical meta-analytic methods which allow for modeling differences in trial-level surrogacy patterns. Trial-level surrogacy patterns may vary across treatment classes due to, for example, the diversity of the mechanisms of action of targeted therapies. A simple way to examine potential differences in surrogacy patterns across treatment classes is by performing subgroup analysis using a bivariate meta-analytic method. However, this approach fails to estimate trial-level association patterns effectively when data are limited in terms of the number of studies. The two hierarchical meta-analytic methods aim to improve the inference about the parameters describing the surrogacy patterns within a treatment class as they borrow information for these parameters across classes.

Secondly, the thesis proposes a new method which is appropriate for modeling correlated binomial aggregate data with very rare or frequent events. Targeted treatments are usually much more successful compared to standard of care resulting in very high numbers of treatment responses and reduced numbers of events. When standard approaches for trial-level validation of surrogate endpoints are applied to such binomial data, they may lead to poor inferences about surrogacy patterns due to inappropriate assumptions. They transform the binomial data on the log odds ratio scale and model the within-study variability using a bivariate normal distribution as data measured on this scale are assumed to be approximately normally distributed. However, this assumption is inappropriate when events occur rarely or very frequently. The proposed hierarchical method allows for modeling the within-study variability on the original binomial scale and accounts for the within-study associations leading to more precise inferences about the trial-level surrogacy patterns.

Finally, this thesis develops a hierarchical method for combining data from randomised control trials and data from single-arm observational studies in a single bivariate meta-analysis. Very often data measured on a short-term final endpoint are not sufficiently mature, or there is limited number of trials published. In these situations trial-level surrogacy patterns cannot be estimated accurately as randomised control trials do not provide sufficient information. The proposed methodology aims to improve inferences for trial-level surrogacy patterns, when randomised control trials offer limited information and evidence from different study designs need to be included for such validation.

The performance of the proposed methodologies were extensively assessed and compared against the standard approaches in various simulated data scenarios. They were also illustrated in data examples where targeted treatments were used.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my primary supervisor Professor Sylwia Bujkiewicz for offering me the opportunity to pursue this doctoral work and for the continuous support and guidance throughout my PhD studies. I would also like to thank Professor John Thompson for his insightful comments, advice and constructive discussions at various stages of my PhD. Special thanks to my second supervisor Professor Keith Abrams for devoting a substantial amount of his time to guiding me.

I would like to extend my sincere thanks to my fellow PhD students in the Biostatistics group for making my three years in Leicester very enjoyable and especially, to Betty and Alessandro for their friendship. My biggest thanks to my partner Danai, my brother George and my parents who supported me and had to put up with my stresses and moans for the past three years of study!

Table of Contents

List of Figures

List of Tables

Abbreviations

aCRC advanced colorectal cancer

anti-EGFR anti-epidermal growth factor receptor

BFs Bayes factors

BRMA bivariate random-effects meta-analysis

BRMA-BC bivariate random-effects meta-analysis with bivariate copulas

BRMA-IB bivariate random-effects meta-analysis with independent binomial likelihoods

CAST Cardiac Arrhythmia Suppression Trial

CCyR complete cytogenetic response

CIs confidence intervals

CML Chronic myeloid leukemia

CMP confidence profile method

CrIs credible intervals

DIC deviance information criterion

DMR deep molecular response

EFS event-free-survival

EMA European Medicines Agency

F-EX Full-exchangeability

FDA Food and Drug Administration

GLMM generalised linear mixed model

HMC Hamiltonian Monte Carlo

ICH International Conference on Harmonisation

IPD individual patient data

MCMC Markov chain Monte Carlo

MMR major molecular response

MRMA multivariate random-effects meta-analysis

NUTS No-U-Turn Sampler

OS Overall survival

P-EX Partial-exchangeability

PE proportion explained

PFS progression-free-survival

PNF product normal formulation

RCTs Randomised controlled trials

RE relative effect

RMSE root mean square error

TKIs tyrosine kinase inhibitors

TR tumor response

TTP time to progression

WAIC widely applicable information criterion

Chapter 1

Introduction

1.1 Aims of the thesis

Bivariate meta-analytic methods provide a natural framework to model the trial-level surrogacy relationships, combining evidence obtained from randomised controlled trials (RCTs). They, by nature, account not only for the between-studies correlation between the treatment effects measured on the surrogate endpoint and the final outcome, but also all related uncertainty required both at within-study and between-studies levels. Accounting for the between-studies correlation, either directly or through some linear relationship between the treatment effects, allows the bivariate meta-analytic methods to quantify the strength of trial-level surrogate relationships. However, the trial-level validation of surrogate endpoints based on data from modern clinical trials of targeted therapies present a number of methodological challenges. This thesis aims to address three methodological challenges related to the trial-level validation of surrogate endpoints in disease areas where precision medicine have played an important role. These challenges are addressed, by proposing novel methodology for each of the following three aims:

- Improve the trial-level validation of surrogate endpoints within a specific class of treatment in disease areas where trial-level surrogacy patterns vary across treatment classes due to, for example, the diversity of the mechanisms of action of targeted therapies.

- Improve the trial-level validation of surrogate endpoints, when such validation is based on correlated binomial aggregate data with high or low proportions of events (such as high response rate or low death rate due to the increase effectiveness of targeted treatments)

- Strengthen the trial-level validation of surrogate endpoints when RCTs offer limited information by allowing for external evidence from different study designs to be included in such validation.

In the remainder of this chapter, we introduce the background of this thesis and the motivating data examples in Section 1.2, the concept of surrogate endpoints in Section 1.3 and the structure of the thesis in Section 1.4.

1.2 Background to the thesis

New advances in molecular science have unveiled a vast genetic heterogeneity among tumors, even within a tumor entity[1]. This knowledge has opened a new area in oncology, shifting cancer treatment from the traditional "one-size-fits all" approach for large groups of population to therapies tailored to specific subgroups of patients according to the genomic signature [2]. Precision medicine is the term that is increasingly being used to describe targeted treatments, diagnosis and disease prevention of individuals or of small groups of patient populations based on the molecular understanding of their disease. Precision medicine became extremely popular in oncology when a Obama's precision medicine initiative was launched in the United states (US) in 2015[3], aiming to accelerate progress towards the development of biomarker-driven treatments. The identification of reliable biomarkers is one of the most important aspects of precision medicine. Biomarkers can be a unique mutated gene, a protein or group of proteins, that allow cancer cells to grow and survive. Personalised treatments target these cancer specific genes resulting in significantly improved overall survival (OS). One of the first targeted therapies was imatinib [4–6] developed as treatment of chronic myeloid leukaemia CML. Imatinib targets a mutant protein which is found only in the cancer cells of CML patients and was approved by the Food and Drug Administration (FDA) in 2001. Since then, a multitude of targeted treatments have been developed in various disease areas, such as anti-epidermal growth factor receptor (anti-EGFR) therapies used in non-small cell lung cancer and some types of colorectal cancer, such as advanced colorectal cancer (aCRC).

Before a new therapy is licensed for market assess, its safety and efficacy has to be assessed. The evaluation process can last several years before the drug can be deemed safe to be released to the population. Usually, it takes considerable time for data measured on the final outcome to become mature enough for the effectiveness to be measured on this outcome. Furthermore, as therapies are targeted, the disease population may be small and therefore trials are often of small size. In addition, such novel targeted therapies are often very successful, resulting in very few events (such as deaths) early in the trial. The small sample size and lack of recorded events lead to considerable uncertainty around early measurements of treatment effectiveness on the final outcome. For instance, in slow progressing diseases with very few events (e.g. deaths) such as CML, the estimation of treatment effect on the final outcomes is obtained with high uncertainty.

There is, also, increased pressure on regulatory agencies, such as FDA and European Medicines Agency (EMA), to approve new promising therapies quickly making approvals based on the long-term final outcomes almost impossible. Regulatory agencies have introduced conditional licensing based on early measurement of treatment effect measured on a surrogate endpoint [7, 8] as it often can be obtained with lower uncertainty.

These challenges generate the need for surrogate endpoints that can be measured more reliably and earlier compared to long-term final outcomes such as OS [9]. Surrogate endpoints should be validated for their predictive value of clinical benefit or harm, by assessing association patterns between the treatment effects on the surrogate endpoint and the treatment effects on the final outcome. The validation of candidate endpoints as reliable surrogate endpoints has been challenging process. The complete validation of surrogate endpoints requires three evaluation levels. Firstly, there must be biological plausibility of the association between the two outcomes (the surrogate and the final), secondly, an individual-level association between the two outcomes needs to be evaluated and lastly, a trial-level association between the treatment effects on the two outcomes needs to be valid to ensure that the surrogate endpoint is a good predictor of clinical benefit (the evaluation levels are discussed in detail in section 1.3.2). Trial-level surrogate relationships can be assessed using bivariate meta-analytic methods as they take into account the between-studies association between treatment effects on the surrogate endpoint and the final outcome and all the necessary sources of uncertainty [10]. However, data from modern clinical trials of targeted therapies present a number of methodological challenges in surrogate endpoint validation, due to, for example, the diversity of the mechanisms of action of new therapies, high rate of response to these therapies and small sample sizes of the trials. As discussed in Section 1.1, the aim of this thesis is to address three methodological challenges related to the trial-level validation of surrogate endpoints in the era of precision medicine. Here, we discuss the aims in more detail, describing also the two motivating data examples:

1. Association patterns between treatment effects on the surrogate endpoint and the final outcome may differ across classes of treatment when targeted therapies are applied to subgroups of population with unique tumour characteristics (mutations). For instance, Ciani et al.[11] found sub-optimal surrogate relationship between treatment effects on progression-free-survival (PFS) and OS based on data from trials of therapies across all treatment classes in aCRC. On the other hand, Buyse et al.[12] found strong surrogacy pattern between the treatment effects on these two outcomes in aCRC using data consisting of only one treatment class. Therefore, the assumption that surrogate relationships remain the same across treatment classes of different mechanisms of action or lines of treatment in aCRC does not seem plausible. A simple way to examine potential differences in surrogate relationships across treatment classes is by performing subgroup analysis using a bivariate meta-analytic method [13–16]. This type of analysis is very practical when there are sufficient data within treatment classes, but it may fail to estimate association patterns effectively when data are limited in terms of the number of studies. The first aim of this thesis is to develop a meta-analytic method allowing for trial-level validation of surrogate endpoints within each treatment class, whilst borrowing of information for the parameters describing surrogate relationships across classes.

2. Standard bivariate meta-analytic methods can be used to investigate trial-level

surrogate relationships. These methods model the observed treatment effects on the surrogate and final outcome jointly using a bivariate normal distribution. When these methods are applied to correlated binomial aggregate data, the numbers of events in each arm across outcomes are transformed to obtain treatment effects on log odds ratio scale which are assumed to be approximately normally distributed. Hamza et al.[17] have shown that (in the univariate case) this assumption is reasonable only when the proportions of events are close to 0.5, otherwise (when the proportions are close to 0 or 1) it leads to biased results and estimates with considerable uncertainty. The high effectiveness of targeted therapies results in large proportions of responders and very small proportions of progressions or deaths. Therefore, modeling binomial aggregate data from trials of such therapies using the normal approximation may lead to poor inferences of trial-level surrogate relationships. For example in CML, which is a slow progressive disease with small death rates, the introduction of tyrosine kinase inhibitors (TKIs) dramatically improved long-term survival. Therefore, use of such approximation is inappropriate and can result in inaccurate surrogate endpoint validation. To overcome this methodological challenge, we aim to develop methods which allow for modelling correlated binomial aggregate data on the original binomial scale, avoiding the normal approximation and also accounting for within-study associations.

3. Often trial-level validation of surrogate endpoints fails due to limited evidence. In many cases, either RCTs do not provide sufficient evidence to validate a candidate endpoint as a surrogate, as data measured on the final endpoint are not mature enough, or there is limited number of trials published. In these situations, the standard meta-analytic methods struggle to estimate the between-studies association between the treatment effects on the surrogate endpoint and the final outcome precisely, affecting the trial-level validation of surrogate endpoints. For instance, in CML, RCTs usually report treatment effects on event-free-survival (EFS) at 2 years and OS at 2 years where the data are not mature enough [18, 19] and only a few trials provide long-term outcomes such as OS at 3 or 4 years. The final aim of the thesis is to explore the use of observational evidence, for example from single-arm observational

cohort studies, as source of suitable (large or long term follow-up) data for trial-level surrogate endpoint evaluation, and to develop methodological tools to incorporate such data into a single meta-analysis.

1.3 The concept of surrogate endpoints

This section outlines the concept of surrogate endpoints highlighting early successes and failures and the necessary steps for the validation of an endpoint as a surrogate.

Probably the most important factor affecting the duration and the complexity of development of new therapies is the choice of the endpoint to measure the efficacy of the new treatment. The sensitivity of the endpoint to detect treatment effects and its clinical relevance are the two main criteria of selection [20]. Clinical relevance depends on whether evidence of biological activity of the therapy is sought or whether a final evaluation of clinical benefit to patients has to be done. For example, in life-threatening diseases like aCRC or cardiovascular diseases, the most clinically relevant endpoint of a therapy is OS.

However, often the most relevant and sensitive clinical endpoint which will be referred to as final clinical outcome or final outcome throughout this thesis, may be difficult to use in clinical trials for a number of reasons [21]:

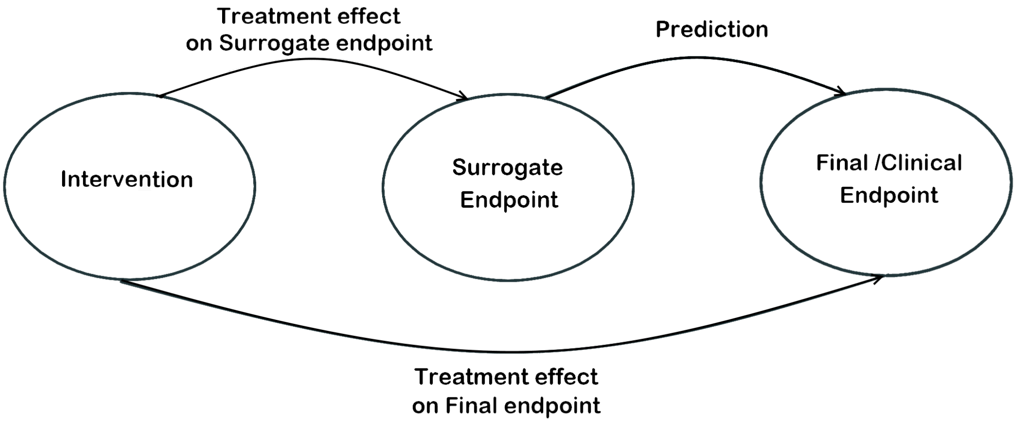
- it may be costly to measure treatment effect on the final outcome (e.g. cachexia is a condition associated with malnutrition, involving muscle and fat tissue loss and requires expensive equipment to measure content of potassium, nitrogen and water in patients)
- measuring treatment effect on the final outcome requires long follow-up times (e.g. in early stage cancers, it take long time to measure overall survival which conflicts with the need to deliver new treatments to patients quickly. In addition, such long follow up time may lead to the treatment effect on the final outcome to be confounded by other therapies.)
- final endpoints may require a large sample size if the event of interest has low incidence (e.g. The increased effectiveness of therapies targeted to specific, often small, patient populations reduce the number of events or deaths making

the measurement of treatment effect difficult)

- final outcomes may be difficult to measure (e.g quality of life or pain assessment includes multi-dimensional instruments that are hard to validate)

In such circumstances, the increased complexity and the duration of the research make use of the final endpoint not feasible. A potential solution is to investigate alternative endpoints which can be measured earlier, more frequently, with lower cost and more conveniently than the final endpoint. Such 'alternative' endpoints are defined as surrogate endpoints [22]. Figure 1.1 provides a graphical representation of the definition of surrogate endpoints and their relationship with the final outcome.

Figure 1.1: Definition of surrogate endpoints


Treatment effect on Surrogate endpoint
Prediction
Intervention
Surrogate Endpoint
Final /Clinical Endpoint
Treatment effect on Final endpoint

The Biomarker Definitions Working Group [23] proposed formal definitions that have been widely adopted:

Definition 1.3.1 *A final outcome is considered the most credible indicator of drug response and defined as a characteristic or variable that reflects how a patient feels, functions, or survives.*

Definition 1.3.2 *A biomarker is defined as a characteristic that can be objectively measured as an indicator of healthy or pathological biological processes, or pharmacological responses to therapeutic interventions. For example, blood or urine measurements and cell mutations can be used as biomarkers.*

Definition 1.3.3 *A surrogate endpoint is a biomarker that is intended for substituting a final outcome. A surrogate endpoint is expected to predict benefit,* *harm, or lack of these.*

1.3.1 Early successes and failures with surrogate endpoints

Due to the potential advantages, surrogate endpoints have been used in medical research under the assumption that high efficacy of a treatment on a surrogate endpoint would imply automatically an impact on a final outcome. Table 1.1 presents several examples of candidate endpoints evaluated based on an established association between treatment effects the potential surrogate endpoints and the effects on the corresponding final outcomes [24].

Table 1.1: Examples of potential surrogate endpoints used in medical research

Disease	**candidate surrogate endpoint**	**Final/ clinical outcome**
advanced cancer	progression free survival	overall survival
hypertension	blood pressure	cardiovascular mortality
Arrhythmia	arrhythmic episodes	survival
glaucoma	intraoccular pressure	vision loss
HIV infection	CD4 counts, viral load	progression to AIDS

However, the existence of an association between a candidate endpoint and a final outcome does not sufficiently imply that the candidate endpoint can be used as a surrogate. As Fleming and DeMets stated, 'a correlate does not make a surrogate'[25]. What is really required is that the treatment effects on the candidate endpoint should reliably predict the treatment effects on the final endpoint. Unfortunately, this condition was not checked sufficiently enough in the early attempts due to lack of appropriate methodology. For example, the most known case of unsuccessful replacement of the final endpoint was the approval of three anti-arrhythmic drugs (encainide, flecainide and moricizine) by the FDA in the US. This decision was based on the fact that they were shown high efficacy on the suppression of arrhythmias. It was believed that, as arrhythmias are associated with an almost fourfold increase in the rate of cardiac-complication-related death, treatments that reduced arrhythmic episodes would also reduce the death rate. However, the Cardiac Arrhythmia Suppression Trial (CAST) study [26], conducted after the drugs had been approved by the FDA, showed that the death rate was twice higher for patients who had treated with the approved drugs compared to the patients who had received placebo. The main reason for this failure was the incorrect assumption of surrogacy. This assumption seemed to be reasonable due to the association between the candidate endpoint and final outcome. This and other unsuccessful examples led to scepticism and negative opinions about the usefulness of surrogate endpoints in the assessment of treatment efficacy [25, 27–30].

Despite the early failures, there were many cases where their application proved vital for the drug development process, having remarkable results in a number of disease areas. For instance, during the first stages of the AIDS epidemic, the impressively early encouraging results obtained with zidovudine led to the use of CD4+ T-lymphocyte counts as a successful surrogate endpoint of progression to AIDS resulting in the fast approval of new successful therapies [13].

1.3.2 Validation of surrogate endpoints

Regardless of the failed attempts in the past, the importance of surrogate endpoints in drug development process requires that candidate endpoints should be validated before deciding on the use of such endpoints. Consequently, formal methodology allowing for validation is required. In practice, the most common way to validate a candidate endpoint as a surrogate is to examine whether or not it satisfies three levels of association proposed by the International Conference on Harmonisation (ICH) Guidelines on Statistical Principles for clinical trials [31].

1. There must be biological plausibility of the association between the candidate surrogate endpoint and the final outcome. This association involves biological rather than statistical considerations.

2. Epidemiological studies should demonstrate the prognostic value of the candidate surrogate endpoint for the final outcome. In other words, treatment effect on a candidate surrogate endpoint may be used to predict the course of a disease in an individual patient. This situation is referred to in the literature as individual-level surrogacy.

3. There must be evidence from multiple clinical trials that treatment effects on the candidate surrogate endpoint correspond to the treatment effects on

the final outcome. This situation is referred to in the literature as trial-level surrogacy.

The second and the third level of surrogacy are independent from each other, which is highlighted in the ICH guidelines [31].

1.3.3 Individual versus trial level surrogacy

Surrogate endpoints can be applied for different purposes depending on the phase of drug development. Early and intermediate endpoints are appropriate in non-randomised phase I or II trials when they have been shown strong individual level surrogacy. However, very few of these endpoints have been also shown acceptable trial level surrogate relationships and can replace a final outcome in multiple phase III clinical trials [21].

When data from only a single study are available, only individual-level surrogacy can be examined unless the size of the trial is substantial and the data can divided into smaller units by, for example, countries or regions. Most of the attempts to validate a Surrogate using data from a single study have been unsuccessful and more recently the attention has shifted to the meta-analytic framework where multiple trials they can be analysed simultaneously when individual patient data (IPD) are available. This allows full surrogate endpoint evaluation both at the individual and trial-level [32].

1.4 Structure of the thesis

This thesis is structured into seven chapters, where the first chapter provides an introduction of the concepts of surrogate endpoints and presents the methodological challenges that this PhD thesis aims to address.

Chapter 2 and 3 highlight the meta-analytic framework, outline important aspects of Bayesian statistics and present the most important meta-analytic methods developed for individual and trial-level surrogate endpoint evaluation. The methods discussed in these two chapters will be used and extended further for development of novel methodology in Chapters 4, 5 and 6 as briefly described below.

Chapter 4 introduces two new methods to address the first of the methodological challenges. The proposed methods investigate surrogate relationships within treatment classes assuming different levels of exchangeability about the parameters describing surrogate relationships, thus facilitating different degrees of borrowing of information across the classes. They extend a standard meta-analytic model for trial-level evaluation of surrogate endpoints by adding another level to its hierarchical structure in order to account for differences in surrogacy patterns across classes of treatment. This leads to more precise inferences about the association patterns between the treatment effects on the surrogate and the final outcome compared to subgroup analysis due to borrowing of information across treatment classes.

Chapter 5 addresses the second methodological challenge by proposing a new method which allows for modelling binomial data on the original scale and accounts for within-study associations. The proposed method models the numbers of events on the first and the second outcome jointly using a bivariate density with binomial marginals constructed with copulas. This allows the model to account for within-study associations between the numbers of events on two (surrogate and final) binomial outcomes. An additional method is also presented in this Chapter to highlight the importance of accounting for within-study associations. This additional method models the within-study variability using binomial likelihoods, but ignores within-study association. Overall, modeling the within-study variability on the original binomial scale and accounting for within-study associations improves the trial-level validation of surrogate endpoints, resulting in reduced bias and increase the precision of the estimates of the parameters describing the surrogate relationships.

Chapter 6 presents two approaches aiming to improve the trial-level validation of surrogate endpoints when evidence from RCTs are limited and observational evidence are required for such validation. It introduces a hierarchical method that incorporates data from all available sources such as observational cohort studies or non-randomised single arm phase II trials in the surrogate endpoint evaluation, combining these data with data from RCTs. Under this approach, non-comparative observational studies contribute to the estimation of the between-studies association of the relative treatment effects and can potentially improve the trial-level surrogate endpoint validation. Two extensions of the method were also developed. The first version accounts for systematic biases, whilst the second one can effectively model correlated binomial aggregate data with high proportions of events using a bivariate density with binomial marginals constructed with copulas (similarly as in Chapter 5).

Chapter 7 concludes the thesis by summarising the findings and the conclusions from Chapters 4, 5 and 6 and discusses how the proposed methodology can improve the trial-level validation of surrogate endpoints. Limitations in the application of the proposed methodologies are outlined and opportunities for further work are also discussed.

Chapter 2

Bayesian statistics and Meta-analytic framework

2.1 Chapter overview

This chapter presents the statistical concepts and methodologies used and extended throughout this PhD thesis. In this thesis, the proposed methodology was developed in the Bayesian framework, as it offers a very flexible way to model all relevant uncertainty. The Chapter begins with a brief review of the Bayesian statistics defining the necessary terminology, the discussion of the Markov chain Monte Carlo sampling methods and the statistical software used to perform Bayesian statistical analysis. This is followed by a review of the fundamental meta-analytic methods highlighting the key assumptions and setting the scene for more complex hierarchical meta-analytic methods discussed later in the thesis.

2.2 Bayesian inference

The origins of Bayes theorem dates back to 1763 when Thomas Bayes' work was published posthumously [33]. In this work Bayes and Price proposed a theorem to relate marginal and conditional probabilities for observed events, which termed as Bayes' theorem. In Bayesian statistics if θ is the unknown parameter of interest and Y is data describing θ, Bayes' theorem takes the following form:

$$p(\theta|Y) = \frac{p(\theta)p(Y|\theta)}{\int p(\theta)p(Y|\theta)d\theta} \tag{2.1}$$

and is commonly expressed such as:

$$p(\theta|Y) \propto p(\theta)p(Y|\theta) \tag{2.2}$$

$$\text{i.e} \quad \text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

where $p(Y|\theta)$ is the likelihood of θ, $p(\theta)$ is the prior probability density of beliefs for θ and $p(\theta|Y)$ is the resulting posterior probability density for θ after combining the likelihood with the prior beliefs. The main characteristic of Bayesian statistics is that data are supplemented using prior beliefs or external evidence. Bayes theorem updates the prior beliefs regarding θ multiplying the likelihood density $p(Y|\theta)$ by the prior distribution $p(\theta)$. By including the normalising constant $\int p(\theta)p(Y|\theta)$ the posterior probability function $p(\theta|Y)$ is a proper probability density function and inferences regarding the parameter θ can be obtained by using this density. In contrast to the frequentist approach, where the parameter θ is a fixed unknown number, in the Bayesian framework θ is treated as random variable and hence probability distributions can be specified for this parameter [34]. A careful selection of a prior distribution is crucial in Bayesian statistics. Prior distributions typically take one of the following forms.

- **non-informative - vague priors:** They do not express any prior belief or information about the parameter of interest considering all values of θ equally likely e.g. $\theta \sim U(a, b)$. These priors are useful when initial beliefs concerning θ are very limited or we prefer to base our inferences on collected data. A characteristic example of a vague prior distribution when it has infinite variance e.g. $\theta \sim N(\mu, s^2)$ where s is very large i.e. $s \to \infty$ [35].
- **informative priors:** These priors are often constructed based on historical data or subjective beliefs elicited from "experts" or external evidence [36]. They have a considerable influence on the posterior distribution especially when data are limited and they should be used with caution[37].

Figure 2.1 illustrates the impact of non-informative and informative prior distributions on the posterior distributions. On the left hand side the non-informative prior distribution plays a minimal role in the posterior distribution allowing the likelihood to dominate the prior. On the other hand, on the right hand side, the informative prior distribution has a substantial impact on the shape of the posterior distribution.

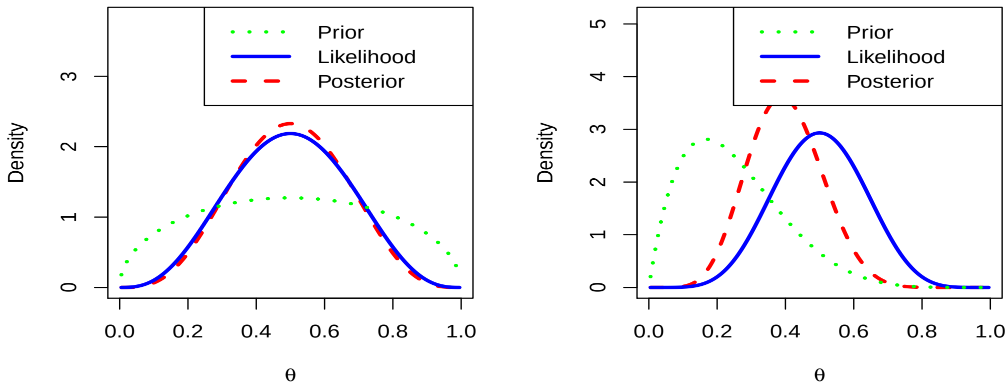


Figure 2.1: Non-informative (left) and informative prior (right) distributions and their impact on the posterior distribution

One of the disadvantages of Bayesian statistics includes the incorporation of prior beliefs. This makes the analysis no longer completely objective, and hence defining a reasonable prior distribution may be a difficult task. When the inclusion of prior beliefs is inappropriate, vague/non-informative prior distributions can be applied to allow the data dominate the prior. However, defining the appropriate vague prior distributions can be a difficult task especially for variance parameters [38]. Therefore, it is important to assess prior distribution specification through sensitivity analysis.

2.2.1 Bayesian point estimates and credible intervals

In this thesis the estimation of posterior distributions is of most interest and specifically the estimation of the parameters describing the surrogacy patterns. Posterior mean is used as point estimate when the posterior is symmetric and posterior median is used in situations where the distribution is skewed. Typically, posterior distributions are presented with Bayesian credible intervals (CrIs) and can be considered as a Bayesian equivalent to frequentist confidence intervals (CIs). However, definition of the CrI is slightly different compared to CI. Specifically, the width of CrIs is determined by the standard deviation of the posterior distribution, whilst the width of CIs is determined by the standard error of the estimate. Therefore, the interpretation of CrIs is also slightly different and more intuitive compared to the CIs. Bayesian CrIs have a probabilistic interpretation, whereas the correct interpretation for frequentist CIs refer to the long-term success rate of the method i.e. after a long series of 95% CIs constructed from replicated experiments, 95% of them will contain the true value of θ [34].

2.2.2 Bayes factors

Bayes factors (BFs) are used to provide a natural way to compare two alternative hypotheses. This approach does not rely on arbitrary significance levels compared to the traditional frequentist hypothesis testing which are also very dependent on sample size. If we consider two hypotheses H_0 and H_1, with prior probabilities $p(H_0)$ and $p(H_1)$ and likelihoods $p(y|H_0)$ and $p(y|H_1)$ respectively, we can compare the two hypotheses by calculating the following relative probabilities:

$$\underbrace{\frac{p(H_0|y)}{p(H_1|y)}}_{\text{posterior odds}} = \underbrace{\frac{p(y|H_0)}{p(y|H_1)}}_{\text{Bayes Factor}} \times \underbrace{\frac{p(H_0)}{p(H_1)}}_{\text{prior odds}} \tag{2.3}$$

The relative likelihood of the two hypotheses is also known as BF and contains all the evidence that can be extracted from the data about the two hypotheses. BFs can vary from 0 to ∞, with small values considered as both evidence against H_0 and for H_1. The following table proposed by Jeffreys [39] provides a scale of the BF range.

2.2.3 Markov chain Monte Carlo

The application of Bayes' theorem is straightforward when a prior probability density combined with the a likelihood function result in a posterior distribution which belongs in the same family of distributions as the prior distribution. Models with such property are termed as conjugate models. However, there are situations where conjugate models are not feasible. This can be the case when the unknown

Table 2.1: Calibration of BFs provided by Jeffreys

Bayes factor range	Strength of evidence in favour of H_0 and against H_1
>100	Decisive
32 to 100	Very strong
10 to 32	strong
3.2 to 10	Substantial
1 to 3.2	Not worth more than a bare mention
	Strength of evidence in against H_0 and in favour of H_1
1 to 1/3.2	Not worth more than a bare mention
1/3.2 to 1/10	Substantial
1/10 to 1/100	Strong
1/32 to 1/100	Very strong
<1/100	Decisive

parameter θ has high dimensionality and the application of Bayes' Theorem yields a multi-dimensional joint posterior probability density for θ, making the analytical calculation of the marginal posterior distribution for each individual parameter a difficult task [40]. Markov chain Monte Carlo (MCMC) methods provide an efficient way to numerically estimate these multi-dimensional integrals [36] and their use dramatically increased in the last two decades. MCMC methods were firstly introduced by Metropolis in 1953 [41] and extended by Hasting in 1970 [42]. However, they remained completely unused until 1984 when Gibbs sampler was proposed by Geman and Geman [43]. MCMC methods are a group of iterative algorithms for drawing random samples from a probability distribution by using the main properties of Markov chains and Monte Carlo integration. A Markov chain is a sequence of random variables $\theta^{(1)}$, $\theta^{(2)}$, $\theta^{(3)}$, $\theta^{(4)}$,... satisfying the following property: $P(\theta^{(t+1)}|\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}) = P(\theta^{(n+1)}|\theta^{(n)})$ i.e the future of the chain depends on the present only. Under regularity conditions, a Markov chain $\theta^{(n)}$ converges to an "equilibrium distribution" as $n \to \infty$ regardless of the initial point $\theta^{(1)}$. Therefore, a sample of the distribution of interest is obtained when the Markov chain reaches its "equilibrium distribution" after a number of iterations in the algorithm. Using the generated sample, point estimates can be obtained for the distribution of interest [44]. Estimating for instance the mean of a sample is much easier task than solving equations analytically. The Markov chain converges to its equilibrium distribution after a considerable number of iterations and convergence in MCMC refers to the final and stable set of samples from the equilibrium or "stationary" posterior distribution (i.e. values should look like a random scatter around a stable mean). It is important to highlight that a series of samples converges to a distribution and not a specific value. A number of iterations that precedes convergence should be discarded from the generated sample and this initial number of iterations defined as "burn-in" period. There is no specific rule about the length of the burn-in period and it exclusively depends on the Markov chain and how fast it converges.

2.2.4 Gibbs sampler

One of the most popular and simplest MCMC methods is Gibbs sampler [43]. In Gibbs sampling the idea is to break the sampling of the multivariate high-dimensional posterior distribution into a series of samples from low-dimensional conditional distributions. The method takes sequentially each parameter of a model and draws a random sample from its posterior distribution, conditional on all the other parameters being fixed. Considering a situation of a model with four parameters $(x, \theta, \psi, \gamma)$, the algorithm takes the following form:

```
Gibbs Sampler algorithm:
  1. Initialisation: initialise the parameter space $(x^{(0)}, \theta^{(0)}, \psi^{(0)}, \gamma^{(0)})$ and the number of samples $N$
  2. for $i = 0$ to $N - 1$ do
       • Simulate $x^{(i+1)} \sim p(x|\theta^{(i)}, \psi^{(i)}, \gamma^{(i)})$.
       • Simulate $\theta^{(i+1)} \sim p(\theta|x^{(i+1)}, \psi^{(i)}, \gamma^{(i)})$.
       • Simulate $\psi^{(i+1)} \sim p(\psi|x^{(i+1)}, \theta^{(i+1)}, \gamma^{(i)})$.
       • Simulate $\gamma^{(i+1)} \sim p(\gamma|x^{(i+1)}, \theta^{(i+1)}, \psi^{(i+1)})$.
  3. return $(\{(x^{(i)}, \theta^{(i)}, \psi^{(i)}, \gamma^{(i)})\}_{i=0}^{N})$
```

Although Gibbs sampler algorithm is a very popular method, it is not without drawbacks and limitations. Gibbs sampling requires the posterior conditional distribution for each of the variables, however, in many cases it is not an easy task. Even if the conditional distributions can be extracted, they may not be in known forms, so samples cannot be drawn from them. Additionally, drawing from multiple conditional distributions may be slow and inefficient. As variables become more correlated, the performance of the Gibbs sampling decreases. This behaviour leads to higher correlations between samples and slow mixing of the chain [45]. In this thesis we focus on the development of complex hierarchical models consisting of many parameters and layers of hierarchy. In this kind of structures, the correlations between the parameters in multiple layers of the hierarchical models can be substantial and the efficiency of sampling methods such as Gibbs sampling become limited.

2.2.5 WinBUGS/ OpenBUGS

WinBUGS and OpenBUGS [46–51] are statistical packages developed for Bayesian analysis using MCMC methods. They use Gibbs sampling as their main MCMC method and require the probability model to be specified in BUGS language. The flexibility of these packages allow of modelling relatively complex methods, however, they suffer from the inefficiencies of Gibbs sampling methods. R2OpenBUGS [50] and R2WinBUGS [49] are very popular statistical packages to perform Bayesian inference, linking R software to OpenBUGS and WinBUGS. They exploit data management functionality of R allowing OpenBUGS code to be executed in R environment. In this thesis, R2OpenBUGS was used for the implementation of the methods discussed in Chapter 4.

2.2.6 Hamiltonian Monte Carlo

As described in section 2.2.4, sampling high-dimensional posterior distributions with Gibbs sampling becomes very inefficient in practice. An alternative and more efficient scheme is called Hamiltonian Monte Carlo (HMC) [52, 53]. HMC is one of the algorithms of the Markov chain Monte Carlo methods that utilises differential geometry techniques to generate transitions spanning the full marginal variance. The method can avoids the random walk behavior endemic to Gibbs sampler and achieves a more consistent and effective exploration of the probability space, being less sensitive to correlated parameters. The algorithm uses a physical system known as Hamiltonian dynamics.

If $p(\theta|y)$ is the target posterior distribution for parameters θ give data y, HMC generates auxiliary momentum variables r drawing it from a joint probability density $p(r,\theta) = p(r|\theta)p(\theta)$. In principle r is drawn from: $r \sim MVN(0, M)$ that does not depend on θ. The joint density $p(r,\theta)$ defines a Hamiltonian,

$$\begin{aligned} H(r,\theta) &= -logp(r,\theta) && (2.4) \\ &= -logp(r|\theta) - logp(\theta) && (2.5) \\ &= \quad T(r|\theta) + V(\theta) && (2.6) \\ &= \quad T(r) + V(\theta) && (2.7) \end{aligned}$$

where $T(r|\theta)$ is defined as "kinetic energy" and $V(\theta)$ as "potential energy". Starting from the current value of the parameter θ, a transition to a new state is generated in two steps before being assessed by a Metropolis accept step. First, a value for the momentum is drawn independently of the current values of θ. Next, the joint system (r,θ) using the current values of θ and r is evolved according to Hamilton's equations:

$$\begin{aligned} \frac{\partial H}{\partial r} &= \quad \frac{\partial T}{\partial r} + \frac{\partial V}{\partial r} = \quad \frac{\partial T}{\partial r} && (2.8) \\ \frac{\partial H}{\partial \theta} &= -\frac{\partial T}{\partial \theta} - \frac{\partial V}{\partial \theta} = -\frac{\partial V}{\partial \theta} && (2.9) \end{aligned}$$

As explained previously, the momentum is independent of θ i.e. $p(r|\theta) = p(r)$ ($r \sim MVN(0, M)$), thus the first term in equation (2.9) is zero ($-\frac{\partial T}{\partial \theta} = 0$). The last part of the method solves the two-stage differential equation using a numerical integration algorithm called leapfrog integrator. The algorithm starts by drawing a new momentum value which is independent of the values of θ or the previous values of the momentum. Next it updates the parameters and the momentum according to the following equations:

$$r^{t+\epsilon/2} = r^t - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}(\theta^t), \quad \theta^{t+\epsilon} = \theta^t + \epsilon\frac{\partial T}{\partial r}(r^{t+\epsilon/2}), \quad r^{t+\epsilon} = r^{t+\epsilon/2} - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}(\theta^{t+\epsilon})$$

where t is time, ϵ is a discrete step of some small time interval and L is the number of repetitions. Finally, to account for numerical errors during the numerical integration, a Metropolis acceptance step is applied, with probability of keeping the proposed values (r^*, θ^*), generated from transition (r, θ) to be :

$$min(1, exp(H(r, \theta) - H(r^*, \theta^*))) \tag{2.10}$$

if the proposed values are not accepted, the previous parameter values are returned for the next draw and used to initialize the next iteration.

Summarising all aforementioned elements, the HMC algorithm can be described as follows:

```
HMC Algorithm :
Given $\theta^0, \epsilon, L, M$
For $i = 1$ to $N$ iterations do
    1. $r^0 \sim N(0, M)$
    2. Set: $\theta^i \leftarrow \theta^{i-1}$, $\theta^* \leftarrow \theta^{i-1}$, $r^* \leftarrow r^0$
    3. For $j = 1$ to $L$ do
       Set $(r^*, \theta^*) \leftarrow$Leapfrog$(\theta^*, r^*, \epsilon)$
       end for
    4. Set $\theta_i \leftarrow \theta^*$ with probability: $min(1, exp(H(r, \theta) - H(r^*, \theta^*)))$
end for
Function Leapfrog$(\theta, r, \epsilon)$
        Set $r^* = r - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}(\theta)$
        Set $\theta^* = \theta + \epsilon\frac{\partial T}{\partial r}(r^*)$
        Set $r^* = r^* - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}(\theta^*)$
return $(r^*, \theta^*)$
```

Unfortunately, the performance of HMC is highly affected by the choice of the hyperparameters for t and ϵ. A poor choice of hyperparameters can potentially dramatically decreases the efficiency of HMC [54]. Hoffman and Gelman developed the No-U-Turn Sampler (NUTS) to mitigate the challenges of tuning the parameters [54]. NUTS uses a recursive algorithm to automatically tune the HMC algorithm without requiring an external intervention or costly tuning runs. These improvements have been packaged into a modelling software called Stan [55–57].

2.2.7 STAN

Stan is a platform for statistical modeling and high-performance statistical computing. Using Stan, a user can perform full Bayesian statistical inference, posterior visualisations and leave-one-out cross-validation. In contrast to WinBUGS, Stan allows for user-defined functions and distributions. This was extremely important, as the proposed methodology in Chapters 5 and 6 uses bivariate copula densities which are not included as build-in distributions in WinBUGS.

In addition to its standard features, Stan also offers a variety of coding techniques such as variable re-parameterisation, multiple indexing, statistical and computational efficiency. In this section, we present a small sample of Stan's statistical and computational efficiency techniques used in the proposed hierarchical models.

There is a main difference between computational and statistical efficiency for Stan programs. Computational efficiency measures the amount of memory or time required for a step in a calculation, such as the evaluation of a posterior distribution. On the other hand Statistical efficiency typically involves requiring fewer steps in algorithms by improving the statistical formulation and making a model better behaved. The standard way to do this is by applying reparameterising variables so that MCMC algorithm is able to mix better. Sampling posterior distributions with difficult geometries is a difficult task for sampler such as NUTS. A typical way to speed up hierarchical models is via reparameterisation [58]. In specific situations, an appropriate parameterisation can dramatically improve chain mixing and convergence.

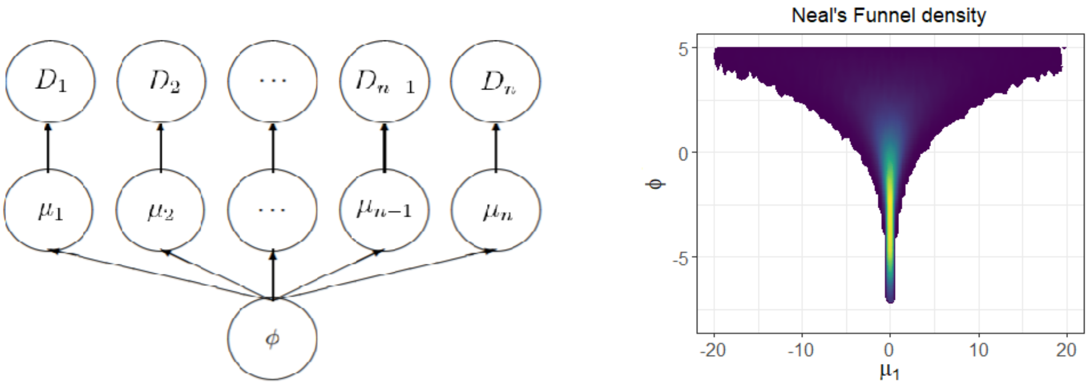
Choosing between centred and non-centred parameterisations

The choice of the correct parameterisation applies to any MCMC sampler, however, it is particularly important in Stan as it has substantial impact on the performance of the NUTS sampler. To illustrate the problem better we considered a simple hierarchical normal model given by:

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma_i^2) \\ \mu_i &\sim N(\mu, \tau^2), \text{ for } i = 1, \ldots, n \end{aligned} \tag{2.11}$$

In terms of notation $\phi = (\mu, \tau)$ will be referred as global parameters, μ_i as local parameters and $D_i = (y_i, \sigma_i)$ as data. Figure 2.2a visualises how the local parameters μ_i interact through a common dependency on the global parameters ϕ and how the interactions allow the data D_i to inform all the local parameters μ_i. As the bottom of the hierarchical structure (D_i) depends on ϕ, a small change in ϕ induces large changes in the density. Consequently, when data are sparse, the posterior density looks like a "funnel" [59] which has a high density and low volume area at the bottom and an area with low density and high volume at the top. In this situations sampling algorithms including NUTS struggle to generate samples from the neck of the funnel (where ϕ is small) and fail to explore the posterior distribution fully (Figure 2.2b) .

Figure 2.2: a)Layers of Hierarchy, b)Neal's Funnel



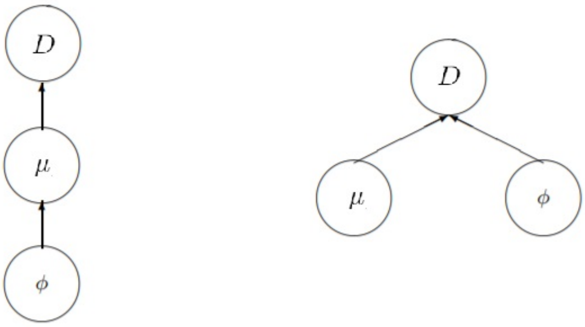
Hierarchical models such as the normal model 2.11 suffer from this kind of inefficiencies as μ_i and $\phi = (\mu, \tau)$ are correlated. The strength of this correlation depends on the amount of data with Neal's funnel being more extreme when data are very sparse [58]. This behaviour is very usual in meta-analysis where the number of studies is often limited. However, in circumstances where data are very informative, centred parameterisation is more efficient [60].

An easy way to reduce the correlation between the successive layers of the hierarchical structure and improve sampling is to separate each layer with auxiliary variables. For instance the second level of the model 2.11 can be written:

$$\mu_i = \mu + \tau z_i, \quad z_i \sim N(0,1) \tag{2.12}$$

This is called non-centred parameterisation [61] and helps the layers to become independent conditioned on z. With this type of parameterisation, the MCMC sampler can efficiently sample from the target distribution (Figure 2.3).

Figure 2.3: a)centred parameterisation, b)non-centred parameterisation

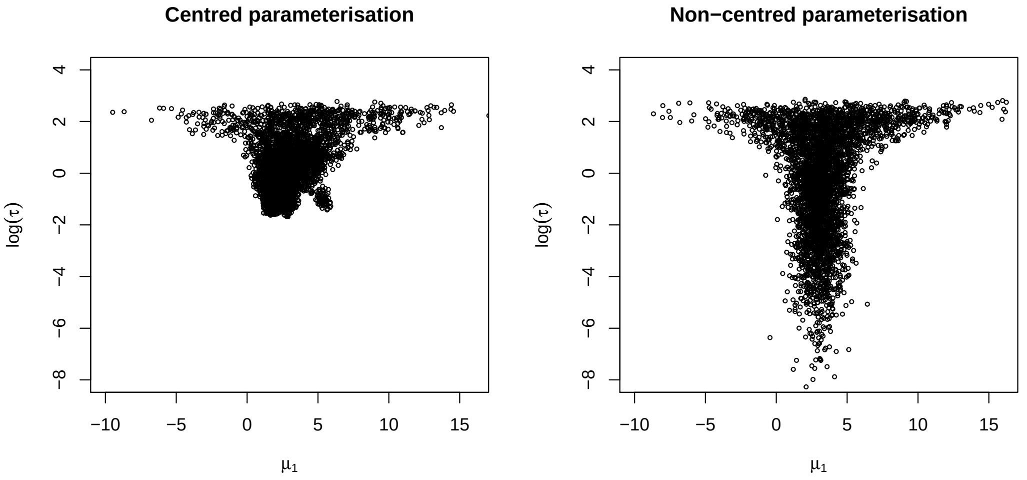


To illustrate the performance of the non-centred parameterisation we generated 10 data points from the normal model using the following values $\mu = 3, \tau = 3$ and $\sigma_i \sim Unif(9, 10)$. Two versions of the normal model (2.11) fitted to this dataset, one with centred parameterisation and one non-centred parameterisation. Figure 2.4 presents the values between $log(\tau)$ and μ_1 under the two versions of the model (the code of the model can be found in the Appendix A.1). Stan under the non-centred parameterisation, samples from the neck of the funnel much more efficiently using the non-centred parameterisation compared to centred.

Vectorisation

Stan spends the vast majority of the time computing the gradient of log probability functions, making gradients an obvious target for optimisation. Gradient calculations of Stan require a template expression to be allocated and constructed for each sub-expression such as the parameters of a Stan model. Vectorising these sub-expressions i.e. parameters of the model, reduce the time of gradient calculations. In this thesis the probability functions of the hierarchical models

Figure 2.4: Posterior distributions



developed in Stan were vectorised. This makes some of the "for loops" in the code redundant speeding-up the model substantially.

2.2.8 Convergence

As discussed in section 2.2.3, a Markov chain converges to its equilibrium distribution after a considerable number of iterations and convergence in MCMC refers to the final stable set of samples from the equilibrium or "stationary" posterior distribution. In this thesis the proposed methodology was implemented in R2OpenBUGS and RSTAN. In both software convergence was assessed visually by checking the trace plots posterior density plots and autocorrelation plots using graphical tools in R and by running multiple chains (n=3). When OpenBUGS was used, where posterior estimates were obtained using MCMC simulations performing 50000 iterations after discarding 20000 iterations as burn-in period. STAN required a much smaller number of iterations and achieved convergence after 5000 iterations (after excluding 1000 iterations as warm-up period). Trace and density plots of the proposed methods can be found in the Appendix B.7, C.8 and D.4.

2.3 Introduction to meta-analysis

Meta-analyses have firstly been discussed by Glass [62] as "the statistical analysis of large collection of analysis results from individual studies for the purpose of integrating the finding". Typically, it is defined as the statistical part of a systematic review process integrating the results of several independent studies considered to be similar. It includes the analysis of extracted data from the primary research by using quantitative methods to explore the heterogeneity of the data (studies) and to estimate overall measures of effects. Additionally, it can be used to assess the validity of the results to possible threats such as publication bias or bad study quality. When many trials investigate the same question, such as the effectiveness of a treatment, it is likely that the smaller studies have conflicting findings due to lack of statistical power. Hence, some may show results favouring a treatment and others show no treatment benefit at all. Pooling all the relevant studies together leads to more reliable and precise estimates of treatment effect. In situations where treatment benefit or harm is small or at best is modest the required sample size for an individual study to detect significant statistical difference between treatment groups may need to be substantial. In these cases, it is the increased power from synthesising findings from a number of trials that make systematic reviews and meta-analysis such important tools [63].

Meta-analytic methods started to be used more frequently in health care after the mid-1980s, when Yusuf [64] published the results from his systematic review and meta-analysis on beta blockers in myocardial infraction and became increasingly popular in the early 90s [65]. Nowadays meta-analyses include several studies which examine the same question, incorporating more patients than any single study potentially reducing the random error in the assessment of a treatment [66]. Furthermore, including results from many studies carried out in different places, having possibly slightly different selection criteria may produce more generalisable results averaging over all settings and contexts.

2.4 Concept of Bayesian meta-analysis and basic meta-analytic methods

The Bayesian meta-analysis involves four fundamental steps [67]:

1. Prior beliefs: The first step of Bayesian meta-analysis is to summarise evidence

external to observed data by identifying appropriate prior distributions. For instance additional evidence from observational studies, systematic reviews, RCTs and expert's opinions can be incorporated as prior distributions which can inform the meta-analytic model. These prior distributions are placed on the unknown parameters of meta-analytic models [37].

2. Observed data: Data collected from different RCTs answering the same question constitute the likelihood function of the parameters.

3. Posterior: External sources of evidence combined with data obtained from RCTs form a current state of knowledge regarding the parameters of interest (e.g. treatment effects). Thus, posterior distributions are obtained from the combination of likelihood functions and prior distributions for the parameters of interest. Any inferences in the Bayesian meta-analysis are based on the posterior distributions.

4. Summarising: The final step is to estimate point estimates from the posterior distribution. As it was mentioned in Section 2.2, summary estimates such as mean, standard deviation, 95% CrIs etc. are estimated from samples obtained from simulation techniques such as MCMC.

Similarly to traditional meta-analysis, two of the most commonly used models in Bayesian meta-analysis are the: fixed effect meta-analysis and random effects meta-analysis.

2.4.1 Fixed-effect meta-analysis of normally distributed data

A fixed effect meta-analysis model assumes homogeneity between studies, hence the observed treatment effect across studies estimate the same underlying pooled effect d. Algebraically, the observed treatment effects y_i follow a normal distribution with a single common pooled effect d and within-study variance σ_i^2.

$$y_i \sim N(d, \sigma_i^2) \tag{2.13}$$

The mean d corresponds to the true treatment effect and it is assumed to be common across all studies [68]. The within-study variances σ_i^2 are assumed known, however, in circumstances where they are not reported a prior distribution can be specified [68]. Implementing the model in the Bayesian framework a prior distribution for the pooled effect d needs to be specified. In the absence of prior beliefs, a vague prior distribution using a normal distribution is suitable when the outcome is continuous or is specified on the log-odds ratio scale. This normal distribution should be centred at zero (no effect) with large variance relative to the scale of the outcome e.g. $d \sim N(0, 10^2)$.

2.4.2 Random-effects meta analysis of normally distributed data

Random effects meta-analysis model assumes that each study has its own true effect δ_i differing from the effects of other studies. Algebraically, the true treatment effects follow a common normal distribution and therefore are called *random* effects as they are drawn randomly from this common normal distribution. The model can be described by the following hierarchical structure. At the within-study level, the observed treatment effects y_i are assumed normally distributed with individual mean true treatment effects δ_i and within-study variances σ_i^2

$$y_i \sim N(\delta_i, \sigma_i^2) \tag{2.14}$$

$$\delta_i \sim N(d, \tau^2). \tag{2.15}$$

The "random" true effects δ_i follow a normal distribution with mean d and variance τ^2 at the between studies level. The parameter d is the pooled treatment effect and τ^2 is the between-studies variance. When there is no heterogeneity between studies i.e. $\tau^2 = 0$, random effects meta-analysis is reduced to fixed effects meta-analysis. As described in section 2.4.1, in the absence of prior beliefs about the pooled effect d a vague prior distribution such as a normal prior distribution centred at 0 with sufficiently large variance should be used. Alternatively when there is suitable evidence external to the meta-analysis regarding d, for instance observational studies, such evidence can be incorporated in the analysis in the form of a the prior distribution. For the between-studies variance τ^2, prior distribution should be selected ensuring that only positive values are sampled. For instance inverse gamma or uniform or half normal distributions can be placed on standard deviation τ as prior distributions [16]. In any case, the use of prior distributions should be assessed via a sensitivity analysis [38].

2.4.3 Bayesian meta-analysis for binomial data

A typical way to perform random effects meta-analysis for binary outcomes is to work with observed treatment effects on the log odds ratio scale. Under this approach the summary data should be transformed to obtain observed treatment effects on the log odds ratio scale using the following formulas:

$$y_i = log(\frac{r_{Bi}}{N_{Bi} - r_{Bi}}) - log(\frac{r_{Ai}}{N_{Ai} - r_{Ai}}) \tag{2.16}$$

with corresponding variances:

$$\sigma_i^2 = \frac{1}{r_{Bi}} + \frac{1}{N_{Bi} - r_{Bi}} + \frac{1}{r_{Ai}} + \frac{1}{N_{Ai} - r_{Ai}} \tag{2.17}$$

where r_{Ai}, r_{Bi} are the numbers of events in treatment arms A and B, in study i and N_{Ai} and N_{Bi} are the total numbers of patients in arm A and B in study i. The observed treatment effects are assumed approximately normally distributed. However, it is questionable whether the within-study variability should be modelled via a normal approximation especially when the events are rare [17, 69].

Another modeling issue occurs when there are no events in either of the treatment arms or the number of events are equal to the number of patients in either of arms A and B. In this situation the log odds ratios y_i and their corresponding variances cannot be defined. A simple way to tackle this problem is to apply a correction, for example, by adding a constant number such as 0.5. However, in some situations the effect of adding a constant may lead to biased results [70, 71].

Smith et al. [72] proposed a different Bayesian univariate random effects model assuming that the number of events in the control arm r_{Ai} and in the new treatment arm r_{Bi} of the i^{th} trial follow independent binomial likelihoods.

$$r_{Ai} \sim Bin(N_{Ai}, p_{Ai}), \quad r_{Bi} \sim Bin(N_{Bi}, p_{Bi}) \tag{2.18}$$

N_{Ai} and N_{Bi} are the total numbers of individuals in arm A and B in study i whilst, the true probabilities of occurrence of an event in the arms A an B are p_{Ai}, p_{Bi} respectively. The baseline effects μ_i are calculated by transforming the true probabilities to the real line scale via a *logit* link function and hence the Bayesian random effects model can be written as follows:

$$logit(p_{Ai}) = \mu_i, \quad logit(p_{Bi}) = \mu_i + \delta_i \tag{2.19}$$

$$\delta_i \sim N(d, \tau^2) \tag{2.20}$$

At the between-studies level, the "random" true effects δ_i are modelled in the same way as the random effects model for normally distributed data (eq. 2.15). Additionally, in the Bayesian framework, a prior distribution needs to be placed on the baseline effects such as $N(0, b)$ (where b is assumed sufficiently large), the pooled effect d and the between-studies variance τ^2.

2.4.4 Heterogeneity

Random-effects model account for differences between the true effects across studies, but it does not sufficiently account for all sources of variation. Between-studies variability in the treatment effects include differences in patient population, administration of interventions, changes in medical practice or design of clinical trials. Systematic differences in treatment effects that are more than this, can be attributed to sampling error alone are termed statistical heterogeneity [68, 73]. Between-studies heterogeneity can be problematic in meta-analysis and sources of variability should be investigated and accounted for in the analysis [10]. When there are evidence of substantial between-studies heterogeneity, subgroup analyses for discrete characteristics, stratified by the covariates of interest can be performed. Alternatively, a meta-regression method can be applied to explore sources of heterogeneity, incorporating covariates such as, for example, average age [74].

2.4.5 Meta-regression

The random effects meta-analysis accounts for between-studies heterogeneity, but does not explicitly explain it. Meta-regression can be used to measure the association between the treatment effects and the measurable characteristics such as age using regression techniques. when the method is applied to random-effects meta-analysis it explains the residual between-studies heterogeneity using study-level covariates [75]. Using the previous notations the random-effects meta-regression can be written as follows:

$$\begin{aligned} y_i &\sim N(\delta_i + \beta(x_i - \bar{x}), \sigma_i^2) \\ \delta_i &\sim N(d, \tau^2) \end{aligned} \tag{2.21}$$

where β is the regression coefficient, x_i is the study-level covariate of interest in study i and $\bar{x}$ is the mean value of the covariate across all studies. Centering the study-level covariate around the mean allows d to be interpreted as the pooled effect for the average study characteristic. For instance, if the study-level covariate is age, d can be interpreted as the pooled effect for a patient of average age in the included studies. To perform this model in the Bayesian framework prior distributions need to be placed on the unknown parameters d, τ and β. As the regression coefficient β is unconstrained, it can be given a normal prior distribution with large variance. The remaining parameters can follow the same prior distributions as in sections 2.4.2 and 2.4.1. However, when the data are sparse, meta-regression may suffer from lack of sufficient power to detect the relationships it intends to estimate [76]. Furthermore, the analysis can be susceptible to unknown confounding factors and aggregation bias if the relationship between summary data do not reflect the true relationship at IPD level. Thus, meta-regression should be treated with caution and the relationship as associative than causative [68].

Chapter 3

Methods for surrogate endpoint evaluation

This chapter reviews existing methodology developed to evaluate surrogate relationships regardless of statistical framework. The first part focuses on methods investigating the individual-level surrogacy. These methods estimate a surrogate relationship between the two outcomes (the surrogate endpoint and the final outcome) using data from single trial, hence they are referred to in the literature as single-trial surrogate endpoint evaluation methods [77–80]. It is important to highlight that all these methods have fundamental theoretical and applied problems and they are discussed simply to set the scene for the multiple-trial surrogate evaluation methods. Meta-analysis provides a robust framework for combining information across studies investigating the same question and has been widely utilised to combine evidence from clinical trials to evaluate treatment efficacy. Bivariate meta-analysis can also be used for trial-level surrogate endpoint evaluation investigating the between-studies association between treatment effects on the surrogate and treatment effects on the final outcome. The second part of this chapter presents currently available meta-analytic methods used for trial-level surrogate endpoint evaluation, as well as, the criteria for surrogacy proposed by different researchers.

3.1 Single-trial evaluation methods for surrogate endpoints

In this section the seminal methods of Prentice [77], Freedman et al. [78], and Buyse and Molenberghs [79] are described to set the scene for more complex research methods.

3.1.1 Prentice's approach

Prentice formally defined a surrogate endpoint as: "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint" [77]. This definition represents the case where the surrogate endpoint should capture any relationship between therapies and final endpoint [81]. This definition can be written as:

$$p(S|Z) = p(S) \Leftrightarrow p(T|Z) = p(T)$$

where $p(S)$ and $p(T)$ denote the probability distributions of the random variables S (surrogate endpoint) and T (final outcome). The distributions, $p(T|Z)$ and $p(S|Z)$ denote the conditional probability distributions of S and T given Z. This definition includes the random variables T, S, Z, so S can be considered as a surrogate endpoint of the final outcome T given only to the effect of some specific treatment Z and not necessarily for a different treatment. Prentice proposed and formulated four operational criteria to validate a candidate surrogate endpoint. The candidate endpoint should fulfil all four criteria at the same time to be deemed a valid surrogate endpoint:

$$p(S|Z) \neq p(S) \tag{3.1}$$

$$p(T|Z) \neq p(T) \tag{3.2}$$

$$p(T|S) \neq p(T) \tag{3.3}$$

$$p(T|S, Z) = p(T|S) \tag{3.4}$$

The equation (eq. 3.1) represents that the treatment Z should have a statistically significant effect on the candidate surrogate endpoint S. Similarly, Z should have a statistically significant effect on the final outcome T (eq. 3.2), the candidate endpoint S should have a significant effect on the final endpoint T (eq. 3.3) The last equation (eq. 3.4) describes that the effect of the treatment Z on T should be captured by the candidate endpoint S.

If the surrogate endpoint S and the final outcome T are assumed to be normally distributed then the first two criteria can be examined by fitting a bivariate linear regression model:

$$S_j = \mu_S + \alpha Z_j + \epsilon_{S_j} \tag{3.5}$$

$$T_j = \mu_T + \beta Z_j + \epsilon_{T_j} \tag{3.6}$$

where α and β are the treatment effects and ϵ_{S_j} and ϵ_{T_j} the error terms on the surrogate endpoint and the final outcome respectively. The errors are also assumed to be normally distributed, following bivariate normal distribution with zero mean:

$$\begin{pmatrix} \epsilon_{S_j} \\ \epsilon_{T_j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix} \right). \tag{3.7}$$

The third criterion (equation 3.3) can be examined by fitting the following univariate liner regression model:

$$T_j = \mu + \gamma S_j + \epsilon_j \tag{3.8}$$

Finally, the fourth criterion (equation 3.4) can be explored by fitting another univariate linear regression model.

$$T_j = \tilde{\mu_T} + \beta_S Z_j + \gamma_Z S_j + \tilde{\epsilon}_{Tj} \tag{3.9}$$

where

$$\beta_S = \beta - \sigma_{ST}\sigma_{SS}^{-1}\alpha \tag{3.10}$$

$$\gamma_Z = \sigma_{ST}\sigma_{SS}^{-1} \tag{3.11}$$

and the variance of $\tilde{\epsilon}_T$ is given by:

$$Var(\tilde{\epsilon}_T) = \sigma_{TT} - \sigma_{ST}^2\sigma_{SS}^{-1} \tag{3.12}$$

For the Prentice's criteria to be satisfied, the hypotheses $H_0 : \alpha = 0$, $H_0 : \beta = 0$ and $H_0 : \gamma = 0$ (eq, 3.5, eq. 3.6 and eq, 3.8) should be rejected and the hypothesis $H_0 : \beta_S = 0$ should not be rejected.

3.1.1.1 Issues with Prentice's approach

The Prentice criteria are very practical and straightforward to be tested but there are some fundamental problems.

Firstly, the fourth criterion (eq. 3.4) requires the null hypothesis $H_0 : \beta_S = 0$ not to be rejected. This criterion works fine for rejection of a poor surrogate endpoint showing that β_S is not significant. However, this criterion is not appropriate for validation of a surrogate endpoint as, this statistical test does not prove that $\beta_S = 0$. When the null hypothesis (H_0: $\beta_S = 0$) is not rejected, this may be due to the lack of statistical power due to small sample size in a study. Secondly, the result of the hypothesis test of the fourth criterion cannot prove that the effect of treatment Z on the final outcome T is fully captured by the surrogate endpoint S [80, 82] and in any practical setting, it would be more realistic for a surrogate endpoint to explain a part of a treatment effect on the final endpoint than the full treatment effect. Thirdly, Buyse and Molenberghs showed [79] that Prentice's surrogate endpoint validation criteria are only equivalent to his definition when both the surrogate endpoint S, and the final outcome T are binary outcomes implying that the criteria do not guarantee that a candidate endpoint is a valid surrogate endpoint unless S, T and Z are all binary. Lastly, a potential surrogate endpoint S can only be validated when the treatment Z has significant impact on both outcomes S and T (eq. 3.1, eq. 3.2). Hence, data from a clinical trial where a treatment does not have a statistically significant effect on S and/or T cannot be used to validate a candidate endpoint as surrogate at the individual-level.

All these issues led Freedman et al. [78] to propose an estimation framework, i.e. a quantitative rating of the appropriateness of a candidate endpoint S shifting the attention from hypothesis testing where a is 'yes' or 'no' answer is expected.

3.1.2 Freedman's et al. approach

Freedman et al. [78] proposed the proportion explained (PE) to quantify surrogate relationship as the proportion of the effect of treatment Z on the final outcome T that it is explained by the surrogate endpoint S

$$PE(T,S,Z) = \frac{\beta - \beta_S}{\beta} = 1 - \frac{\beta_S}{\beta}, \tag{3.13}$$

where β is the estimated effect of treatment Z on the final outcome T without correction for S and β_S is the estimates of the effect of treatment Z on the final outcome T with correction for S. The idea behind PE is that if all of the effect is mediated by S i.e., $\beta_S = 0$ then $PE = 1$. PE is a ratio of parameters and its confidence interval can be calculated using delta method or Fieller's theorem. [80]. We expect that a good surrogate endpoint S should have a lower limit of the (1-α)% CI for PE close to 1. Using Filler's theorem the (1-α)% CI of PE is given by

$$1 - \frac{A \pm \sqrt{A^2 - BC}}{B} \tag{3.14}$$

where,

$$
\begin{aligned}
A &= \beta\beta_S - Z_\alpha^2 Cov(\beta, \beta_S) \\
B &= \beta^2 - Z_\alpha^2 Var(\beta) \\
C &= \beta_S - Z_\alpha^2 Var(\beta_S)
\end{aligned}
$$

where Z_α is the (1-$\alpha/2$) percentile of the normal distribution while, the variances of β and β_S can be obtained by fitting the models described in equations (eq. 3.9 and eq. 3.10). In addition, Freedman et al. [78] proposed an efficient way to calculate the covariance between β and β_S.

3.1.2.1 Issues with Proportion Explained ratio

Similarly as the Prentice's approach, there are several problems with PE. The idea behind this quantity is that $PE = 1$ ($\beta_S = 0$) when all of the treatment effect is mediated and $PE = 0$ when there is not mediation ($\beta = \beta_S$). Unfortunately this idea is problematic, as β_S is not necessarily zero when there is full mediation, and β is not always equal to β_S when there is no mediation. Freedman et al. showed that if treatment effect on final outcome T is small and the size of study is not large, the CI of PE tends to be wide, spanning almost the entire [0,1] interval. Hence, it is not possible to draw any inferences from such an interval. Freedman reported [78] that to achieve 80% power for a hypothesis test that the surrogate explains more than 50% of treatment effect the ratio $\beta/SE(\hat{\beta})$ should be ≥ 5. This requirement makes the use of PE infeasible. Also, PE approach is problematic when the assumption of normality for the surrogate endpoint S and final outcome T is incorrect. In this case PE ceases to have a simple interpretation and the validation process stops. Moreover, in practice after validating an outcome as a surrogate endpoint we should be able to make predictions about the treatment effect on the final outcome T. Such predictions should be obtained using the treatment effect on the surrogate endpoint S. It is not clear how it can be achieved within PE setting.

3.1.3 Buyse and Molenberghs's approach

Considering complications with the PE approach, Buyse and Molenberghs [79] introduced two new measures for quantifying strength of the surrogate relationship, relative effect (RE) and the adjusted association (ρ_Z) under the assumption that both outcomes S and T are normally distributed. RE is defined as the ratio of the effects of treatment Z upon the surrogate endpoint S and the final outcome T.

$$RE(T, S, Z) = \frac{\beta}{\alpha} \tag{3.15}$$

Where α and β are the effects of treatment Z measured on the surrogate endpoint and the final outcome respectively. RE can be interpret as the slope of a regression line between β and α and is expected to be equal to 1 if the effect of treatment Z on the surrogate endpoint S is identical to the effect of treatment Z on the final outcome T. If the multiplicative relationship (eq. 3.15) could be assumed and if RE were known exactly, then it could be used to predict the effect of treatment Z on the final outcome T based on an observed effect of treatment Z on the surrogate endpoint S. In reality RE needs to be estimated and the precision of the estimate of RE will be relevant for the precision of prediction.

Buyse and Molenberghs also introduced the association ρ_Z after adjustment for treatment Z to quantify the strength of the individual-level association between the surrogate endpoint S and final outcome T.

$$\rho_Z = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}},} \tag{3.16}$$

where σ_{ST}, σ_{SS}, σ_{TT} are the variances of the covariance matrix of the normal distribution in equation 3.7. RE quantifies the strength of association between S and T at the individual level thus, when $\rho_Z = 1$ then, the effect of treatment Z on the final endpoint T for an individual patient can be perfectly predicted based on the effect of treatment Z on the surrogate endpoint S. However in practice, perfect individual-level surrogacy is unrealistic and it is important to judge whether or not the correlation is considered sufficiently high to validate an outcome as a surrogate endpoint.

Similarly to PE, RE is a ratio of two parameters and its CI can be computed based on the delta method or Fieller's theorem [79]. The CI of ρ_Z quantity can be computed using the general Fisher transformation procedure for correlations or by bootstrapping [80].

3.1.3.1 Problems with Buyse and Molenberghs' approach

As for the previous methods, there are also a few issues with this approach. The problems occurring with the adjusted association and the RE are more applied, whereas the Prentice approach and the PE face fundamental problems. The main assumption of the adjusted association ρ_Z is that both T and S are continuous normally distributed outcomes and in practice, it is simply the correlation between these two outcomes. The estimation of this quantity is straightforward and can easily be interpreted, its CI remains within the unit interval and it is relatively narrow if a study has a reasonably large sample size. However, when the assumption of normality is violated, the calculation of the adjusted association ρ_Z requires using different approaches (for details see [80]). The problem with RE originates from the fact that this quantity is based on data from a single study. If the multiplicative relationship 3.15 is plausible then the ratio should be constant across other clinical trials implying that the relationship between β and α is linear and passes through the origin. However, in practice this can not be proved having data only from a single trial.

Considering all the aforementioned problems of the 'individual-level surrogacy' methods, it becomes clear that data from multiple trials are necessary for the evaluation of surrogate endpoints. Meta-analysis provides a useful framework for combining evidence across relevant studies and has been widely utilised to combine evidence from clinical trials. The next section presents the key meta-analytic methods for trial-level surrogate endpoint evaluation.

3.2 Meta-analytic methods for surrogate endpoint evaluation

Most efficient meta-analytic methods for surrogate endpoint evaluation have a bivariate form (or multivariate for multiple surrogate endpoints) [13, 16, 83–87]. Bivariate meta-analysis can be utilised to model treatment effects measured on the surrogate endpoint and the final outcome jointly, resulting in estimates quantifying the association between the treatment effects. Moreover, it can be used to predict an unreported treatment effect in a study, with the advantage of obtaining the estimates of clinical effectiveness early. When the treatment effect of a new intervention (being under consideration) on the final outcome is not yet reported, regulatory licensing decision for the new treatment can be made conditional on a surrogate endpoint. Bivariate meta-analytic methods can be used to evaluate the trial-level surrogacy patterns as by nature take into account not only the between-studies correlation between the treatment effects measured on the surrogate endpoint and the final outcome, but also all related uncertainty required in decision modelling. As stated in Section 2.21, other methods such as meta-regression do not take into account all relevant uncertainty, whereas the bivariate meta-analytic methods appropriately account for all relevant uncertainty, both at within-study and between-studies levels [16]. Furthermore, without taking the correlation into account (either directly or through some functional relationship between the correlated effects), it is not possible to quantify the strength of the trial-level surrogate relationship. In addition, bivariate meta-analysis can be used to predict a likely treatment effect on the final outcome from the treatment effect measured on the surrogate endpoint. This will allow for making conditional licensing decisions.

A plethora of meta-analytic methods have been proposed to evaluate trial-level surrogacy. This section introduces the key models for the trial-level evaluation of surrogacy patterns. The first formal proposal was from Daniels and Hughes [13] in 1997. They developed a Bayesian two-level meta-analytic method using a linear relationship to describe the trial-level association between the true treatment effects on the surrogate endpoint and the final outcome. The second model that is discussed, is the bivariate random-effects meta-analysis (BRMA) model [88] and an alternative parameterisation of this method proposed by Bujkiewicz et al. [10, 87, 89]. Available are also extensions of the methods to multivariate random-effects meta-analysis (MRMA) [89]. The MRMA approach can be used to evaluate multiple trial level surrogate relationships modelling multiple surrogate endpoints (or treatment effects on multiple surrogate endpoints) simultaneously. The last method that is presented is an approach developed by Buyse et al. [84], who proposed a two-stage model which evaluates surrogacy both in individual and trial levels. It relies on IPD being available from all of the studies and therefore, it can be used only when this is the case.

3.2.1 Daniels and Hughes model

The following method was introduced by Daniels and Hughes [13] to model jointly correlated outcomes or correlated treatment effects under the Bayesian framework. Starting from the within-study variability (eq. 3.17), the observed treatment effects y_{1i} and y_{2i} are assumed to follow a bivariate normal distribution estimating underlying true treatment effects on the surrogate and the final outcomes δ_{1i}, δ_{2i} respectively, for each study i, with corresponding within-study standard deviations σ_{1i}, σ_{2i} and within-study correlation ρ_{wi}

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 \end{pmatrix}\right). \tag{3.17}$$

In comparison to other approaches, the true treatment effects on the surrogate endpoint δ_{1i} are assumed to be fixed effects (which here means they are independent effects across studies). Furthermore, they used a simple linear regression model (eq. 3.18) to describe the association between the true treatment effects on the surrogate endpoint δ_{1i} and the final outcome δ_{2i}.

$$\delta_{2i}|\delta_{1i} \sim N(\lambda_0 + \lambda_1\delta_{1i}, \psi^2), \qquad i = 1, 2, ..., N \tag{3.18}$$

where the slope λ_1, the intercept λ_0 and the conditional variance ψ^2 are used as criteria for surrogate endpoint validation as discussed in Section 3.2.1.1

By combining the within-study (eq. 3.17) and the between-studies distributions (eq. 3.18) we can obtain the marginal distribution:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N \left(\begin{pmatrix} \delta_{1i} \\ \lambda_0 + \lambda_1 \delta_{1i} \end{pmatrix} , \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 + \psi^2 \end{pmatrix} \right) \tag{3.19}$$

Implementing this method under a Bayesian setting, "non-informative" prior distributions can be placed on the fixed effects and the regression parameters:

$$\begin{cases} \delta_{1i} \sim N(0, A_{\mu_{1i}}) \\ \lambda_1 \sim N(0, A_{\lambda_1}) \\ \lambda_0 \sim N(0, A_{\lambda_0}) \end{cases} , \tag{3.20}$$

considering each of $A_{\delta_{1i}}$, A_{λ_1}, A_{λ_0} to be sufficiently large. For the conditional variance ψ^2 they considered three possibilities.

1. DuMouchel prior [90]: $\pi(\psi^2) = \frac{\sigma_c}{(\sigma_c+\psi)^2}\frac{1}{2\psi}$ where σ_c^2 is the harmonic mean of within-study variances σ_{2i} of treatment effects on the final outcome.
2. Shrinkage prior [91]: $\pi(\psi^2) = \frac{\sigma_c^2}{(\sigma_c^2+\psi^2)^2}$ where σ_c is the same as above.
3. Flat prior [92]: $\pi(\psi^2) = \partial\psi^2$

More recently other researchers have suggested that vague prior distributions, such as half normal distributions $N(0, b)I(0,)$ (where $N(0, b)I(0,)$ denotes a truncated at mean normal distribution and b is sufficiently large [93]) or uniform distributions $U(0, b)$, could also be used [10]. If IPD are available, the correlations ρ_{wi} can be obtained by bootstrapping [13]. Otherwise prior distributions can be placed on the within-study correlations, such as, for example, uniform distributions ranging between -1 and 1 $\rho_{wi} \sim U(-1, 1)$ (or weakly informative prior allowing positive/negative values only), or a normal distribution for Fisher's z transformation: $\rho_{wi} = tanh(z)$ where $z \sim N(0, 1)$.

Subgroup analysis with Daniels & Hughes model is used as the standard approach to investigate trial-level surrogacy patterns within-treatment classes in the Chapter 4. Two extensions of Daniels & Hughes model are also proposed to account for differences in the association patterns across classes in the Chapter 4.

3.2.1.1 Criteria for surrogacy

As we mentioned previously, the parameters λ_0, λ_1, ψ^2 play a very important role, as they are used to evaluate surrogacy. A valid surrogate relationship should imply that $\lambda_1 \neq 0$ as slope establishes the association between treatment effects on the surrogate and the final outcome. Subsequently, having $\psi^2 = 0$ implies that δ_{2i} could be perfectly predicted given δ_{1i}. The parameter λ_0 corresponds to the intercept and is expected to be zero for a good surrogate relationship. This ensures that no treatment effect on the surrogate will imply no effect on the final outcome.

3.2.2 BRMA model

Another meta-analytic method that can be used in the context of surrogate endpoints is BRMA model. BRMA models correlated and normally distributed treatment effects y_{1i} and y_{2i} on two outcomes. It was firstly introduced by McIntosh [94] and since then many extensions have been proposed. One of the most popular and practical forms was described by van Houwelingen *et al.* [88] and Riley *et al.* [15]:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 \end{pmatrix}\right) \tag{3.21}$$

$$\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_b \\ \tau_1\tau_2\rho_b & \tau_2^2 \end{pmatrix}\right). \tag{3.22}$$

Here, the treatment effects on the first and the surrogate endpoint y_{1i} and the final outcome y_{2i}, which for example, can be log odds ratios, are assumed to be normally distributed. They estimate the correlated true treatment effects δ_{1i} and δ_{2i} with corresponding within-study variances σ_{1i}^2 and σ_{2i}^2 and within-study correlations ρ_{wi}. At the between-studies level, the true treatment effects are also modelled jointly following a bivariate normal distribution with means (d_1, d_2) and corresponding to the two outcomes between-studies variances τ_1^2 and τ_2^2 and a between-studies correlation ρ_b. In the context of surrogate endpoints the between-studies correlation ρ_b is the main parameter of interest and it quantifies the strength of the trial-level association between the treatment effects on the surrogate endpoint and the final outcome. Equation (3.21) represents the within-study variation and (3.22) is the between-studies model.

The elements of the within-study covariance matrix, σ_{1i}^2, σ_{2i}^2 and ρ_{wi} are assumed known. Whilst the estimates of the variances are easily obtained by taking the square of the standard error for each outcome, the estimates of the within-study correlations between the treatment effects on the two outcomes are more difficult to obtain as they would not be reported in the original articles. When IPD are available, the correlation can be obtained by bootstrapping [13] or alternatively by fitting a regression model for the two outcomes with correlated errors [95]. Other methods of estimating the within-study correlations have been discussed elsewhere and were summarized in Bujkiewicz et al. [10]. Implementing the model in the Bayesian framework the unknown parameters τ_1^2, τ_2^2, d_1, d_2 and ρ_b have to be estimated and therefore, prior distributions should be specified on them. Typically, non-informative prior distributions can be placed on the these parameters: $d_{1,2} \sim N(0, 10^2)$, $\tau_{1,2} \sim U(0, 5)$, to implement the natural constrain of $-1 \leq \rho_b \leq 1$ the Fisher's z transformation can be used as: $\rho_b = tanh(z)$, $z \sim N(0, 1)$.

3.2.2.1 Criteria for surrogacy

The main parameter of interest in this model is the between-studies correlation ρ_b as it quantifies the strength of a trial-level association pattern between the treatment effects on the surrogate endpoint and the final outcomes. For perfect surrogacy, the between-studies correlation should be $\rho_b = \pm 1$. Additionally, it is important to ensure that no treatment effect on the surrogate endpoint will imply no effect on the final outcome - this suggest that the intercept should be very close to zero. Although, BRMA method models the between-studies level without explicitly using a parameter of the intercept, the between-studies parameters of BRMA have been linked with the parameters forming the surrogacy criteria in Section 3.2.1.1 [10]. The slope λ_1, the intercept λ_0 and the conditional variance ψ^2 can be expressed in terms of the parameters of the between-study level of BRMA model as follows:

$$\lambda_1 = \rho_b \frac{\tau_2}{\tau_1} \tag{3.23}$$

$$\lambda_0 = d_2 - d_1 \rho_b \frac{\tau_2}{\tau_1} \tag{3.24}$$

$$\psi_2^2 = \tau_2^2 (1 - \rho_b^2) \tag{3.25}$$

Therefore, we are able to draw inferences for the intercept by expressing it in terms of the between-studies parameters (eq. 3.24). Furthermore, from (eq. 3.25), $\rho_b^2 = 1 - \frac{\psi_2^2}{\tau_2^2}$, which implies that $\rho_b = \pm 1$ when $\psi_2^2 = 0$. This means that the criteria for BRMA model have the same interpretation in terms of surrogacy as the criteria proposed by Daniels and Hughes.

3.2.3 BRMA in product normal formulation

Bujkiewicz et al.[89] proposed and extended in the context of surrogate endpoints[87] BRMA model. They introduced an alternative form of the BRMA model by reparameterising the between-studies level. This alternative model can be used in a very similar way as the model proposed by Daniels and Hughes [13]. More specifically, the between-studies model 3.22 can be presented as a product of univariate conditional normal distributions in a product normal formulation (PNF), whilst the within-study model 3.21 remains exactly the same.

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N \left(\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 \end{pmatrix} \right) \tag{3.26}$$

$$\begin{cases} \delta_{1i} \sim N(\eta_1, \psi_1^2) \\ \delta_{2i} | \delta_{1i} \sim N(\eta_{2i}, \psi_2^2) \\ \eta_{2i} = \lambda_0 + \lambda_1 \delta_{1i} \end{cases} \tag{3.27}$$

As in the BRMA model, y_{1i}, y_{2i} are the observed treatment effects measured by two correlated outcomes (the surrogate endpoint and the final outcome), δ_{1i} and δ_{2i} are the true treatment effects which are correlated and are assumed exchangeable and normally distributed. Therefore, they are modelled as random effects with a linear relationship. Instead of placing independent non-informative prior distributions on all the unknown parameters of model 3.27, relationships between these parameters and the elements of the between-studies covariance matrix (eq. 3.22) are derived to allow for the inter-relationship between the parameters of the two parameterisations and to ensure that the between-studies covariance matrix of the model (eq. 3.22) is positively defined.

$$\psi_1^2 = \tau_1^2,\ \psi_2^2 = \tau_2^2 - \lambda_1^2\tau_1^2,\ \lambda_1 = \frac{\tau_2}{\tau_1}\rho_b \tag{3.28}$$
$$d_1 = \eta_1,\ d_2 = \lambda_0 + \lambda_1 d_1.$$

After establishing these three relationships, we can now place a range of different non-informative prior distributions directly on the following between-studies parameters: $\tau_{1,2} \sim U(0,a)$ or $\tau_{1,2} \sim N(0,b)I(0,)$, $\rho_b = tanh(z)$, $z \sim N(0,1)$ or $\rho \sim U(-1,1)$. The above relationships (eq. 3.28) give implied prior distributions on λ_1, ψ_1^2 and ψ_2^2. Vague prior distributions should also be placed on $\eta_1 \sim N(0,c)$ and $\lambda_0 \sim N(0,c)$ (where a, b and c are sufficiently large).

3.2.3.1 Criteria for surrogacy

The evaluation framework proposed by Daniels & Hughes in the Section 3.2.1.1 applies also to BRMA PNF model. A valid surrogate relationship should imply that $\lambda_1 \neq 0$, the parameter λ_0 is expected to be zero an ψ^2 is expected to be very close to 0.

3.2.4 Multivariate meta-analytic model

Most meta-analytic methods for surrogate endpoint evaluation are designed to evaluate a single surrogate endpoint. However, methods for multiple surrogate endpoints evaluated as joint predictors of clinical benefit or harm have also been proposed. The idea of evaluating multiple surrogate endpoints jointly is not new. In the summary of a National Institutes of Health Workshop on the use of surrogate endpoints, Gruttola et al. [96] suggested the development of models that can allow for modelling multiple surrogate endpoints and/or multiple final clinical outcomes. Approaches for evaluating multiple surrogate endpoints simultaneously were also proposed by Xu and Zeger [97] for time-to-event data modelled jointly with multiple biomarkers measured longitudinally.

BRMA model can be straightforwardly generalised to a bivariate or multivariate model allowing for modelling of multiple outcomes. Bujkiewicz et al.[87, 89] proposed a Bayesian multivariate meta-analytic model aiming to include multiple surrogate endpoints with the potential benefit of reducing the uncertainty of the parameters of interest when making predictions. They showed that the between-study covariance matrix of the multivariate model could also reparameterised in a PNF setting.

Analogous to the bivariate case, at the within-study level $\mathbf{Y_i} = (y_{1i}, y_{2i},...,y_{(N-1)i},y_{Ni})$ are the observed treatment effects on each of the $N-1$ surrogate endpoints and y_{Ni} is the observed treatment effect on the final outcome. These estimates follow a multivariate normal distribution given by:

$$\mathbf{Y_i} \sim MVN(\mathbf{\Delta_i}, \Sigma_i) \tag{3.29}$$

where $\mathbf{\Delta_i} = (\delta_{i1}, ..., \delta_{in})$ is the vector of the true treatment effects on the surrogate endpoints and the final outcome for each study i and Σ_i is within-study variance covariance matrix. As in the bivariate case, the between-studies model makes the same assumption of normality

$$\mathbf{\Delta_i} \sim MVN(\mathbf{D}, T) \tag{3.30}$$

with $\mathbf{D} = (d_1, ..., d_n)$ being the vector of average effects and T an unknown covariance matrix to be estimated.

3.2.5 A two-stage mixed model

3.2.5.1 Full model

In the presence of individual patient data from several RCTs, Buyse et al. [84] considered a two-stage model which evaluates both individual and trial-level surrogacy. They considered two distinct modelling strategies for each stage, the first one is based upon a fixed effects model while the other on random effects.

The linear predictors of the surrogate and the final outcomes are given by:

$$\begin{cases} E(S_{ij}|Z_{ij}) = \mu_{Si} + \alpha_i Z_{ij} \\ E(T_{ij}|Z_{ij}) = \mu_{Ti} + \beta_i Z_{ij} \end{cases} \tag{3.31}$$

where, α_i and β_i are the study-specific fixed treatments effects and μ_{Si} and μ_{Ti} are the intercepts in this setting. These two equations are termed as a full fixed-effects model. Consequently, it is assumed that both outcomes are normally distributed

$$\begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{Si} + \alpha_i Z_{ij} \\ \mu_{Ti} + \beta_i Z_{ij} \end{pmatrix} \quad, \Sigma\right), \tag{3.32}$$

where, the covariance matrix Σ is given by

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \tag{3.33}$$

The second stage (between-studies level) of the model includes:

$$\begin{cases} \mu_{Si} = \mu_S + m_{Si} \\ \mu_{Ti} = \mu_T + m_{Ti} \\ \alpha_i = \alpha + a_i \\ \beta_i = \beta + b_i \end{cases}, \tag{3.34}$$

where m_{Si}, m_{Ti}, a_i, b_i are normally distributed parameters with mean zero and variance-covariance matrix D has the following formulation:

$$D = \left(\begin{array}{cc|cc} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ \hline d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Tb} & d_{ab} & d_{bb} \end{array}\right). \tag{3.35}$$

By combining the equations 3.32 and 3.34 we obtain the mixed-effects model

$$\begin{pmatrix} S_{ij} \\ T_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_S + m_{Si} + (\alpha + a_i)Z_{ij} \\ \mu_T + m_{Ti} + (\beta + b_i)Z_{ij} \end{pmatrix}, \ \Sigma\right). \tag{3.36}$$

To quantify the trial-level association they proposed to use the coefficient of determination defined as:

$$R^2_{\text{trial(f)}} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}, \tag{3.37}$$

where the (f) index indicates that the full model is used to evaluate the surrogacy.

This modelling approach assumes that IPD are available across studies, hence the individual level surrogacy can also be evaluated using the following measure:

$$R^2_{\text{indiv}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}. \tag{3.38}$$

When $R^2_{\text{trial}} = 1$ and $R^2_{\text{indiv}} = 1$ indicate perfect surrogacy both at trial and individual level and thus, the final endpoint can be perfectly predicted using the surrogate endpoint at both levels. However, it is unrealistic to expect this coefficient to be 1 making the need for a realistic threshold for R^2 necessary.

3.2.5.2 Frequentist framework

Implementing the above method in a frequentist setting Buyse et al. proposed the calculation of 95% confidence intervals for both coefficients. The $R^2_{\text{trial(f)}}$ coefficient cannot take the value 1 and remains within the unit interval when D is positive definite. The 95% confidence interval for $R^2_{\text{trial(f)}}$ is given by

$$R^2_{\text{trial(f)}} \pm 1.96\sqrt{\frac{4R^2_{\text{trial(f)}}(1 - R^2_{\text{trial(f)}})^2}{N_{\text{trials}} - 3}} \tag{3.39}$$

,

The variance of $R^2_{\text{trial(f)}}$ is estimated using the delta method (for details see [80, 84]). A value of $R^2_{\text{trial(f)}}$ close to 1 implies strong surrogacy between the treatment effects on the surrogate and final endpoints at the trial-level.

Similarly to the trial-level coefficient, a 95% confidence interval for R^2_{indiv} can be obtain when Σ is positive definite.

$$R^2_{\text{indiv}} \pm 1.96\sqrt{\frac{4R^2_{\text{indiv}}(1 - R^2_{\text{indiv}})^2}{N_{\text{total}} - 3}} \tag{3.40}$$

where N_{total} represents the number of patients in a study. To estimate the variance of R^2_{indiv} the delta method can be used. A value of R^2_{indiv} close to 1 indicates strong surrogacy at individual level and thus, the final endpoint can be perfectly predicted using the surrogate endpoint at this level. However in general, it is unrealistic to expect these coefficients to be close to 1 making the need for a realistic threshold for R^2 necessary.

Chapter 4

Improving the validation of surrogate endpoints in a specific treatment class, whilst borrowing information across classes

4.1 Introduction

This Chapter aims to improve the trial-level validation of surrogate endpoints within a specific class of treatment in disease areas where trial-level surrogacy patterns vary across treatment classes, by proposing novel methodology.

Potential surrogate endpoints have been investigated in clinical trials in a number of disease areas. These candidate endpoints need to be validated [98] as reliable predictors of clinical benefit. This can be done by exploring the three levels of association as described in Section 1.3. Evaluation of the association between treatment effects on a candidate surrogate endpoint and a final outcome requires data from a number of RCTs and it can be investigated by carrying out a bivariate meta-analysis. A strong association pattern between treatment effects on a candidate endpoint and treatment effects on a final outcome implies the existence of a trial-level surrogate relationship between the two outcomes and validates the candidate endpoint as a suitable surrogate endpoint.

Traditionally, trial-level surrogate relationships between treatment effects on a surrogate endpoint and treatment effects on a final outcome have been investigated in given a disease area using data from all trials regardless of treatment classes or limiting data to trials of the same class of treatments. For instance, in aCRC, PFS, tumor response (TR) or time to progression (TTP) have been investigated as potential surrogate endpoints for OS [11, 12, 99, 100]. In previous work, Buyse et al. [12] found a strong association between treatment effects on PFS and OS in this disease area, by including in their meta-analysis studies on one treatment class only (cytotoxic treatments).

However, surrogacy association patterns may differ across classes of treatment when targeted treatments are applied to subgroups of population with unique tumour characteristics (mutations). These potential differences can simply be investigated by performing subgroup analysis using a bivariate meta-analytic model. This type of analysis is very practical when there are sufficient data within treatment classes, or a specific class of interest. However, the analysis may fail to estimate a trial-level surrogate relationship effectively when data are limited in terms of the number of studies, resulting in estimates of the parameters describing surrogate relationships obtained with considerable uncertainty [101]. For example, after the introduction of targeted treatments in aCRC, Giessen et al.[99] investigated the association patterns across treatment classes by performing subgroup analysis. They inferred that further research was needed at that point to establish surrogate relationship between treatment effects on PFS and treatment effects on OS in the new treatment classes consisting of targeted treatments as the data were very sparse.

In this Chapter, we use subgroup analysis adopting a model proposed by Daniels and Hughes (see details in Section 3.2.1) as the standard approach to investigate association patterns (surrogate relationships) within each treatment class. To improve the validation of surrogate endpoints in a particular treatment class and address the limitations of subgroup analysis, we developed two meta-analytic methods allowing for trial-level validation of surrogate endpoints within each treatment class, whilst borrowing information across classes. Specifically, instead of carrying out subgroup-analysis, we propose two extensions of the model proposed by Daniels and Hughes [13] by adding another level to its hierarchical structure.

This additional level accounts for differences in surrogate relationships between treatment classes but at the same time assumes some level of similarity between them. The first extension allows for full borrowing of information for surrogate relationships across treatment classes assuming exchangeability for the parameters describing the surrogate relationships. The second one relaxes this assumption, as it allows for partial borrowing of information for the surrogate relationships across treatment classes by assuming partial exchangeability [102]. In this model one or more of the parameters describing the surrogate relationships can be either exchangeable or non-exchangeable giving more flexibility when the assumption of exchangeability is not reasonable. The proposed methodology and the results of the data analysis, described in this chapter, have also been published in Statistics in Medicine [103].

To investigate whether the proposed methods can improve the trial-level validation of surrogate endpoints compared to subgroup analysis, we carried out a extensive simulation study. The aim of the simulation study was to assess whether or not the proposed methods result in less biased and more precise estimates of the parameters describing the surrogate relationship and improve the accuracy and the precision of the predictions of the true treatment effects on the final outcome. Furthermore, we applied the methods (subgroup analysis with the standard model and the two proposed methods) to a data example in aCRC which consist of three treatment classes. As discussed previously, Buyse et al. [12] found a strong association between treatment effects on PFS and OS in this disease area, by including in their meta-analysis studies on one treatment class only (cytotoxic treatments). However, more recently Ciani et al.[11] found sub-optimal surrogate relationship between treatment effects for PFS-OS pair of outcomes concluding that trial-level surrogate relationships could vary across treatment classes in aCrC (or equally in other diseases) and a surrogate relationship observed in a specific treatment class may not directly apply across other treatment classes or lines of treatment. This may be particularly important for targeted treatments used only in subsets of population. For instance in aCRC, anti-EGFR treatments are recommended for patients without a KRAS/panRAS mutation as these mutations are associated with resistance to the anti-EGFR therapies [104, 105] and the association pattern might be different for this particular treatment class in this subset of population with this unique characteristic.

The existing and the proposed modeling approaches are introduced in Section 4.2, the simulation study is presented in Section 4.3. A detailed description of the data-set and the results of the evaluation on PFS-OS and TR-PFS pairs can be found in section 4.4. The Chapter concludes with a discussion in Section 4.5.

4.2 Methods for trial-level surrogate endpoint evaluation across treatment classes

This section presents existing and the proposed approaches to evaluate association patterns of potential surrogate markers across treatment classes. Firstly, subgroup analysis with Daniels and Hughes model is presented as the standard approach to investigate potential differences in surrogate relationships in each treatment class separately. Secondly, two alternative Bayesian meta-analytic methods are introduced as alternative approaches to subgroup analysis. They allow for the association patterns to vary across classes taking advantage of exchangeability . The use of exchangeability across treatment effects consisting of multiple studies has a long history in evidence synthesis. Recently, exchangeability has been assumed across certain nodes, doses levels or treatment groups in hierarchical network meta-analysis models [106–109]. These approaches take advantage of the attractive statistical properties of exchangeability [110–112] leading to more precise inferences for the parameters of interest as they allow strength to be borrowed across/within groups.

4.2.1 Subgroup analysis with a standard surrogacy model

The simplest way to investigate differences in trial-level surrogate relationships within treatment classes is to perform subgroup analysis using a bivariate meta-analysis. In this chapter, the Bayesian meta-analytic model proposed by Daniels and Hughes [13] was used as evaluation method.

To perform subgroup analysis across treatment classes we applied the model (eq. 4.1 and eq 4.2) to subsets of data that consist of only one class of treatment j, examining surrogate relationship in each subgroup separately, taking motivation from similar analyses in clinical trials [101, 113].

$$\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} \sim N \left(\begin{pmatrix} \delta_{1ij} \\ \delta_{2ij} \end{pmatrix} , \begin{pmatrix} \sigma^2_{1ij} & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma^2_{2ij} \end{pmatrix} \right) \tag{4.1}$$

$$\delta_{2ij}|\delta_{1ij} \sim N(\lambda_{0j} + \lambda_{1j}\delta_{1ij}, \psi^2_j) \tag{4.2}$$

Equation (4.1) corresponds to the within-study model where y_{1ij}, y_{2ij} are the observed treatment effects on surrogate endpoint and on the final outcome for each study i and treatment class j. At the between-studies level (4.2) a linear relationship between the true effects on the surrogate δ_{1i} and the true effects on the final outcome describes the trial-level surrogate relationship of the j^{th} treatment class. In practice this linear relationship can be used to predict the true effect on the final outcome from a known true effect on the surrogate in a new study i and from a specific treatment class j. The parameters λ_{0j}, λ_{1j}, ψ^2_j correspond to the intercept, the slope and the conditional variances of the linear relationship of the j^{th} treatment class and measure the shape of the relationship and the strength of the association between the true treatment effects on the surrogate endpoint and the effects on the final outcome. These parameters form the criteria of surrogate endpoint evaluation as described in section 3.2.1.1 and in section 4.2.1.1 and should be met for each treatment class separately.

Implementing this model in the Bayesian framework, no prior knowledge was assumed about the parameters describing the surrogate relationships, thus non-informative priors were placed on all the unknown parameters. This allows the data to dominate the posterior distribution even if the data-set is relatively small, hence the following prior distributions were used in the simulation study and the motivating example: $\delta_{1i} \sim N(0, 100)$, $\lambda_0 \sim N(0, 100)$, $\lambda_1 \sim N(0, 100)$, $\psi \sim N(0, 100)I(0,)$.

This kind of analysis is very practical when association patterns in a given disease area are different and the treatment classes consist of many studies. By performing subgroup analysis, potential differences in the association patterns across treatment classes can be explored.

4.2.1.1 Criteria for surrogacy

In this chapter the evaluation framework proposed by Daniels and Hughes[13] was used to investigate surrogate relationship within each treatment class. As described in section 3.2.1.1 a strong association (surrogate relationship) requires the slope to be non zero, as it establishes the association between treatment effects on the surrogate and the final outcome. The conditional variance should be approximately zero as this implies that δ_{2i} could be perfectly predicted given δ_{1i}. The parameter λ_0 corresponding to the intercept is expected to be zero which ensures that no treatment effect on the surrogate endpoint will imply no effect on the final outcome. These three simple rules will be referred to as surrogacy criteria in this chapter. A simple way to examine these surrogacy criteria is to check whether or not zero is included in the 95% credible intervals (CrIs) of λ_0, λ_1 and to compute the Bayes factor for the hypothesis H_1: $\psi^2 = 0$. The model with $\psi^2 = 0$ is a nested model within the standard model [114], so in order to compare these models, Bayes factors can be computed using the Savage Dickey density ratio [115]. To implement the Savage Dickey density ratio, proper prior distributions for ψ are needed. A moderately informative half normal prior distribution $N(0, 2)I(0,)$ was used for the conditional standard deviation. The R code of the Bayes factors calculation can be found in the Appendix B.6. A strong association pattern (surrogate relationship) requires zero to be included in the CrI of λ_0, zero not to be included in the CrI of λ_1 and the Bayes factor of ψ^2 to be greater than 3.3 [39].

4.2.2 Hierarchical model with full exchangeability (F-EX)

When subgroup analysis is used to investigate the trial-level surrogate relationships within treatment classes the validation process may fail due to limited data resulting in estimates of the parameters describing surrogate relationships obtained with considerable uncertainty [101]. Our first approach (Full-exchangeability (F-EX) model) extends the standard model 4.2.1 by adding another level of hierarchy to its hierarchical structure. By doing this, the proposed model accounts for differences in trial-level association patterns across different treatment classes [116–118]. The method can be applied to continuous and normally distributed aggregate data or it can be used for binomial or time to event data when they are transformed to the log odds ratio scale or log hazard ratio scale respectively. Similarly as in the standard model, at the within-study level we assume that correlated and normally distributed observed treatment effects y_{1ij} and y_{2ij} (logHR or logOR) in each study i and treatment class j estimate the true treatment effects δ_{1ij} and δ_{2ij} on the surrogate and final outcomes respectively.

$$\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_{1ij} \\ \delta_{2ij} \end{pmatrix}, \begin{pmatrix} \sigma_{1ij}^2 & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma_{2ij}^2 \end{pmatrix}\right) \tag{4.3}$$

$$\delta_{2ij}|\delta_{1ij} \sim N(\lambda_{0j} + \lambda_{1j}\delta_{1ij}, \psi_j^2) \tag{4.4}$$

$$\lambda_{0j} \sim N(\beta_0, \xi_0^2), \lambda_{1j} \sim N(\beta_1, \xi_1^2) \tag{4.5}$$

The parameters σ_{1ij}^2, σ_{2ij}^2, ρ_{wij} correspond to the within-study variances and within-study correlations for each study i in treatment class j. The observed estimates y_{1ij}, y_{2ij}, σ_{1ij}, σ_{2ij} are aggregate data extracted from systematic review RCTs whilst, the within-study correlations ρ_{wij} can be calculated using a bootstrap method from IPD [119]. The true effects δ_{1ij} on the surrogate endpoint are modelled as fixed effects.

In contrast to subgroup analysis with the standard model, F-EX is fitted to full data-set and not only to a specific subgroup of data. It also assumes unique linear relationships between true treatment effects on the surrogate endpoint and the final outcome across treatment classes. Each relationship between the true effects on the surrogate endpoint δ_{1ij} and the final outcome δ_{2ij} is described by the same linear model (eq. 4.4) as in Daniels and Hughes model, where λ_{0j} denotes the intercept of the j^{th} treatment class and λ_{1j} establishes the relationship between treatment effects on surrogate and final outcomes within the treatment class j. Furthermore, the intercepts λ_{0j} and the slopes λ_{1j} are assumed exchangeable across treatment classes leading to full borrowing of information across treatment classes (eq. 4.5). This is implemented by assuming exchangeability of these parameters and placing common normal distributions on λ_{0j} and λ_{1j} with means and variances β_0, ξ_0^2 and β_1,

ξ_1^2. Hence, when a slope and an intercept are estimated within a specific class, this assumption allows for full borrowing of information across all treatment classes. To evaluate whether a candidate endpoint is considered a valid surrogate endpoint in a given treatment class, all three surrogacy criteria need to be met for this particular class.

Implementing this model in the Bayesian framework, non-informative prior distributions were placed on the unknown parameters of the model such as: $\beta_0, \beta_1 \sim N(0, 100)$ and $\xi_0, \xi_1 \sim N(0, 100)I(0,)$, $\delta_{1ij} \sim N(0, 100)$ and $\psi_j \sim N(0, 100)I(0,)$. The WinBUGS code of the model can be found in the Appendix B.4.

Overall, F-EX model extends the standard model (described in section 3.2.1) by including an additional layer of hierarchy to the linear relationship (eq. 4.5) between true effects on the surrogate and the final outcome, assuming that slopes and intercepts are exchangeable across treatment classes.

The exchangeable estimates, however, are shrunk towards the means β_0, β_1 and the amount of shrinkage depends on the number of studies within each class, the between treatment class heterogeneity [102] and the number of treatment classes. Although these statistical properties are very attractive in terms of potential reduction of uncertainty around the parameters of interest, they are advantageous only when the assumption of exchangeability is reasonable, otherwise there is a danger of excessive shrinkage.

4.2.3 Hierarchical model with partial exchangeability (P-EX)

As discussed in the previous section, when the assumption of exchangeability of the intercepts and the slopes is not reasonable, F-EX may give biased estimates due to excessive shrinkage towards the pooled mean [110–112]. Hence, F-EX can be extended further allowing for tailored borrowing of information by assuming partial exchangeability for a/some of the parameters of interest, similarly as in the method proposed by Neuenschwander et al. [102]. Partial-exchangeability (P-EX) model relaxes the assumption of exchangeability allowing a parameter of interest in a specific class to be either exchangeable with all or some of the parameters from other treatment classes or non-exchangeable throughout the estimation process. This model is more flexible compared to F-EX, in particular in data scenarios where the assumption of exchangeability is not fully reasonable for some of the treatment classes.

The within-study level (eq. 4.6) of P-EX is exactly the same as in F-EX where, y_{1ij}, y_{2ij} are the treatment effects on the surrogate endpoint and the final outcome in study i and treatment class j. These effects follow a bivariate normal distribution with mean values corresponding to the true treatment effects δ_{1ij} and δ_{2ij} on the two outcomes.

$$\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_{1ij} \\ \delta_{2ij} \end{pmatrix}, \begin{pmatrix} \sigma^2_{1ij} & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma^2_{2ij} \end{pmatrix}\right) \tag{4.6}$$

$$\delta_{2ij}|\delta_{1ij} \sim N(\lambda_{0j} + \lambda_{1j}\delta_{1ij}, \psi^2_j) \tag{4.7}$$

$$\lambda_{0j} \sim N(\beta_0, \xi^2_0)$$

$$\lambda_{1j} = \begin{cases} \lambda_{1j} \sim N(\beta_1, \xi^2_1) & \text{if } p_j = 1 \\ \lambda_{1j} \sim N(0, 100) & \text{if } p_j = 0 \end{cases} \tag{4.8}$$

In the between-studies model (eq. 4.8), the parameters of slopes are modelled in a different way compared to F-EX model. In this approach two possibilities arise for these parameters for each treatment class j. When $p_j = 1$ the parameter λ_{1j} in a specific treatment class j can be exchangeable with some or all the parameters of the slopes from the other treatment classes via an exchangeable component (i.e. follows a common normal distribution with other slopes as in F-EX model). On the other hand, when $p_j = 0$ the slope can be non-exchangeable with any slopes from the other treatment classes. In this case a vague prior distribution can be placed on the parameter, as in the standard model.

The method uses both components during the estimation process of λ_{1j}. Specifically, in each MCMC iteration, the sampler chooses between the two components by using a Bernoulli distribution $p_j \sim Bernoulli(\pi_j)$. By calculating the posterior mean of this Bernoulli distribution we derive the mixture weights of each treatment class and we are able to evaluate the degree of borrowing of information throughout the estimation process. The main advantage of this method is that the degree of borrowing of information is inferred based on the similarity of the data. The hyper-parameters π_j of the Bernoulli prior distribution can be either fixed or, in a fully Bayesian framework, they can follow a prior distribution, for example, a Beta distribution $\pi_j \sim Beta(1, 1)$.

In a special case where $p_j = 1$ for all treatment classes, P-EX model reduces to full exchangeability model as it uses only the exchangeable component. Having $p_j = 0$ for all treatment classes makes the P-EX model equivalent to subgroup analysis using the standard model as only the non-exchangeable component is used to estimate λ_{1j} in this case. Implementing the model in the Bayesian framework the same non-informative prior distributions can be placed on the unknown parameters β_0, β_1, ξ_0, ξ_1, δ_{1ij} as in F-EX model. The WinBUGS code of the model can be found in the Appendix B.5.

4.2.4 Cross-validation

One of the main aims of this chapter was to explore whether the methods we proposed in the sections 4.2.2, 4.2.3, improve the predictions of the true treatment effects on the final outcome (by reducing bias and/or uncertainty) compared to subgroup analysis using a standard surrogacy model. To evaluate this, a cross-validation procedure was carried out. It is a similar approach to the 'leave-one-study-out' procedure that was described by Daniels & Hughes [13] and it is repeated as many times as the number of studies in the data-set.

During the cross-validation procedure, for each study i $(i = 1, .., N)$, the treatment effect on the final outcome y_{2i} is omitted and assumed unknown. This effect is then predicted from the effect on the surrogate endpoint and by taking into account the treatment effects on both outcomes from the remaining studies. In a Bayesian framework it can be achieved by performing MCMC simulation. In a simulated data scenario, this procedure can be used to draw inferences about predicting the true effect on the final endpoint δ_{2i} in a 'new' study i, however, in real data scenarios the true values of the treatment effects are unknown and therefore, we can only compare the predicted values of the true treatment effect with the values of the observed treatment effect y_{2i}. In this situation the mean predicted effect is equal to the true effect $\hat{\delta}_{2i}$ predicted by MCMC simulation and the variance of the predicted effect is equal to $\sigma^2_{2i} + var(\hat{\delta}_{2i}|y_{1i}, \sigma_{1i}, y_{1(-i)}, y_{2(-i)})$, where $y_{1,2(-i)}$ denote the observed treatment effects from the remaining studies without the study that is omitted in i^{th} iteration [13]. Then it can be examined whether the 95% predictive interval (constructed using the variance) include the observed value of the treatment difference y_{2i} on the final outcome.

4.3 Simulation study

We carried out a simulation study to assess the performance of the methods and to compare them with subgroup analysis conducted using Daniels and Hughes model. We evaluated the performance of the methods in distinct data scenarios generated assuming different strengths of association within classes, different levels of similarity of the association patterns across classes and different number of studies per class. We evaluated the models' ability to identify treatment classes with strong association patterns and to make predictions of the treatment effect on the final outcome in a new study from a treatment effect measured on the surrogate endpoint.

4.3.1 Simulation scenarios and generation process

Nine data scenarios were simulated with 1000 replications per scenario. In each scenario we simulated 5 treatment classes varying the number of studies. Different heterogeneity patterns in each treatment class were assumed hence, in order to have a control over such heterogeneity patterns when simulating the data, an assumption about the distribution of the true effects both on the surrogate and the final endpoints was made. The standard model by Daniels & Hughes assumes fixed effect for the true effects on the surrogate endpoint (no common distribution) making difficult to control the heterogeneity patterns when simulating the data. To avoid this issue, the data were simulated using a product normal formulation of bivariate random effect meta-analysis (PNF of BRMA) (discussed in Section 3.2.3), assuming normal random effects on the surrogate endpoint. Apart from this additional assumption, this method is the same as Daniels & Hughes model using a bivariate normal distribution to describe the within-study variability and a linear relationship to model the association between the surrogate and the final outcome. However, simulating data from this model can lead to results obtained with increased uncertainty, as the models used to analyse the data make fewer distributional assumptions.

To generate the data, the following steps were pursued:

1. Set the number of classes $N = 5$
2. Create three designs (a short description about the designs can be found below)
3. Create three sets of scenarios: two with fixed number of studies (n_j=16 and n_j=8, j=1,...,5) per treatment class and one with unbalanced classes ($n_1 = 4$, $n_2 = 8$, $n_3 = 6$, $n_4 = 10$, $n_5 = 7$). We applied the three sets of scenarios to each design having in total 9 scenarios (3 designs × 3 sets = 9 scenarios).
4. Simulate the true effects of the surrogate endpoint and the final outcome using the following distributions: $\delta_{1ij} \sim N(\eta_{1j}, \psi_{1j}^2)$, $\delta_{2ij}|\delta_{1ij} \sim N(\eta_{2ij}, \psi_{2j}^2)$ with $\eta_{2ij} = \lambda_{0j} + \lambda_{1j}\delta_{1ij}$ and $\psi_{1j} = \frac{\psi_{2j}}{|\lambda_{1j}|\sqrt{(1/\rho_{bj}^2)-1}}$
5. Simulate the estimates of treatment effect from the following distribution for each class j separately: $\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_{1ij} \\ \delta_{2ij} \end{pmatrix}, \begin{pmatrix} \sigma_{1ij}^2 & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma_{2ij}^2 \end{pmatrix}\right)$

The values of the parameters are listed in Table 1 and a short description of each design can be found below:

Table 4.1: Simulation designs

1^{st} design	2^{nd} design	3^{rd} design
$\lambda_{11} = 0.40$, $\rho_{b1} = 0.89$	$\lambda_{11} = 0.60$, $\rho_{b1} = 0.93$	$\lambda_{11} = 0.40$, $\rho_{b1} = 0.90$
$\lambda_{12} = 0.45$, $\rho_{b2} = 0.90$	$\lambda_{12} = 1.55$, $\rho_{b2} = 0.99$	$\lambda_{12} = 0.50$, $\rho_{b2} = 0.70$
$\lambda_{13} = 0.50$, $\rho_{b3} = 0.91$	$\lambda_{13} = 1.60$, $\rho_{b3} = 0.99$	$\lambda_{13} = 0.60$, $\rho_{b3} = 0.93$
$\lambda_{14} = 0.55$, $\rho_{b4} = 0.92$	$\lambda_{14} = 1.65$, $\rho_{b4} = 0.99$	$\lambda_{14} = 0.70$, $\rho_{b4} = 0.75$
$\lambda_{15} = 0.60$, $\rho_{b5} = 0.93$	$\lambda_{15} = 1.70$, $\rho_{b5} = 0.99$	$\lambda_{15} = 0.80$, $\rho_{b5} = 0.95$
$\lambda_{0j} = 0$	$\lambda_{0j} = 0$	$\lambda_{0j} = 0$
$\sigma_{1ij,2ij} = 0.1$	$\sigma_{1ij,2ij} = 0.1$	$\sigma_{1ij,2ij} = 0.1$
$\rho_{wij} = 0.4$	$\rho_{wij} = 0.4$	$\rho_{wij} = 0.4$
$\psi_{2j} = 0.08$	$\psi_{2j} = 0.08$	$\psi_{21,23,25} = 0.08$ $\psi_{22,24} = 0.30$
$\eta_{1j} = 0.3$	$\eta_{1j} = 0.3$	$\eta_{1j} = 0.3$

Design 1:

In the first design, the aim was to illustrate the properties of exchangeability. Five treatment classes were generated having high degree of similarity for their slopes and intercepts. We simulated data assuming strong association (see surrogacy criteria in Section 4.2.1.1) in each individual class but weak overall.

Design 2:

The second design illustrates the situation where the assumption of exchangeability is in doubt for one of the parameters describing the surrogate relationships in a particular class. To achieve this, we simulated one treatment class with very different slope compared to the other four treatment classes classes. Similarly as in the first scenario, strong association patterns in each individual class were assumed.

Design 3:

In the last design we focus on the association patterns of strengths that vary across treatment classes, investigating whether the proposed methods can estimate a strong association pattern better compared to subgroup analysis with the standard model and whether they can distinguish between the different association patterns despite borrowing of information across treatment classes. To achieve this, three out of five treatment classes were generated with strong association and the remaining two classes with a weak association.

4.3.2 Estimands and performance measures

The primary estimand of the simulation study was the parameter of the slope λ_{1j}. Subgroup analysis with the standard model and the proposed hierarchical models make different assumptions about this parameter, hence, by estimating the slopes across the simulated scenarios we assessed the performance of these models under different settings. The second group of estimands of the simulation study were the predicted true treatment effects δ_{2ij} on the final outcome, in each study i and each treatment class j. These effects can be predicted by carrying out a cross-validation procedure in each data scenario. In the simulation study, the true value of true treatment effect on the final endpoint δ_{2ij} was known, as it had been simulated, therefore here we were able to compare the predicted effects with the true values of true treatment effects (in real data scenarios we compare the predicted effect with the observed effect). The last estimand of the simulation study reflects the ability of the models to estimate a strong association pattern in each treatment class. To assess this, we estimated 95% CrIs of slopes, 95% CrIs of intercepts and Bayes factors of the conditional variances in each treatment class, and used the three surrogacy criteria (Section 4.2.1.1) proposed by Daniels and Hughes (a strong association pattern can be inferred when the three surrogacy criteria are satisfied).

To evaluate the performance of the models we calculated and monitor the following measures across all the simulated scenarios: coverage probability of the 95% CrIs of λ_{1j} and the 95% predictive intervals of δ_{2ij}, absolute bias and root mean square error (RMSE) of $\hat{\lambda}_{1j}$ and $\hat{\delta}_{2ij}$. Furthermore, to investigate potential decrease in the degree of uncertainty of the estimates as a result of borrowing of information across treatment classes, we calculated ratios of the width of the 95% CrIs. Two width ratios were defined and used across the simulation study. The first width ratio corresponded to F-EX model and it was defined as: $w_{\lambda_{1j}^{F-EX}}/w_{\lambda_{1j}^{subgr}}$, the ratio of the widths of the CrIs of λ_{1j} from F-EX to the width of the CrIs of λ_{1j} from subgroup analysis using the standard model. The second width ratio corresponded to P-EX model and it was defined as: $w_{\lambda_{1j}^{P-EX}}/w_{\lambda_{1j}^{subgr}}$, the ratio of the widths of the CrIs of λ_{1j} from P-EX to the width of the CrIs of λ_{1j} from subgroup analysis with the standard model. Similarly, two width ratios were calculated and monitored across scenarios about the predictive intervals. We evaluated the ability of subgroup analysis and the proposed hierarchical methods to estimate a strong association pattern, by calculating probabilities of estimating a strong association pattern, in each treatment class across the scenarios and the models.

4.3.3 Results

This section presents the results of simulation study, reporting the average coverage probabilities across treatment classes of the CrIs of λ_{1j} and δ_{2ij} in each scenario, the average absolute bias and RMSE of $\hat{\lambda}_{1j}$ and $\hat{\delta}_{2ij}$ across treatment classes, the average width ratios of λ_{1j} and δ_{2ij} across treatment classes in each scenario and the probability to estimate a strong association pattern by fitting each model. The Figures and the Tables in the results section list the performance of the posterior means of $\hat{\lambda}_{1j}$, the performance of the posterior means of $\hat{\delta}_{2ij}$, and the probabilities of estimating a strong association pattern (see definition in Section 4.2.1.1) in each class across methods. Detailed results for the performance of $\hat{\lambda}_{1j}$ and $\hat{\delta}_{2ij}$ in each class separately were listed in the Appendix (see sections B.1 and B.2).

4.3.3.1 Slope λ_{1j}

Figure 4.1 presents the results across the nine scenarios reporting averages of the performance measures over the five classes of treatment and across 1000 replications. Table 4.2 displays the values of the mixture weights p_j obtained from P-EX. These values are used to illustrate the degree of borrowing of information in each treatment classes.

Starting from the first design (scenarios 1, 2, 3), where the treatment classes were very similar in terms of patterns (similar slopes), F-EX and P-EX were superior compared to subgroup analysis as they gave estimates of slopes with lower average absolute bias, average RMSE and reduced uncertainty (narrower 95% CrIs) due to borrowing of information across classes. Specifically when 16 studies where available in each treatment class, F-EX and P-EX models performed marginally better compared to subgroup analysis in term of average absolute bias and average RMSE. However, in the scenario with 8 studies in each treatment class and the scenario with unbalanced treatment classes, the proposed methods resulted in estimates of λ_{1j} with significantly lower average absolute bias and RMSE across treatment classes compared to subgroup analysis. Additionally, substantial improvement was observed in the precision of the estimates of λ_{1j} across classes. In the scenario with 8 studies in each treatment class, the average width ratio of F-EX versus subgroup analysis was 0.60 and the average width ratio of P-EX versus subgroup analysis was 0.61. In the scenario with unbalanced treatment classes, the average width ratio of for F-EX was 0.50 and the average width ratio for P-EX was 0.51. Focusing on the results of Table 4.2, P-EX model achieved almost the same level of borrowing of information as F-EX model in the first three scenarios (design 1) - the mixtures weights were very close to 1 across treatment classes (Table 4.2). Overall, the proposed hierarchical models performed better compared to subgroup analysis but the difference was more pronounced in the scenarios with smaller number of studies.

Moving to the second design (scenarios 4, 5, 6), where the exchangeability assumption was not reasonable for one of the classes, P-EX model yielded the most robust results. The model resulted in estimates with the smallest average absolute bias and average RMSE. It is important to highlight that P-EX reduced the degree of borrowing of information in the 1st treatment class were the true slope was distinctly different, whilst achieved almost the same degree of borrowing of information across the remaining classes. Table 4.2 displays the mixture weights in these scenarios across all treatment classes. On the other hand, F-EX achieved inferior performance compared to P-EX in terms of average absolute bias and average RMSE and average width ratio in the scenarios with 16 and 8 studies per treatment class. Furthermore, F-EX performed poorer compared to all the other methods in the 6th scenario with unbalanced and relatively small number of studies per class, leading to more biased results. This indicates that F-EX model was not an appropriate modeling approach when the assumption of exchangeability was not reasonable. Subgroup analysis using the standard model achieved a reasonable performance only in the forth scenario where there were sufficient data.

In the third design (scenarios 7, 8, 9), where the strength of the association patterns varied, the proposed models achieved superior performance compared to subgroup analysis resulting in lower average absolute bias and lower average RMSE, similarly as in the first three scenarios. Furthermore, they gave estimates of λ_{1j} with reduced uncertainty, as the average width ratios were 0.79, 0.67 and 0.56 for both methods in each scenarios respectively.

The performance of the models varied in terms of the coverage probability of the 95% CrIs of λ_{1j} across all scenarios. In the scenarios 1, 4 and 7 where the number of studies per class was relatively high, the models achieved 95% coverage probabilities. However, in the scenarios where the number of studies was smaller the coverage probability was higher, as the 95% CrIs were wider (more conservative) due to the sparsity of the data and likely to the fact that the model we used in the generation process makes an additional distributional assumption compared to Daniels and Hughes model and the proposed methods.

Figure 4.1: Absolute bias of $\hat{\lambda}_{1j}$ averaged over the 5 treatment classes (first row), coverage of λ_{1j} averaged over the 5 classes (second row), RMSE of $\hat{\lambda}_{1j}$ averaged over the 5 classes (third row) and width ratios of λ_{1j} averaged over the 5 classes (forth row)

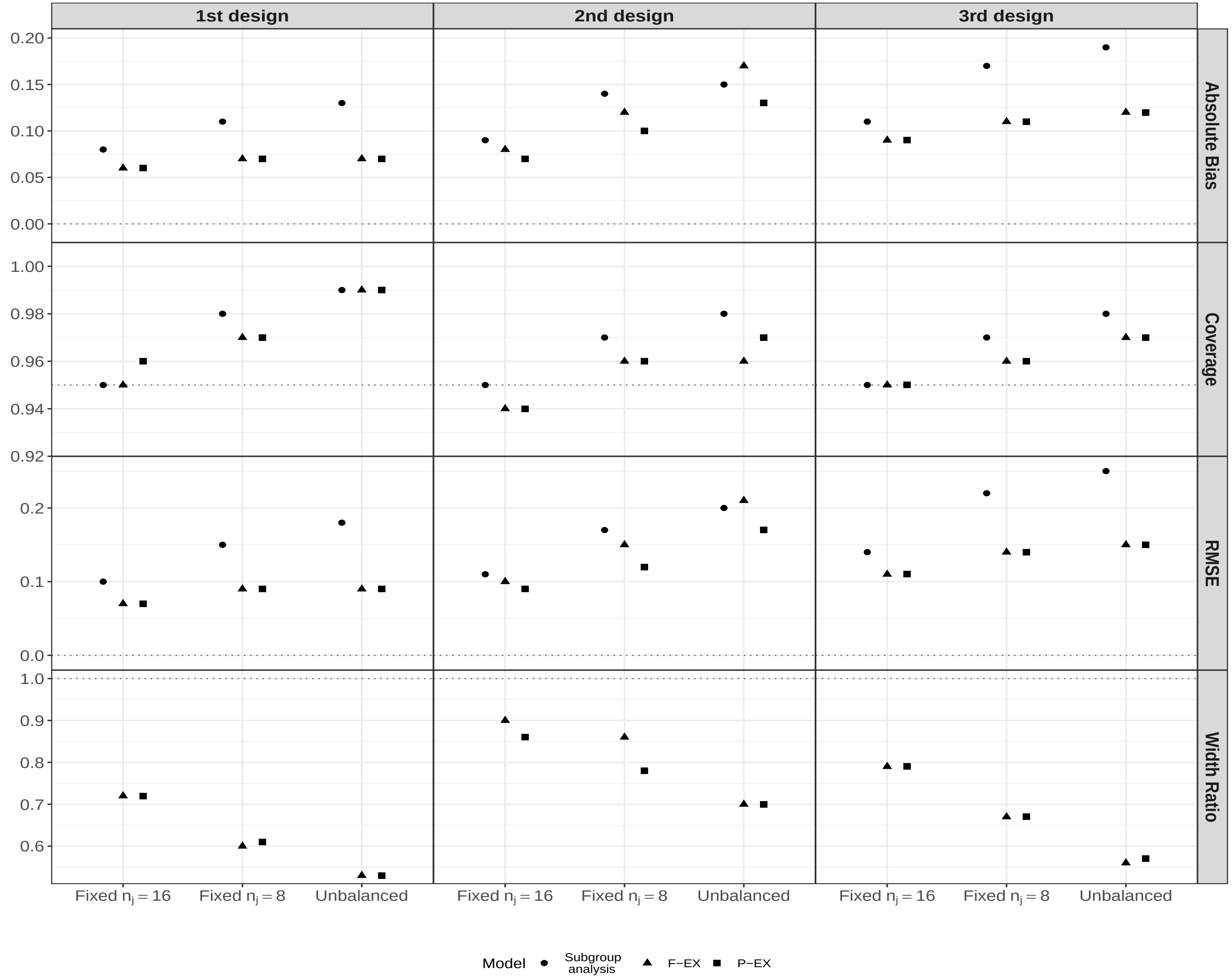


Table 4.2: Mixture weights p_j across all scenarios

Scenarios	1st treatment class	2nd treatment class	3rd treatment class	4th treatment class	5th treatment class
1^{st} scenario	0.99	0.99	0.99	0.99	0.99
2^{nd} scenario	0.99	0.99	0.99	0.99	0.99
3^{rd} scenario	0.99	0.99	0.99	0.99	0.99
4^{th} scenario	0.56	0.98	0.98	0.98	0.98
5^{th} scenario	0.31	0.88	0.88	0.88	0.88
6^{th} scenario	0.80	0.98	0.98	0.98	0.98
7^{th} scenario	0.99	0.99	0.99	0.98	0.98
8^{th} scenario	0.98	0.99	0.99	0.98	0.98
9^{th} scenario	0.98	0.98	0.98	0.98	0.98

4.3.3.2 Predictions of the true treatment effect on the final outcome $\hat{\delta}_{2ij}$

Figure 4.2 shows the results from the cross-validation procedure which resulted in predictions of the true treatment effects ($\hat{\delta}_{2ij}$) and 95% predictive intervals of the true effects δ_{2ij}. It presents the same measures as Figure 4.1 averaged over the five classes and over the number of studies.

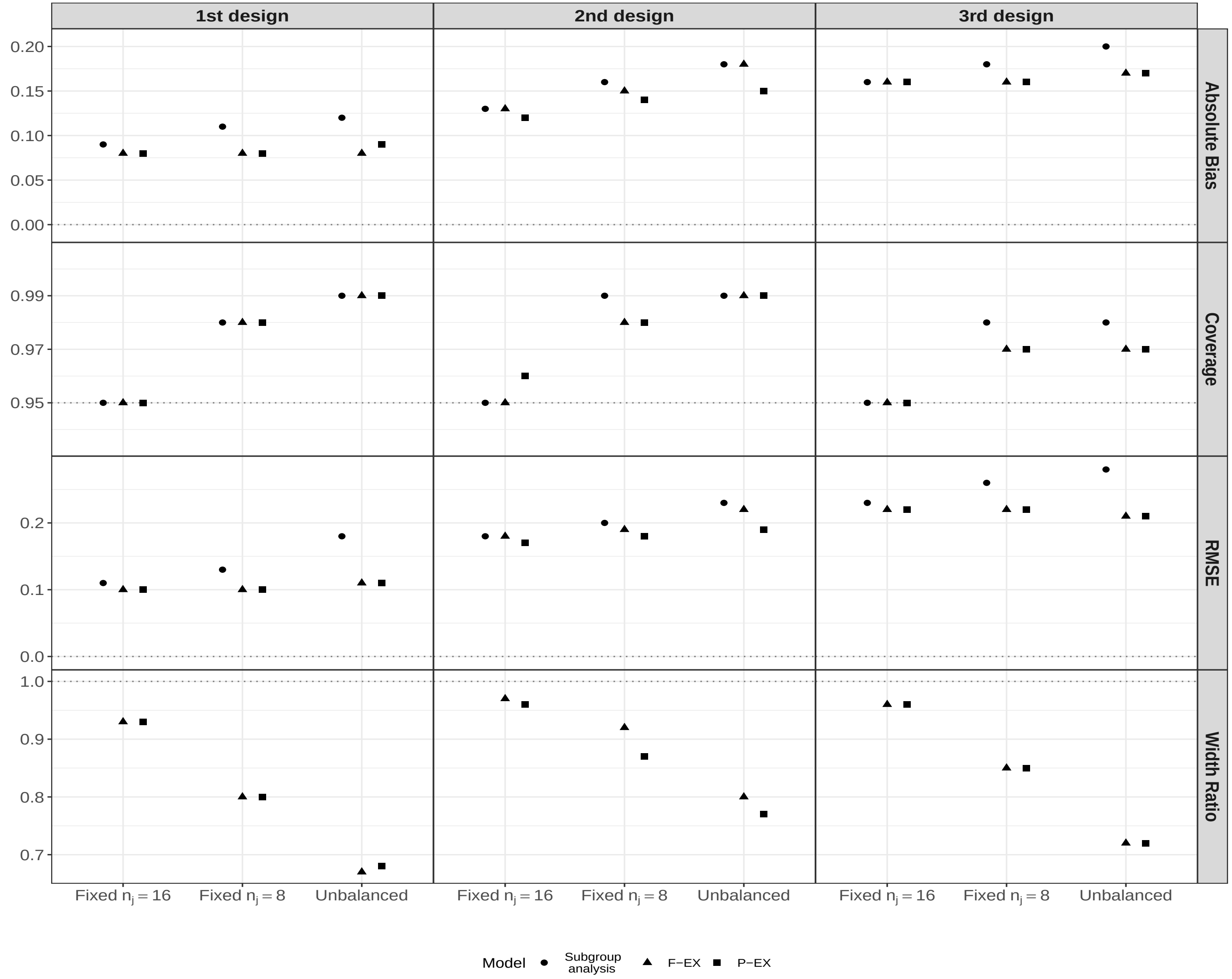
In the first design (scenarios 1, 2, 3), F-EX and P-EX models outperformed subgroup analysis in terms of the average absolute bias, the average RMSE and the uncertainty around the estimate $\hat{\delta}_{2ij}$. However, there was no winner between them as both methods had almost the same degree of borrowing of information resulting in 7%, 20%, and 33% narrower on average predictive intervals compared to subgroup analysis in these three scenarios respectively - the width ratios for both methods were 0.93 in the 1st scenario, 0.8 in the 2nd scenario, 0.67, in the 3rd one.

In the second design (scenarios 4, 5, 6), P-EX yielded predictions of the true treatment effect on the final outcome with the smallest average absolute bias, average RMSE and 95% CrIs of δ_{2ij} with the smallest average width ratio. Furthermore, P-EX method gave the most robust results for the 'extreme' treatment class, reducing by 44%, 69% and 20% the borrowing of information in this class across the scenarios (see the mixture weights of the 1st class in Table 4.2). In the 6th scenario F-EX performed poorer compared to P-EX model leading to higher average absolute bias and average RMSE. Subgroup analysis performed almost equally well as the P-EX model in the 4th scenario where the number of studies per class was relatively large.

The third design (scenarios 7, 8, 9) gave similar results as the first three in terms of the uncertainty (average width ratios), the average absolute bias and the average RMSE of $\hat{\delta}_{2ij}$. F-EX, P-EX models performed equally well, whilst subgroup analysis with the standard model was the worst approach resulting in inflated predictive intervals and larger RMSE in all cases.

In scenarios 1, 4 and 7, the models achieved 95% coverage due to the large amount of data, however, in the remaining scenarios where the number of studies was smaller the models resulted in higher than 95% coverage probabilities.

Figure 4.2: Absolute bias of $\hat{\delta}_{2ij}$ averaged over the 5 treatment classes and the simulated studies (first row), coverage averaged over the 5 classes of δ_{2ij} and the simulated studies (second row), RMSE of $\hat{\delta}_{2ij}$ averaged over the 5 classes (third row) and width ratios of δ_{2ij} averaged over the 5 classes and the number of simulated studies (forth row)



4.3.3.3 Probabilities of estimating a strong association pattern

Another aim of the simulation study was to assess the ability of the models to identify treatment classes with a strong association pattern. To evaluate this, we calculated probabilities of estimating a strong association pattern obtained from the 1000 replications in each scenario. These probabilities were based on the three surrogacy criteria proposed by Daniels and Hughes. Table 4.3 shows the average probabilities of estimating a strong association pattern over the five treatment classes across the first 6 data scenarios. Overall, F-EX and P-EX methods estimated the association patterns better compared to subgroup analysis across all scenarios. In the first design (scenarios 1, 2, 3) where the association was designed to be strong for all the classes, F-EX and P-EX models estimated a strong association pattern in more than 85% of the simulations. Subgroup analysis estimated the 81% of them in the 1st scenario but its performance reduced noticeably in the 2nd and 3rd scenario where the data were more sparse. In the second design (scenarios 4, 5, 6) with strong association patterns across all classes, P-EX and F-EX estimated more than 87% of the association patterns across these three scenarios. Subgroup analysis performed well only in the 4th scenario predicting the 89% of the association patterns but its performance gradually reduced as the number of studies was decreased in scenario 5 and 6. The probabilities of estimating a strong association per class in designs 1 and 2 are presented in the Appendix in section B.1 (last column of the tables).

Table 4.3: Probabilities of estimating a strong association pattern averaged over the five treatment classes, in the 1st and the 2nd design

Design	Scenario	No of studies across classes	Subgroup Analysis	F-EX model	P-EX model
1st	1st	Fixed ($n_j = 16$)	0.81	0.85	0.85
	2nd	Fixed ($n_j = 8$)	0.71	0.89	0.90
	3rd	Unbalanced	0.56	0.89	0.88
2nd	4th	Fixed ($n_j = 16$)	0.89	0.91	0.91
	5th	Fixed ($n_j = 8$)	0.88	0.92	0.92
	6th	Unbalanced	0.72	0.88	0.87

The last design focuses on the association patterns, consisting of three treatment classes with strong association patterns and two classes with weak surrogate relationship. In this design, it is important to present the probabilities of estimating a strong association pattern in each treatment class separately as the association patterns varied across classes.. Table 4.4 presents the probabilities in each treatment class separately across the three models. F-EX and P-EX methods were able to estimate a strong association pattern with higher probability compared to subgroup analysis in the classes where the association was designed to be strong. At the same time, the methods successfully identified classes with strong association patterns from a mixture of classes with weak and strong association patterns, even for the scenarios with relatively few studies per class where subgroup analysis failed almost completely to identify. Subgroup analysis performed similarly to the F-EX and P-EX methods in the 7th and 8th scenario but its performance was substantially decreased in the scenario with unbalanced treatment classes and especially in the first treatment class which consists of 4 studies.

Table 4.4: Probabilities of estimating a strong association pattern per class in the 3rd design

Scenario	No of studies across classes	Treatment classes	Subgroup Analysis	F-EX model	P-EX model
7th	Fixed ($n_j = 16$)	1^{st} class	0.82	0.84	0.84
		2^{nd} class*	0.00	0.00	0.00
		3^{rd} class	0.83	0.85	0.85
		4^{th} class*	0.00	0.00	0.00
		5^{th} class	0.80	0.80	0.80
8th	Fixed ($n_j = 8$)	1^{st} class	0.78	0.89	0.89
		2^{nd} class*	0.04	0.05	0.05
		3^{rd} class	0.80	0.90	0.90
		4^{th} class*	0.06	0.06	0.06
		5^{th} class	0.85	0.87	0.86
9th	Unbalanced	1^{st} class	0.06	0.82	0.80
	($n_1 = 4$,$n_2 = 8$,	2^{nd} class*	0.06	0.07	0.07
	$n_3 = 6$,$n_4 = 10$,	3^{rd} class	0.65	0.91	0.91
	$n_5 = 7$)	4^{th} class*	0.03	0.03	0.03
		5^{th} class	0.82	0.89	0.89

*Treatment classes with weak association pattern

4.3.4 Key findings

This section presents a short summary of the key finding of the simulation study:

- The aim of the simulation study was to illustrate and assess the performance of the methods under different scenarios. The models gave 95% coverage probabilities in the scenarios 1, 4, and 7 where the number of studies was sufficiently large (16 for each class). However, in the remaining scenarios the coverage probabilities were higher than 95%, which means that the methods derived more conservative CrIs of parameters than expected. This is largely due to the sparsity of the data in these scenarios but may also be partly due to different models being used to simulate and analyse the data as explained in section 4.3.1.
- In the first design (scenarios 1, 2, 3) where the assumption of exchangeability was reasonable, F-EX and P-EX models achieved similar performance and performed better compared to subgroup analysis giving on average narrower 95% CrIs of λ_{1j} and 95% predictive intervals of δ_{2ij}. This indicates that P-EX model successfully identified the correct level of borrowing of information inferring that the mixture weights should be very close to 1.
- P-EX model was the best choice in all the scenarios of the second design (scenarios 4, 5, 6) where there was a treatment class with distinctly different slope. It reduced the degree of borrowing of information for the 'extreme' treatment class, resulting in the most accurate posterior means of the slopes in the scenarios 4, 5 and 6 and the most precise 95% CrIs of the slopes in the scenarios 4 and 5. P-EX model was the best choice in terms of predictions of the true effect on the final endpoint, reducing the width of predictive intervals by 4%, 13% and 23% compared to subgroup analysis in each scenario respectively.
- Another aim of the simulation study, was to investigate whether the proposed methods identified treatment classes with the strong association patterns better compared to subgroup analysis across all data scenarios. In particular, in scenarios 3, 6 and 9, where the data were sparse, the proposed hierarchical methods were able to estimate surrogacy significantly better compared to the subgroup analysis. This illustrates well the benefits of using hierarchical

methods when data are limited. Furthermore, as illustrated by scenarios 7, 8 and 9, F-EX and P-EX models could easily distinguish between the different association patterns, as they identified treatment classes with strong association patterns and at the same time did not overestimate the strength of the association in the classes where the association was designed to be weak.

4.4 Data example

4.4.1 Advanced colorectal cancer, treatment classes and candidate endpoints

Advanced colorectal cancer is among the most common types of cancer worldwide [120]. It has the third highest incident rate in the UK after breast and lung cancer. Although the incident rates remained relatively stable for over a decade in the UK, recent reports have shown increased rates of aCRC in economically transitioning countries around the world [121]. Among all patients diagnosed with colorectal cancer approximately the 20% will develop a metastasis and treatment is palliative rather than curative as the 5-year overall survival approximately is 10% [122].

Traditional therapies form a treatment class of chemotherapy which consist of cytotoxic agents. The most efficacious cytotoxic agents in aCRC are 5-fluorouracil (5-FU) folinic acid, irinotecan, oxaliplatin and capecitabine. These drugs combined consist of the chemotherapy treatment class. Some of them are: FOLFOX (folinic acid, fluorouracil, oxaliplatin), FOLFIRI (folinic acid, fluorouracil, irinotecan) or XELOX (capecetabine and oxaliplatin). Two newer classes of targeted treatments have shown to improve treatment outcomes such as OS or PFS for aCRC when combined with cytotoxic agents [123–127]. The first one is the class of anti-angiogenic treatments. These treatments focus on stopping angiogenesis, which is the process of making new blood vessels. Tumours use blood vessels to grow, therefore, by blocking the flow of nutrients and oxygen to tumours they can halt the tumour growth and stop the spread. Some of the most popular angiogenesis inhibitors in aCrC are Bevacizumab, Regorafenib and Ziv-aflibercept. The newer class of targeted treatments are the anti-epidermal growth factor receptor anti-EGFR monoclonal antibodies. Researches have found that therapies which block EGFR can potentially be effective for stopping or slowing tumour growth for aCRC. Anti-EGFR agents such as cetuximab and panitumumab are widely used in combination with cytotoxic for aCRC. However, the efficacy of EGFR inhibitors was found to vary as several studies have shown limited benefits to patients who have tumours with K-RAS mutations [104].

The primary and long term final outcome in this disease area is OS whereas, the most investigated surrogate endpoints are PFS, TTP and TR. These endpoints can be defined as continuous variables, however, TR can be considered also as categorical with four ordered categories (complete response, partial response, stable disease, progressive disease). Several authors have investigated the validity of the above endpoints as surrogate endpoints over the last decade [11, 12, 24, 99, 128]. In 2007, Buyse et al. [12] showed that PFS is an acceptable surrogate endpoint for OS using trials which compare only traditional types of chemotherapy with modern types of chemotherapy. After the introduction of targeted treatments Giessen et al. [99] investigate the study-level surrogacy patterns across treatment classes by performing subgroup analysis. They inferred that further research was needed to establish a strong surrogate relationship between treatment effects on PFS and treatment effects on OS. Ciani et al. [11] investigated surrogate relationships between treatment effects on potential surrogate endpoints (TR and PFS) and on the final clinical outcome (OS) using a more recent and diverse in terms of treatment classes data-set. They carried out a systematic review which consists of 101 RCTs including the following 5 treatment classes: the class of chemotherapy, the anti-EGFR class, angiogenesis inhibitors, other molecular-targeted agents and intrahepatic arterial chemotherapies. They found that the surrogate relationships between the treatment effects on PFS/TR and the OS were suboptimal. Specifically, they stated that PFS was an acceptable surrogate endpoint for OS however, they found weaker association pattern between the treatment effects on PFS and OS compared to the findings by Buyse et al. [12]. They concluded that a very strong association pattern observed in a specific treatment class (class of chemotherapy) in aCRC, may not apply directly across other classes of treatments. More details about the studies and how the systematic review was designed can be found in Ciani et al. [11]. We refer these data as 'Ciani data' in the remainder of this chapter.

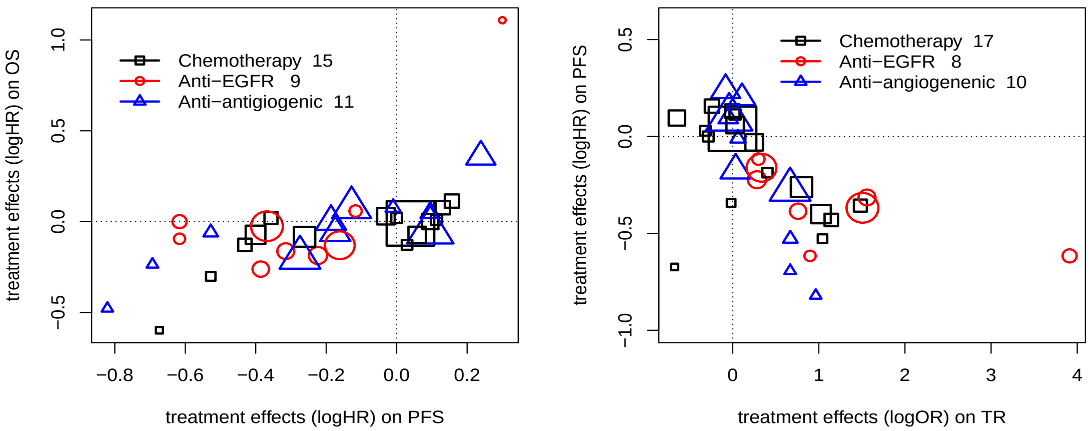
4.4.2 Data-extraction

To illustrate the proposed methodology, we focused on a subset of the 'Ciani data', examining the association patterns between treatment effects on TR and PFS and treatment effects on PFS and OS including data from three treatment classes. We extracted data from 35 studies reporting treatment effect on PFS and OS. 15 of them belonged to the chemotherapy treatment class, 9 of them investigated anti-EGFR therapies and 11 anti-angiogenic treatments. To investigate the association patterns between treatment effects on TR and PFS, we extracted data from 35 studies reporting treatment effects on these endpoints; 17 of them investigated chemotherapies, 8 and 10 studies anti-EGFR and anti-angiogenic treatments respectively. TR can be evaluated as a surrogate endpoint to treatment effect on PFS, as treatment effects on TR is typically measured earlier compared to treatment effects on PFS. In this data-set, IPD were available from four RCTs [124, 129–131]. Two of the studies [130, 131] belonged to the class of chemotherapy and the other two [124, 129] to the anti-angiogenic treatment class.

Treatment effects on PFS and OS were obtained on the log hazard ratio scale $logHR(OS)$, $logHR(PFS)$, whereas the treatment effects on TR were calculated from the reported number of events an measured on log odds ratio scale $logOR(TR)$. We also retrieved the corresponding standard errors in each study and on each outcome.

Figure 4.3 provides a graphical representation of the data-set we used, where the size of each data point corresponds to the size of each study. It illustrates the association patterns between the treatment effects on each pair of outcomes across classes.

Figure 4.3: Scatterplots of treatment effects on PFS-OS and TR-PFS



4.4.3 Data synthesis

The method proposed by Daniels and Hughes, F-EX and P-EX models account for within-study associations across studies via the within-study correlations ρ_{wi}. Within-study correlations can be estimated using a bootstrap method (the implementation of the bootstrap method can be found in the Appendix B.3) from IPD.

Table 4.5 presents the within-study correlations between treatment effects on each pair of outcomes for each of the studies where IPD were available.

Table 4.5: Within-study correlations across the 4 RCTs were IPD were available

	Endpoints	
Studies	PFS-OS	TR-PFS
AVF2107g Study[124]	0.52	-0.43
ML18147 Study [129]	0.54	-0.31
NO16966 study [130]	0.55	-0.39
NO16967 Study [131]	0.55	-0.38

The within-study correlations estimated for each pair of outcomes were very similar across these 4 studies. Hence, due to the lack of IPD for the remaining studies, we decided to use a fixed within-study correlation for each pair of outcomes across all the studies of the data-set. By taking the mean of the 4 correlations for each pair of outcomes, we created the following two within-study correlations: $\rho_{w(PFS-OS)} = 0.54$, $\rho_{w(TR-PFS)} = -0.38$.

4.4.4 Data analysis

To explore potential differences in association patterns across treatment classes, we performed subgroup analysis using standard model and applied the proposed hierarchical models to the extracted data. We estimated posterior distributions for the parameters describing the surrogate relationships in each treatment class, monitoring the posteriors means of the intercepts $\hat{\lambda}_{0j}$, the slopes $\hat{\lambda}_{1j}$ and posterior medians of conditional variances $\hat{\psi}_j^2$ with their corresponding 95% CrIs and BFs of $\hat{\psi}_j^2$ using the Savage Dickey density ratio [115]. By using the evaluation framework proposed by Daniels and Hughes (discussed in section 4.2.1.1), we were able to infer whether or not a candidate endpoint is a valid surrogate in each treatment class. A cross-validation procedure was also carried out to investigate the performance of the prediction of the true treatment effects on the final clinical outcome in terms of precision across methods. Throughout the cross-validation procedure, we monitored and reported the following measures: mean absolute error of the predictions across studies, ratio of the width of the 95% predictive intervals of P-EX or F-EX to the width of the 95% predicted interval of subgroup analysis, averaged over the studies.

4.4.4.1 Association patterns across models and treatment classes

Subgroup analysis with the standard model

Table 4.6 presents the estimates of the parameters describing the surrogate relationship of subgroup analysis with the standard method. Strong association patterns were found between the treatment effects on PFS and the effects on OS in the class of chemotherapy and the anti-angiogenic treatment class, as the three criteria for surrogacy were satisfied (the 95% CrIs of λ_{01} and λ_{03} included zero, the 95% CrIs of λ_{11} and λ_{13} did not contain zero and there was substantial evidence using BFs in favour of the hypotheses $H_1 : \psi_1^2 = 0$, and $H_1 : \psi_3^2 = 0$, the BFs of ψ_1^2 and ψ_3^2 were larger than 10). On the other hand, we can infer that the association pattern between the treatment effects on PFS and the effects on OS in the anti-EGFR treatment class was weak, as the 95% CrI of the posterior distribution of the slope included zero. Therefore, PFS was a valid surrogate endpoint of OS only in the anti-angiogenic treatment class and the class of chemotherapy.

Investigating the surrogacy on TR-PFS pair of outcomes we found a similar pattern, thus we can infer that there was an acceptable surrogate relationship between treatment effects on TR and PFS in the chemotherapy and the anti-angiogenic classes. The relationship was negative overall, as the slopes were negative across classes. Additionally, the surrogacy criteria indicated poor surrogacy between the treatment effects on TR and the treatment effects on PFS for anti-EGFR class, since the 95% CrI of the slope λ_{12} included zero.

Table 4.6: Estimates of the parameters defining the surrogacy criteria of subgroup analysis with the standard model

Treatment Class	parameter	PFS-OS	TR-PFS
chemotherapy		N=15	N=17
	λ_{01}	-0.00 (-0.06, 0.05)	-0.05 (-0.16, 0.03)
	λ_{11}	0.31 (0.08, 0.55)	-0.26 (-0.40,-0.10)
	ψ_1^2	0.00 (0.00, 0.01)	0.02 (0.00, 0.07)
	BF of ψ_1^2	310.43	7.89
anti-EGFR		N=9	N=8
	λ_{02}	-0.05 (-0.29, 0.29)	-0.20 (-0.42, 0.03)
	λ_{12}	0.12 (-0.55, 1.02)	-0.14 (-0.37, 0.02)
	ψ_2^2	0.01 (0.00, 0.10)	0.01(0.00, 0.13)
	BF of ψ_2^2	25.33	14.42
anti-angiogenic		N=11	N=10
	λ_{03}	0.05 (-0.04, 0.15)	0.07 (-0.08, 0.23)
	λ_{13}	0.48 (0.18, 0.80)	-0.786 (-1.20,-0.46)
	ψ_3^2	0.01 (0.00, 0.04)	0.01 (0.00,0.09)
	BF of ψ_3^2	20.98	19.83

F-EX model

Table 4.7 displays results of the parameters describing the surrogate relationships of F-EX across the three treatment classes. PFS was deemed an valid surrogate endpoint of OS in the anti-angiogenic and chemotherapy treatment classes as the surrogacy criteria were satisfied in both treatment classes (i.e. the 95% CrIs of the slopes did not contain zero, the 95% CrIs of the intercepts λ_{01} and λ_{03} and the BFs of the conditional variances of ψ_1^2 and ψ_3^2 were larger than 3.3). On the other hand, the association between the treatment effects on PFS and the treatment effects on OS was weak in the anti-EGFR treatment class, failing to meet one of the criteria, as the 95% CrI of the slope λ_{12} included zero.

Fitting F-EX model to the data extracted on TR-PFS pair of outcomes resulted in different inferences about the parameters describing the surrogate relationships compared to the analysis based on subgroup analysis. Here, the three surrogacy criteria were satisfied across all the three treatment classes. This is due to the assumption of exchangeability of the parameters λ_{0j} and λ_{1j}. As a result, TR was an acceptable surrogate endpoint of PFS across the three treatment classes in this data-set.

Table 4.7: Estimates of the parameters defining the surrogacy criteria of F-EX model

Treatment Class	parameter	PFS-OS	TR-PFS
		N=15	N=17
	λ_{01}	-0.00 (-0.05, 0.05)	-0.05 (-0.15, 0.03)
chemotherapy	λ_{11}	0.33 (0.13, 0.53)	-0.27 (-0.41,-0.11)
	ψ_1^2	0.00 (0.00, 0.01)	0.02 (0.00, 0.07)
	BF of ψ_1^2	300.80	9.09
		N=9	N=8
	λ_{02}	0.00 (-0.15, 0.15)	-0.14 (-0.34, 0.06)
anti-EGFR	λ_{12}	0.28 (-0.16, 0.63)	-0.19 (-0.42,-0.03)
	ψ_2^2	0.01 (0.00, 0.08)	0.01(0.00, 0.13)
	BF of ψ_2^2	18.04	14.02
		N=11	N=10
	λ_{03}	0.03 (-0.04, 0.11)	0.03 (-0.13, 0.18)
anti-angiogenic	λ_{13}	0.41 (0.16, 0.68)	-0.67 (-1.06,-0.27)
	ψ_3^2	0.01 (0.00, 0.04)	0.02 (0.00,0.12)
	BF of ψ_3^2	26.97	13.23

P-EX model

P-EX model allows the parameters of slope of each treatment class to be either exchangeable or non-exchangeable with parameters of slopes from other classes throughout the estimation process. For both pairs of outcomes, fixed values for the hyper-parameters $\pi_j = (0.5, 0.5, 0.5)$ were chosen assuming that exchangeability and non-exchangeability were *apriori* equally likely.

As in the case of F-EX model, the parameters describing the surrogate relationships were estimated for each class. Additionally, this model regulates the degree of borrowing of information for the parameter of the slope by using an exchangeable and a non-exchangeable component. To estimate the degree of borrowing of information across classes, we also monitored the mixture weights by calculating the posterior means of p_j. Table 4.8 presents the estimates of the parameters defining the surrogacy criteria.

For the PFS-OS pair, P-EX estimated the parameter of the slope using the exchangeable component in the 97% of the MCMC iterations in the class of chemotherapy in the 96% of the MCMC iterations in the anti-EGFR class and in the 97% of the iterations in the anti-angiogenic treatment class. This indicates that P-EX reduced borrowing of information approximately 3% in the anti-angiogenic treatment class and the class of chemotherapy and 4% in the anti-EGFR class compared to F-EX model. Focusing on the estimates of the parameters describing the surrogate relationships of P-EX, we drew the same inferences as from F-EX model. PFS was deemed as a valid surrogate endpoint of OS in the anti-angiogenic and the chemotherapy classes. On the other hand there were not enough evidence to validate PFS as a surrogate endpoint of OS in the anti-EGFR treatment class, as the 95% CrI of the slope λ_{12} included zero.

For TR-PFS pair of outcomes, the degree of borrowing of information was smaller compared to PFS-OS pair of outcomes. P-EX estimated the slopes, reducing the borrowing of information by 7% in the anti-angiogenic class, by 5% in the anti-EGFR class and 6% in the class of chemotherapy compared to F-EX model. The three surrogacy criteria were fulfilled in each treatment classes despite the decrease in levels of borrowing of information, indicating that TR was an acceptable surrogate of PFS across the treatment classes of the Ciani data.4.8

Table 4.8: Estimates of the parameters defining the surrogacy criteria of P-EX model

Treatment Class	parameter	PFS-OS	TR-PFS
		N=15	N=17
	p_1	0.97	0.94
	λ_{01}	0.01 (-0.04, 0.05)	-0.05 (-0.16, 0.03)
chemotherapy	λ_{11}	0.33 (0.13, 0.54)	-0.27 (-0.40,-0.11)
	ψ_1^2	0.00 (0.00, 0.01)	0.02 (0.00, 0.07)
	BF of ψ_1^2	309.29	8.31
		N=9	N=8
	p_2	0.96	0.95
	λ_{02}	0.01 (-0.16, 0.14)	-0.14 (-0.34, 0.06)
anti-EGFR	λ_{12}	0.27 (-0.17, 0.63)	-0.18 (-0.42,-0.02)
	ψ_2^2	0.01 (0.00, 0.08)	0.01(0.00, 0.13)
	BF of ψ_2^2	16.94	13.52
		N=11	N=10
	p_3	0.97	0.93
	λ_{03}	0.03 (-0.04, 0.11)	0.03 (-0.13, 0.18)
anti-angiogenic	λ_{13}	0.41 (0.16, 0.69)	-0.69 (-1.08,-0.28)
	ψ_3^2	0.01 (0.00, 0.04)	0.02 (0.00,0.11)
	BF of ψ_3^2	27.90	14.18

As discussed in Section 4.2.3, we fitted P-EX to the data using fixed values for the hyper-parameters $\pi_j = (0.5, 0.5, 0.5)$ assuming that exchangeability and non-exchangeability were *apriori* equally likely. To evaluate how sensitive were the results of the parameters describing the surrogate relationship to the values of π_j, we performed a sensitivity analysis placing a non informative prior distribution to each $\pi_j \sim Beta(1, 1)$ instead of assigning a fixed value to each parameter. Table 4.9 compares the results of P-EX having the hyperparameters π_j fixed with the results of P-EX when the hyperparameters assumed random. It can be seen that both versions of P-EX model gave very similar results regardless off whether we used fixed or random hyperparameters π_j.

Table 4.9: Estimates of the parameters defining the surrogacy criteria of P-EX model

		P-EX model with fixed π_j		P-EX model with random π_j	
Treatment class	parameter	PFS-OS	TR-PFS	PFS-OS	TR-PFS
		N=15	N=17	N=15	N=17
	p_1	0.97	0.94	0.98	0.94
	λ_{01}	0.01 (-0.04, 0.05)	-0.05 (-0.16, 0.03)	0.00 (-0.05, 0.05)	-0.05 (-0.16, 0.03)
chemotherapy	λ_{11}	0.33 (0.13, 0.54)	-0.27 (-0.40,-0.11)	0.33 (0.13, 0.53)	-0.27 (-0.40,-0.11)
	ψ_1^2	0.00 (0.00, 0.01)	0.02 (0.00, 0.07)	0.00 (0.00, 0.01)	0.02 (0.00, 0.07)
	BF of ψ_1^2	309.29	8.31	306.53	8.11
		N=9	N=8	N=9	N=8
	p_2	0.96	0.95	0.96	0.95
	λ_{02}	0.01 (-0.16, 0.14)	-0.14 (-0.34, 0.06)	0.01 (-0.16, 0.14)	-0.15 (-0.35, 0.05)
anti-EGFR	λ_{12}	0.27 (-0.17, 0.63)	-0.18 (-0.42,-0.02)	0.27 (-0.17, 0.63)	-0.18 (-0.42,-0.02)
	ψ_2^2	0.01 (0.00, 0.08)	0.01(0.00, 0.13)	0.01 (0.00, 0.08)	0.01(0.00, 0.13)
	BF of ψ_2^2	16.94	13.52	17.70	13.92
		N=11	N=10	N=11	N=10
	p_3	0.97	0.93	0.97	0.92
	λ_{03}	0.03 (-0.04, 0.11)	0.03 (-0.13, 0.18)	0.03 (-0.04, 0.12)	0.03 (-0.12, 0.18)
anti-angiogenic	λ_{13}	0.41 (0.16, 0.69)	-0.69 (-1.08,-0.28)	0.41 (0.16, 0.70)	-0.70 (-1.10,-0.29)
	ψ_3^2	0.01 (0.00, 0.04)	0.02 (0.00,0.11)	0.01 (0.00, 0.04)	0.02 (0.00,0.11)
	BF of ψ_3^2	27.90	14.18	27.67	14.93

4.4.4.2 Results of the cross-validation procedure across models

After estimating the parameters describing the surrogate relationship and applying the validation framework proposed by Daniels and Hughes, we carried out cross-validation procedure to predict the treatment effects δ_{2i} on the final outcome (discussed in Section 4.2.4). Table 4.10 displays the results of cross-validation procedure across models. Overall, all the methods performed almost equally well in terms of absolute error.

Subgroup analysis with the standard model gave predictive intervals of the effects on the final outcome containing the corresponding observed estimates y_{2i} in the 97% of the studies for both pairs of outcomes confirming good fit of the model (first row last column). It outperformed F-EX and P-EX models in the anti-angiogenic class resulting in predictions of the true treatment effects with smaller on average absolute error on TR-PFS pair of outcomes. It also performed marginally worse compared to the proposed methods in terms of absolute error in the treatment class of chemotherapy, where the number of studies was large on both pairs of outcomes. In contrast to this, it performed poorly in terms of accuracy of predictions in the anti-EGFR class (large absolute error), as the number of studies small for both pair of outcomes.

The results from the cross-validation procedure of F-EX model showed that the method fitted the data well. All of the predicted intervals of δ_{2ij} contained the observed values of the treatment effects on the final outcome on PFS-OS pair and all but one on TR-PFS pair. The cross-validation procedure of F-EX performed slightly better compared to subgroup analysis in chemotherapy treatment class on both pairs of outcomes resulting in smaller on average absolute error. F-EX also was the best method in terms of its accuracy in the anti-EGFR treatment class on PFS-OS pair. In contrast to this, higher average absolute errors were observed in the anti-angiogenic class on both pair of outcomes indicating that the assumption of exchangeability of the parameters describing the surrogate relationships was fairly strong and it was likely to affect the predictions in this particular class. The overall results of the width ratios (second row, last two columns) imply that F-EX method gave intervals of the true effect on the final endpoint with smaller degree of uncertainty compared to subgroup analysis especially on PFS-OS pair of outcomes. There was a small decrease in the uncertainty of the predictions of δ_{2ij} on PFS-OS pair for the chemotherapy treatment class, as the cross-validation procedure of F-EX model yielded 1% narrower intervals compared to subgroup analysis. Furthermore, significantly reduced uncertainty was observed in the other two treatment classes on PFS-OS pair of outcomes, 14% in the anti-EGFR treatment class and 7% in the anti-angiogenic, where the number of studies was smaller. On the contrary, very limited decrease in the degree of uncertainty was observed for the TR-PFS pair of outcomes across all classes. Overall on TR-PFS pair, the predictive intervals were only 1% narrower compared to subgroup analysis. The benefit was small (3.2% reduction of the width of the predictive interval) even for the anti-EGFR treatment class where there were only 8 studies for this pair.

Focusing on the results from the cross-validation procedure of P-EX model, all the intervals of the predicted treatment effects on the final outcome contained the observed treatment effects on PFS-OS pair and all but one on the TR-PFS pair. The absolute error was smaller in chemotherapy treatment class on the PFS-OS pair and significantly higher in the other two classes. P-EX gave almost equally accurate estimates in the anti-EGFR and the chemotherapy treatment classes on TR-PFS pair compared to F-EX. However, the absolute error was higher in the anti-angiogenic treatment class where the association was much stronger compared to the other two classes indicating potential excessive borrowing of information from the other classes. This is likely due to that the assumption of partial exchangeability was only applied to the slopes (the intercepts were assumed to be exchangeable across classes). The method predicted the effects on the final outcome with reduced uncertainty giving more precise predictions of the true effect on the final outcome compared to subgroup analysis in the anti-EGFR class on PFS-OS pair reducing the uncertainty by 13%. On the other hand, the predicted effects $\hat{\delta}_{2ij}$ had almost the same degree of uncertainty as those of subgroup analysis on TR-PFS pair. The intervals were only 1% narrower on average across all classes compared to the subgroup analysis.

Table 4.10: Predictions of δ_{2ij} across treatments and models

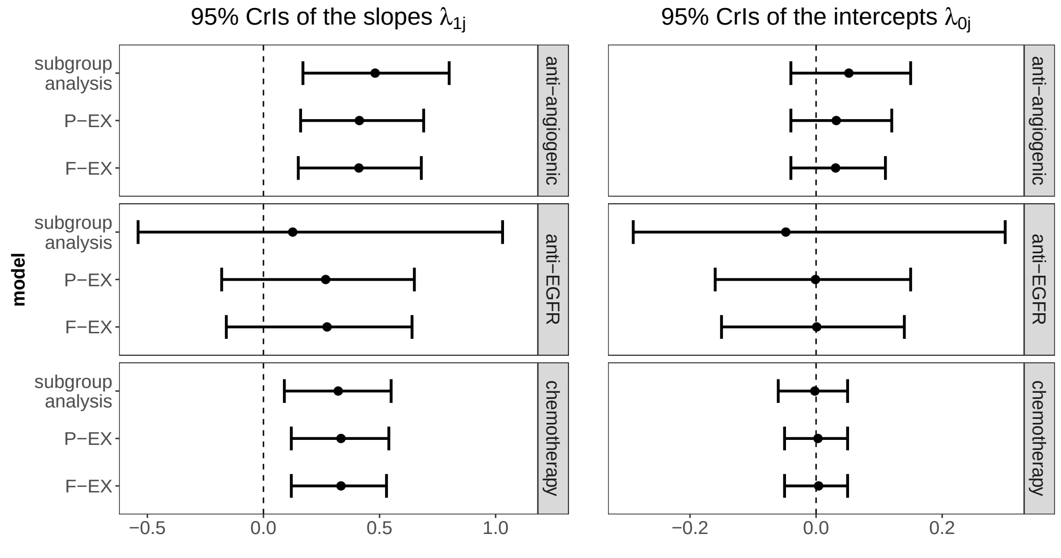
		chemotherapy		***anti-EGFR***		***anti-angiogenic***		***Overall***	
Models	Measures	PFS-OS	TR-PFS	PFS-OS	TR-PFS	PFS-OS	TR-PFS	PFS-OS	TR-PFS
Standard Model	Performance of 95% predictive intervals	1.00	0.94	0.89	1.00	1.00	1.00	0.97	0.97
	Mean absolute error	0.05	0.11	0.14	0.13	0.10	0.15	0.09	0.12
F-EX	Performance of 95% predictive intervals	1.00	0.94	1.00	1.00	1.00	1.00	1.00	0.97
	Mean absolute error	0.04	0.10	0.10	0.11	0.12	0.21	0.09	0.13
	Mean width ratio	0.99	0.99	0.86	0.97	0.93	1.00	0.95	0.99
P-EX	Performance of 95% predictive intervals	1.00	0.94	1.00	1.00	1.00	1.00	1.00	0.97
	Mean absolute error	0.04	0.10	0.13	0.11	0.11	0.21	0.09	0.13
	Mean width ratio	0.99	0.99	0.87	0.98	0.93	1.00	0.96	0.99

4.4.5 Discussion of the results of the data example across models

In this section we compare the estimates (posterior means and 95% CrIs) of λ_{1j} and λ_{0j} across the three methods, as each model makes different assumptions about these parameters. Subgroup analysis with the standard model assumes that both parameters are non-exchangeable across classes, F-EX assumes full exchangeability of both parameters across treatment classes and P-EX assumes that the slopes are partially exchangeable and the intercepts fully exchangeable across classes.

Figure 4.4 presents 95% CrIs of the slopes λ_{1j} and intercepts λ_{0j} on PFS-OS pair of outcomes across the treatment classes and methods of estimation. Comparing the aforementioned methods in regards to the surrogacy criteria on the PFS-OS pair, we can conclude that F-EX model estimated the parameters of the surrogate relationships with reduced uncertainty compared to the subgroup analysis and P-EX model taking advantage of borrowing of information across classes. P-EX relaxes the assumption of exchangeability reducing the effect of borrowing of information on average by 3%. It gave narrower CrIs of the parameters of interest compared to subgroup analysis but marginally wider than those obtained form F-EX model. F-EX and P-EX models could distinguish between the different association patterns avoiding to give over-shrunk estimates of the slopes and the intercepts, although they

Figure 4.4: 95% Credible intervals of λ_{1j} and λ_{0j} for the PFS-OS pair of outcomes



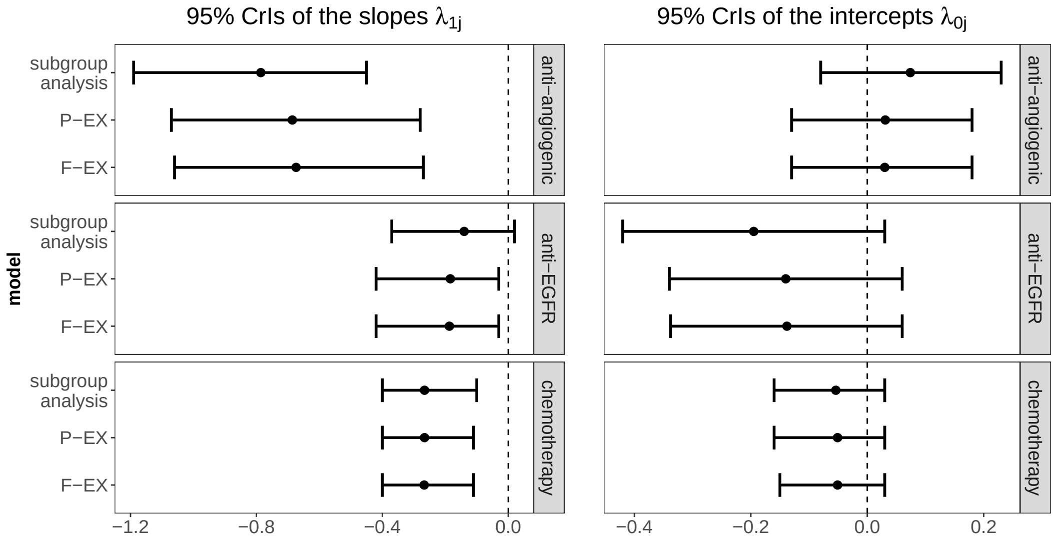
allowed different degrees of borrowing of information for the slopes. In particular, this pair of outcomes (PFS-OS) illustrates well the impact of number of studies per class on the degree of borrowing of information. In general, borrowing of information is determined by the number of studies within treatment classes, between treatment classes heterogeneity, as well as the number of treatment classes. In this case, the fewer studies we have within a treatment class, the bigger is the impact of borrowing of information resulting in higher reduction in uncertainty of the estimates of surrogate relationships. This effect was particularly strong for the anti-EGFR treatment class.

On the other hand, TR-PFS pair of outcomes was a good example to illustrate the performance of the hierarchical methods when between treatment class heterogeneity is relatively large. In this case, subgroup analysis with the standard model performed equally well as the proposed methods in terms of uncertainty of the CrIs of the parameters describing the surrogate relationships. For instance by fitting F-EX and P-EX models, we did not observe any decrease in uncertainty around λ_{1j} and λ_{0j} across classes (Figure 4.5). This is because the between treatment classes heterogeneity was relatively large for TR-PFS pair and hence there was not much shrinkage. By performing subgroup analysis, the surrogacy criteria failed in the anti-EGFR class (zero was included in the 95% CrI of the slope). However, the 95% CrI in the anti-EGFR class just contained zero and substantially overlapped with the 95% CrI of the slope in the chemotherapy treatment class. By fitting P-EX and F-EX models, we were able to draw different inferences for the association patterns in the anti-EGFR class, as these methods allow for borrowing of information for the parameters describing the surrogate relationships from the other classes. As illustrated in Figure 4.5, both hierarchical models moved the 95% CrI of the slope in the direction of the CrIs of the other two classes resulting in the surrogacy criteria being satisfied across all treatment classes.

When carrying out cross-validation procedure, we wish to ensure that not only predictive intervals contain the observed values but also that they are sufficiently narrow. In general, adding a hierarchical structure to slopes and intercepts reduces the uncertainty and leads to more precise predictions compared to those obtained from subgroup analysis.

Starting from the findings on PFS-OS pair of outcomes, the accuracy of the

Figure 4.5: 95% Credible intervals of λ_{1j} and λ_{0j} for the TR-PFS pair of outcomes



predictions was very similar across all methods (similar absolute errors) but the uncertainty varied depending of the level of borrowing of information. F-EX model gave on average the most precise predictions of the true treatment effect, having the narrowest 95% predictive intervals (smallest width ratio seen in Table 4.10) reducing the overall uncertainly by 5%. The benefit was smaller in the chemotherapy class where the number of studies was much larger compared to the anti-EGFR treatment class where we had only 8 studies available. Overall, P-EX performed better compared to subgroup analysis with the standard model and equally well as F-EX model regarding the uncertainty of the predictions. This indicates that the assumption of exchangeability seems to be plausible for this pair of outcomes and P-EX model was able to identify this.

Moving to the predictions obtained on TR-PFS pair of outcomes, subgroup analysis with the standard model was a robust approach in terms of the accuracy of its predictions. Although the overall absolute error was very similar across models, F-EX and P-EX yielded higher absolute error compared to subgroup analysis in the anti-angiogenic class. This implies that the posterior means of the true effects were to some extent 'overshrunk' due to excessive borrowing of information from the other classes. P-EX model was implemented allowing for partial exchangeability of the slopes only, this decision is likely to affect the performance of the model in terms of its predictions on TR-PFS pair of outcomes. However, the model can be extended allowing for partial exchangeability also of the intercepts or the conditional variances and different combinations of these assumptions can be explored and models compared using DIC. Similarly, there was no significant decrease in the degree of uncertainty of the estimates $\hat{\delta}_{2ij}$ of F-EX and P-EX models. The results indicated that the proposed hierarchical models performed slightly better compared to subgroup analysis in terms of uncertainty only in the class of chemotherapy and the anti-EGFR treatment class giving 1% and 3% narrower predictive intervals respectively. This kind of behaviour might be caused by the relatively large between treatment class heterogeneity (due to different definitions of TR across RCTs) [132] and the assumption of full exchangeability of the intercepts.

4.5 Discussion

We developed two hierarchical models allowing to account for distinct treatment classes when examining the association patterns within each treatment class. The proposed models may be particularly useful in surrogate endpoint evaluation in complex diseases where different treatment classes of different mechanism of action and potential different association patterns within those classes exist. These models investigate potential differences in trial-level surrogacy across treatment classes in a particular disease area and can help to effectively identify treatment classes with strong association patterns, even when data are relatively sparse. F-EX model is somewhat restrictive, assuming full exchangeability of the parameters describing the surrogate relationships across treatment classes. However, in many situations the assumption of exchangeability may be too strong given the heterogeneity between treatment classes. In such circumstances, a more flexible model such as P-EX may be a better choice. P-EX model can infer an appropriate level of exchangeability from the data. It regulates the degree of borrowing of information by using an exchangeable and a non-exchangeable component throughout the estimation process, thus relaxing the assumption of exchangeability when it is not fully reasonable. It evaluates whether the association pattern between treatment effects on the surrogate and the final endpoint in a specific treatment class differs from the other patterns in other classes.

F-EX model is appropriate only when the degree of similarity of surrogate relationships is relatively high. It can offer substantial gains in precision, reduced RMSE of the posterior means of the parameters describing surrogate relationships and it can improve the predictions of the true effects on the final endpoint. For example, F-EX model gave posterior means of the slopes and predicted effects with reduced uncertainty (smaller credible intervals) compared to subgroup analysis in the first design of the simulation study and for the illustrative example on PFS-OS pair, where the parameters describing the surrogate relationship were similar and the assumption of full exchangeability was reasonable. These findings are consistent with the results from other hierarchical Bayesian methods which assume full exchangeability and were developed in other research areas [116, 117]. In addition to this, P-EX model achieved the same degree of borrowing of information in such data scenarios making fewer assumptions compared to F-EX model. Furthermore, when between treatment class heterogeneity is relatively large or there is a treatment class with distinctly different pattern, P-EX model has the advantage of avoiding the excessive borrowing of information, as illustrated in the second design of the simulation study. All the above illustrate the benefits of partial exchangeability, as described by Neuenschwander et al. [102] in their work. Subgroup analysis using the standard model is a simple method which performs well when there are sufficient data available for each treatment class, but it produces estimates with higher bias and uncertainty when data within a treatment class are limited.

Although the proposed methods provide additional robustness to the CrIs and the posterior means of the parameters describing the surrogate relationships compared to subgroup analysis, potential limitations should always be kept in mind. First, in real data scenarios it can be challenging to find data-sets with sufficient number of treatment classes. The small number of treatment classes can affect the performance of hierarchical methods substantially [133] reducing the impact of borrowing of information. For instance, fitting P-EX model to the illustrative example (in aCRC with three treatment classes) led to a situation where in some of the MCMC iterations only one class was deemed exchangeable by the model which is not possible since there were no other classes to exchange information with. However, in our example it did not affect the performance of the model as it occurred only in the 0.5% of the MCMC iterations. On the other hand, there is no upper limit to the number of classes we can have. In general, the larger number of classes the easier it is for the models to borrow information across classes.

Another limitation of the illustrative example is that treatment switching was applied in a subset of trials in this data-set. Patients were allowed to switch from the treatment that was initially assigned to them to the other treatment arm in the trial. Most commonly patients switched after progression from control to experimental arm in particular, when there was sufficient evidence during the trial that the experimental treatment was better than control [134]. Treatment switching has diminishing effect on the difference in treatment effects on OS when applying intention-to-treat analysis, and the treatment effect is often obtained with larger uncertainty. This makes the estimation of surrogacy between treatment effects on the surrogate and treatment effects on the final outcome very challenging. Many adjustment methods have been proposed, however, their validity is often questionable[134]. Additionally, the evaluation of PFS as a surrogate endpoint is distinctive compared to other surrogate endpoints as PFS can be considered as nested outcome within OS outcome. These factors may explain the different findings for the two pairs of outcomes (PFS-OS and TR-OS).

Furthermore, as it was mentioned in section 4.4, each treatment class consist of studies with multiple treatment comparisons. According to Daniels and Hughes [13] and Shanafelt et. al [135] different treatment comparisons and the use of active or inactive control interventions may influence the surrogate relationship. This could potentially be resolved by classifying treatment according the treatment class comparison (for example anti-angiogenic therapies versus chemotherapy) which potentially would lead to more treatment classes, but with reduced number of studies per class. To continue with the same issue, in this data-set the treatment classes were defined according to the class of the experimental treatment regardless of the control. Alternatively, we could classify them according to the treatment contrasts taking into account the class of the control group, however, this could result in fewer studies per class. A network meta-analysis model was developed for this problem by Bujkiewicz et al. [136].

Additionally, the evaluation framework proposed by Daniels and Hughes (see section 4.2.1.1) examines whether zero is contained in the CrIs of λ_1 and λ_0. However, the sparsity of data may lead to increased uncertainty around the intercept and slope. This increased uncertainty is also likely to manifest itself in increased conditional variance, thus invalidating the third criterion. Unsurprisingly, for sparse data it is unlikely that all the surrogacy criteria hold and this problem is more likely to occur in subgroup analyses. The proposed methods alleviate this problem as shown in some of the scenarios of the simulation study. However, we used the criteria mainly for the purpose of model comparison. In real life scenarios, when evaluating a potential surrogate endpoint for use in regulatory decision making or clinical trials, the decision of whether the surrogate endpoint should be used to predict clinical benefit or harm should be based on the balance between the strength of the surrogate relationship and the need for the decision to be made about the effectiveness of the new treatment [21]. Typically, the strength (or weakness) of the surrogate relationship is manifested in the width of the predicted interval of the treatment effect on the final outcome. i.e. a wider 95% interval of intercept and slope will imply a wider interval around the predicted effect and hence increased uncertainty about the regulatory or clinical decision made based on such prediction. This suggests that perhaps an evaluation framework should focus on the predictions [136]. The quality of predictions can be evaluated through a cross-validation procedure (see section 4.2.4).

A possible extension of these methods is to add another layer of hierarchy accounting for the different treatments within a treatment class. However, a relatively large number of studies for each treatment and number of treatments per class would be required to fit such model. As we mentioned in section 6.4, P-EX model could also be extended by making additional partial-exchangeability assumptions about the intercepts and the conditional variances, however, this may lead to over-parameterising the model. Furthermore, taking advantage of the setting proposed by Bujkiewicz et al. [87], both hierarchical models can be extended to allow for modeling multiple surrogate endpoints (or the same surrogate endpoint but reported at multiple time points) as joint predictors of treatment effect on the final outcome.

In summary, we developed hierarchical Bayesian methods for evaluating surrogate relationships within treatment classes whilst borrowing of information for surrogate relationships across treatment classes. We believe that the proposed methods have a lot of potential for improving the validation of surrogate endpoints in the era of personalized medicine, where the surrogacy may depend on the mechanism of action of specific targeted therapies.

Chapter 5

Improving the validation of surrogate endpoints on binary outcomes when the proportions of events occur rarely or very frequently

5.1 Introduction

This Chapter discusses about the methodological challenges in the trial-level validation of surrogate endpoints, when such validation is based on binomial aggregate data with high or low proportions of events.

Bivariate meta-analysis of treatment effects on a surrogate endpoint and a final outcome allows for the trial-level validation of a surrogate endpoint. In a bivariate meta-analysis of correlated outcomes, two sources of association exist in the data (one at the individual level and one at the study level). Specifically, within each study, the treatment effects on the two outcomes are measured on the same individuals and are therefore correlated (within-study correlation). Additionally, at the between-studies level, the between-studies variability on both outcomes (due to, for example, the differences in study population or treatment dose) generate correlation at the between-studies level (between-studies correlation) [137]. This correlation needs to be estimated with good precision and high accuracy in order to appropriately validate a candidate surrogate endpoint at the trial-level.

A standard way to validate the trial-level surrogate relationship is to perform a form of bivariate meta-analysis, such as BRMA[15, 88] and estimate the between-studies correlation parameter (or express the between-studies parameters in terms of other parameters describing surrogacy pattern such as intercept, slope and conditional variance). As discussed in Section 3.2.2, BRMA method models the treatment effects on both outcomes jointly with a bivariate normal distribution accounting for the within-study correlations. When this approach is applied to binomial data, the proportions of events are transformed to obtain treatment effects on log odds ratio scale which are assumed to be approximately normally distributed. However, when modeling binomial data on log odds ratio scale, the assumption of normality may not always be reasonable. Hamza et al. [17] showed that the normal approximation, used for binomial data in univariate meta-analysis of diagnostic test accuracy studies, leads to biased results, in particular when the proportions of events are very close to zero or one and the variance is large.

When trial-level validation of surrogate endpoints is based on data from modern clinical trials assessing personalized treatments, the high effectiveness of such targeted therapies results in large proportions of responders and very small proportions of progressions or deaths. Therefore, the assumption of normality when modeling binomial aggregate data on effectiveness of such therapies may lead to poor inferences about the parameters describing the surrogate relationship and may affect the trial-level validation of a surrogate endpoint.

To address this issue, we present two alternative meta-analytic methods for trial-level evaluation of surrogate relationships of the treatment effects on binomial outcomes. The first approach is a modification of a generalised linear mixed model (GLMM) applied to meta-analysis of diagnostic test accuracy studies [138]. It uses the exact independent binomial likelihoods across outcomes to model the within-study variability. This method, however, ignores potential within-study associations. In previous work, Riley et al. [139] highlighted the importance of taking into account the within-study correlation when using BRMA model.

To account for the within-study association on the original binomial scale, we developed a method which models the numbers of events on each outcome jointly using a bivariate distribution with binomial marginal distributions constructed with a bivariate copula. This model takes into account the within-study association between the numbers of events on the surrogate endpoint and the final outcome through the copula dependence parameter. This makes the model a more appropriate approach as the events on the surrogate endpoint and the final outcome are obtained from the same patients and therefore, they are correlated. Copulas are flexible tools for modeling bivariate/multivariate data, as they account for dependencies between multiple outcomes, allow for different dependence structures and for use of exact likelihoods, such as binomial or Poisson. In the past, copulas have been used to model individual-level surrogacy patterns modeling dependencies between, for example, time to event surrogate and final outcomes in IPD based methods[128].

We carried out a simulation study to investigate whether the proposed method improves the validation of surrogate endpoints compared to BRMA model using a normal approximation. It allowed us to investigate how sensitive were the estimates of the between-studies parameters (and in particular the between-studies correlation) were to assumptions made when modeling the within-study variability, and in particular when the proportions of events (such as responses to treatment or deaths) were close to zero or one. We also applied the methods (the standard BRMA model using log OR scale and the two proposed models) to a data example in CML, which consists of treatment effects from RCTs of targeted treatments.

CML is a myeloproliferative neoplasm of hematipoietic stem cells associated with a characteristic chromosomal translocation called the Philadelphia chromosome [140]. The main characteristic of CML is that it is regarded as a slow progressive disease [18]. Before the molecular pathogenesis of the disease was well understood, the median survival was 6 years, with a predicted 5-year overall survival (OS) of 47.2% [141]. However, the introduction of TKI therapies [142] have led to dramatically improved patients outcomes with high rates of complete cytogenetic response (CCyR) at 1 year and very few events at 2-year OS and EFS [4]. In the data example we assessed the whether CCYR at 1 year can be considered as a valid surrogate endpoint for EFS or OS at 2 years.

The existing and the proposed modeling approaches are introduced in Section 5.2 and the simulation study is presented in Section 5.3. A detailed description of the data example and the data analysis can be found n Section 5.4. The Chapter concludes with a discussion in Section 5.5.

5.2 Methods for trial-level surrogate endpoint evaluation on binomial data

In this section we present methods for evaluation of trial-level association patterns of potential surrogate endpoints, based on binomial aggregate data. Firstly, we recall the standard BRMA model and show how this model can be applied to binomial data. Secondly, we presents meta-analytic approaches for modeling binomial aggregate data on the original binomial scale, giving also a brief overview of the copula theory.

5.2.1 BRMA model

A standard way to investigate trial-level surrogate relationships is to carry out a form of bivariate meta-analysis, such as BRMA [15, 88] and estimate the between-studies correlation, which is the main parameter of interest as it quantifies the trial level association between the treatment effects on the surrogate endpoint and the treatment effects on the final (see surrogacy criteria for this model in section 3.2.2.1). BRMA method models correlated and normally distributed treatment effects on the surrogate endpoint and on the final outcome, and a detailed description of the method can be found in Section 3.2.2.

To apply this model to binomial data, the numbers of events (r_{1Ai}, r_{1Bi}, r_{2Ai}, r_{2Bi}) and the numbers of patients (N_{1Ai}, N_{1Bi}, N_{2Ai}, N_{2Bi}) in each arm, and for each outcome are transformed to obtain treatment effects (y_{1i}, y_{2i}) and their corresponding variances (σ^2_{1i}, σ^2_{2i}) on the log odds ratio scale, which are assumed to be approximately normally distributed (eq.5.1-5.4).

A modeling issue occurs when there are no events in either of the treatment arms as the log odds ratios (y_{1i}, y_{2i}) and their variances cannot be defined. A very simple way to tackle this problem is to apply a continuity correction, for instance, by adding 0.5. However, in some situations the effect of adding 0.5 may lead to biased results

[70, 71].

$$y_{1i} = log(\frac{r_{1Bi}}{N_{Bi} - r_{1Bi}}) - log(\frac{r_{1Ai}}{N_{Ai} - r_{1Ai}}) \tag{5.1}$$

$$y_{2i} = log(\frac{r_{2Bi}}{N_{Bi} - r_{2Bi}}) - log(\frac{r_{2Ai}}{N_{Ai} - r_{2Ai}}) \tag{5.2}$$

$$\sigma_{1i}^2 = \frac{1}{r_{1Bi}} + \frac{1}{N_{Bi} - r_{1Bi}} + \frac{1}{r_{1Ai}} + \frac{1}{N_{Ai} - r_{1Ai}} \tag{5.3}$$

$$\sigma_{2i}^2 = \frac{1}{r_{2Bi}} + \frac{1}{N_{Bi} - r_{2Bi}} + \frac{1}{r_{2Ai}} + \frac{1}{N_{Ai} - r_{2Ai}} \tag{5.4}$$

BRMA model accounts for the within-study correlation ρ_{wi} between the treatment effects on the surrogate endpoint and on the final outcome. As discussed in Section 3.2.2, when IPD are available, an estimate of ρ_{wi} can be obtained by bootstrapping. Otherwise an weakly informative prior distribution assuming the likely direction (positive or negative) can be placed on these parameters .

To implement the model in the Bayesian framework, the prior distributions should be specified on the unknown parameters. For instance, the following prior distributions can be placed on the heterogeneity parameters $\tau_{1,2} \sim U(0,5)$ and on the pooled treatment effects on the surrogate endpoint and the final outcome $d_1, d_2 \sim N(0, 10^2)$. To implement the natural constrain of the between-studies correlation $-1 \leq \rho_b \leq 1$, we used the Fisher's z transformation as, $\rho_b = tanh(z)$, $z \sim N(0,1)$. The implementation of the model in Stan can be found in the Appendix in Section C.1.

5.2.1.1 Criteria for Surrogacy

The primary parameter of interest is the parameter of between-studies correlation ρ_b as it establishes a strong association pattern between the treatment effects on the surrogate endpoint and on the final outcomes. For perfect surrogacy, the between-studies correlation should be ±1. However, in practice it is difficult either to achieve perfect surrogacy or to define a specific threshold for the correlation in order to consider the surrogate endpoint suitable for predictions. Typically, we expect the correlation to be relatively close to ±1. Additionally, it is important to ensure that no treatment effect on the surrogate endpoint will imply no effect on the final outcome - this suggest that the intercept should be very close to zero. Although, BRMA models the between-studies level without using the parameter of the intercept λ_0, it can be expressed in terms of the between-studies parameters (eq. 5.5) of BRMA [10].

$$\lambda_0 = d_2 - d_1 \rho_b \frac{\tau_2}{\tau_1}. \tag{5.5}$$

Therefore, we are able to draw inferences about the intercept by deriving λ_0 and checking whether or not a 95% CrI of λ_0 contains zero (and that it is relatively narrow).

5.2.2 Bivariate random effect meta-analysis with independent binomial likelihoods (BRMA-IB)

In this section, we present a bivariate meta-analytic model with independent binomial likelihoods for the first and the second outcomes at the within-study level. This approach is very similar to a standard model for meta-analysis of diagnostic test accuracy studies [138, 143] (where true positive and true negative observations are not correlated within a study as they are obtained from different patients). To adapt the model for diagnostic test accuracy studies (which are single arm studies) to the context of bivariate meta-analysis of RCTs, we assumed that the numbers of events r_{1Ai}, r_{2Ai}, in the control arm A and r_{1Bi}, r_{2Bi} in the experimental arm B, on the two outcomes (the surrogate and the final outcome respectively) follow independent Binomial distributions with the corresponding true probabilities of events p_{1Ai}, p_{2Ai}, p_{1Bi} and p_{2Bi}:

$$\begin{aligned} r_{1Ai} &\sim Bin(p_{1Ai}, N_{Ai}), \quad r_{2Ai} \sim Bin(p_{2Ai}, N_{Ai}), \\ r_{1Bi} &\sim Bin(p_{1Bi}, N_{Bi}), \quad r_{2Bi} \sim Bin(p_{2Bi}, N_{Bi}) \end{aligned} \tag{5.6}$$

At the between-studies model (5.7), the true probabilities of events are transformed using a link function $g(\cdot)$ (such as logit).

$$\begin{aligned} g(p_{1Ai}) = \mu_{1i}, \quad g(p_{1Bi}) = \mu_{1i} + \delta_{1i} \\ g(p_{2Ai}) = \mu_{2i}, \quad g(p_{2Bi}) = \mu_{2i} + \delta_{2i} \\ \begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_b \\ \tau_1\tau_2\rho_b & \tau_2^2 \end{pmatrix}\right) \end{aligned} \tag{5.7}$$

Where μ_{ji} are study specific baseline effects (i.e. the log-odds for the control group A and outcome $j = 1, 2$ in study i) while, δ_{ji} are the study specific true treatment effects on the log OR scale for outcome $j = 1, 2$ in study i and (d_1, d_2) are the pooled treatment effects on the surrogate endpoint and on the second outcome, τ_1 and τ_2 are the between-studies heterogeneity parameters and ρ_b the between-studies correlation.

To implement the model in the Bayesian framework, prior distributions need to be placed on unknown parameters which are, the baseline treatment effects $\mu_{1i,2i} \sim N(0, 10^2)$, the mean effects $d_{1,2} \sim N(0, 10^2)$, the between-studies standard deviations $\tau_{1,2} \sim U(0, 5)$, to implement the natural constrain of the between-studies correlation $-1 \leq \rho_b \leq 1$, we used the Fisher's z transformation as, $\rho_b = tanh(z)$, $z \sim N(0, 1)$. The Stan code of the model can be found in the Appendix in section C.2.

The key difference between this method and the BRMA method is the within-study model (eq. 5.6). Here, the within-study variability is modeled using the exact likelihood approach based on the binomial distribution avoiding to make the assumption of normality. Another advantage of this approach is that it does not to require continuity corrections. However, the model ignores the within-study association, which is restrictive as within each study the treatment effects on the two outcomes are measured on the same individuals and are therefore correlated. At the between-studies level, the between-studies variability on both outcomes generate the correlation at the between-studies level [137]. Therefore, when modeling aggregate data obtained from correlated binary outcomes, two sources of association exist and bivariate random-effects meta-analysis with independent binomial likelihoods (BRMA-IB) model accounts only for the second one.

In section 5.2.4 we propose a extension of the BRMA-IB model using a copula representation to model the within-study variability - in such a way to allow for the association between the numbers of events in each arm on the first and the second outcome to be taken into account. Copulas are flexible tools for modeling multivariate data as they account for the dependencies between multiple outcomes and allow for different dependence structures. Firstly, in section 5.2.3 we introduce some background on copula theory and then in section 5.2.4 we present the model based on copulas.

5.2.3 Overview of copula theory

This sections presents the main concepts of copulas for the bivariate case.

A copula is a bivariate cumulative density function (cdf) restricted to the unit square with standard uniform marginal distributions [144–146], which satisfies the following properties:

- $C(u_1, 1) = u_1$ or $C(1, u_2) = u_2$
- if $u_i = 0, \ \forall i \leq 2$ then $C(u_1, u_2) = 0$
- C is always monotonic to ensure that the joint probability is not be negative.
- C has to satisfy the Frechet-Hoeffding inequality. This means that copulas are bounded by:

$$max(u_1 + u_2 - 1, 0) \leq C(u_1, u_2) \leq min(u_1, u_2), \tag{5.8}$$

where the upper and lower Frechet-Hoeffding bounds describes perfect positive and negative dependence respectively.

If H is a bivariate cdf with univariate cdf margins F_1, F_2 then according to the Sklar's theorem [147] for every bivariate distribution, a copula representation C exists, such that:

$$H(x_1, x_2, \theta) = C(F_1(x_1), F_2(x_2), \theta). \tag{5.9}$$

The copula C is unique if F_1, F_2 are continuous random variables. However if some of the margins have discrete components , there are many possible copulas as emphasized by Genest and Neslehova [148], but all coincide on the closure of

$Ran(F_1) \times Ran(F_2)$ where $Ran(F)$ denotes the range of F. If H is continuous and $(X_1, X_2) \sim H$ then the unique copula is the distribution of $(U_1, U_2) \sim (F_1(x_1), F_2(x_2))$ leads to

$$C(u_1, u_2, \theta) = H(F_1^{-1}(u_1), F_2^{-1}(u_2), \theta), \quad 0 \leq u_j \leq 1, \quad j = 1, 2. \tag{5.10}$$

The joint probability density function (pdf) of the specified distribution H can be obtained using partial derivatives:

$$h(x_1, x_2, \theta) = \frac{\partial H(x_1, x_2, \theta)}{\partial x_1 \partial x_2} = c(F_1(x_1), F_2(x_1), \theta) f_1(x_1) f_2(x_2), \tag{5.11}$$

where $c(\cdot, \cdot)$ is the copula density distribution and f_1, f_2 are the univariate marginal density distributions. While the derivation of the joint density is easy for the continuous case through partial derivatives, it is not that simple in the discrete case. For the discrete variables, the joint probability mass function (pmf) is obtained using finite differences

$$\begin{aligned} h(x_1, x_2, \theta) = {} & C(F_1(x_1), F_2(x_2), \theta) - C(F_1(x_1 - 1), F_2(x_2), \theta) \\ & - C(F_1(x_1), F_2(x_2 - 1), \theta) + C(F_1(x_1 - 1), F_2(x_2 - 1), \theta). \end{aligned} \tag{5.12}$$

The key benefit of this theory is that avoids the assumption of normality when modeling non-normal data, allows for different dependence structures and provides a natural way to study and measure the dependence among variables. The correlation between two random variables x_1 and x_2 is captured by the dependence parameter θ.

5.2.3.1 Families of copulas

Having a variety of copulas can be extremely useful for building models having different properties such as heavy tails or asymmetries. More specifically, a bivariate copula C is symmetric if its density satisfies $c(u_1, u_2) = c(1 - u_1, 1 - u_2)$ for all $0 \leq u_1, u_2 \leq 1$. Otherwise, the joint density is asymmetric with more probability in the upper tail or the lower tail. Tail dependence is another useful copula-based measure indicating stronger dependence in extreme values.

In this section, we present three copulas which were used in this thesis. The first one belongs to the family of elliptical copulas, whereas the other two to the family of Archimedean copulas.

Elliptical copulas

The Elliptical copulas are simply the copulas of elliptically contoured distributions. The main advantage of elliptical copulas is that they allow for modeling the full range of correlation between the marginal distributions, however, they cannot be expressed in a close form.

The bivariate Gaussian copula is the most commonly used copula of the Elliptical family. It is a symmetric copula with weak tail dependence (see Figure 5.1c) and is given by:

$$C^G_\Sigma(u_1, u_2, \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\ \Sigma), \tag{5.13}$$

where $\Phi_2(\cdot|\Sigma)$ is the cdf of a bivariate standard normal distribution $N(0, \Sigma)$ with covariance matrix Σ and Φ^{-1} is the inverse cdf of the standard univariate normal distribution. The Gaussian copula interpolates from the Frechet lower bound $\rho \to -1$ (perfect negative dependence) to the Frechet upper bound $\rho \to 1$ (perfect positive dependence). Song et al.[149] showed that ρ is equal to Pearson correlation.

Archimedean copulas

A bivariate copula, constructed with a generator function ϕ, and specified as:

$$C(u_1, u_2, \theta) = \phi(\phi^{-1}(u_1, \theta) + \phi^{-1}(u_2, \theta), \theta) \tag{5.14}$$

is called Archimedian copula [144]. The generator function $\phi(u, \theta)$ is the Laplace transform of a univariate family of distributions of positive random variables and its inverse has a closed form. Archimedean copulas are very attractive as most of them allow for modeling wide range of dependencies, tail-dependencies, asymmetries and in contrast to elliptical copulas they can be expressed in a closed form.

Frank copula [150] is a symmetric copula without tail dependence (see Figure 5.1a), and it is given by:

$$C_F(u_1, u_2, \theta) = \theta^{-1} log\Big\{1 + \frac{(e^{-u_1\theta} - 1)(e^{-u_2\theta} - 1)}{(e^{-\theta} - 1)}\Big\}, \quad \theta \in (-\infty, \infty) \setminus \{0\}. \tag{5.15}$$

Frank copula interpolates from the Frechet lower bound $\theta \to -\infty$ (perfect negative dependence) to the Frechet upper bound $\theta \to \infty$ (perfect positive dependence) and hence, it is appropriate to model both kind of dependencies (negative and positive) between a surrogate endpoint and a final outcome.

Gumbel copula is an asymmetric copula with upper tail dependence (see Figure 5.1b). The bivariate Gumbel copula is given by:

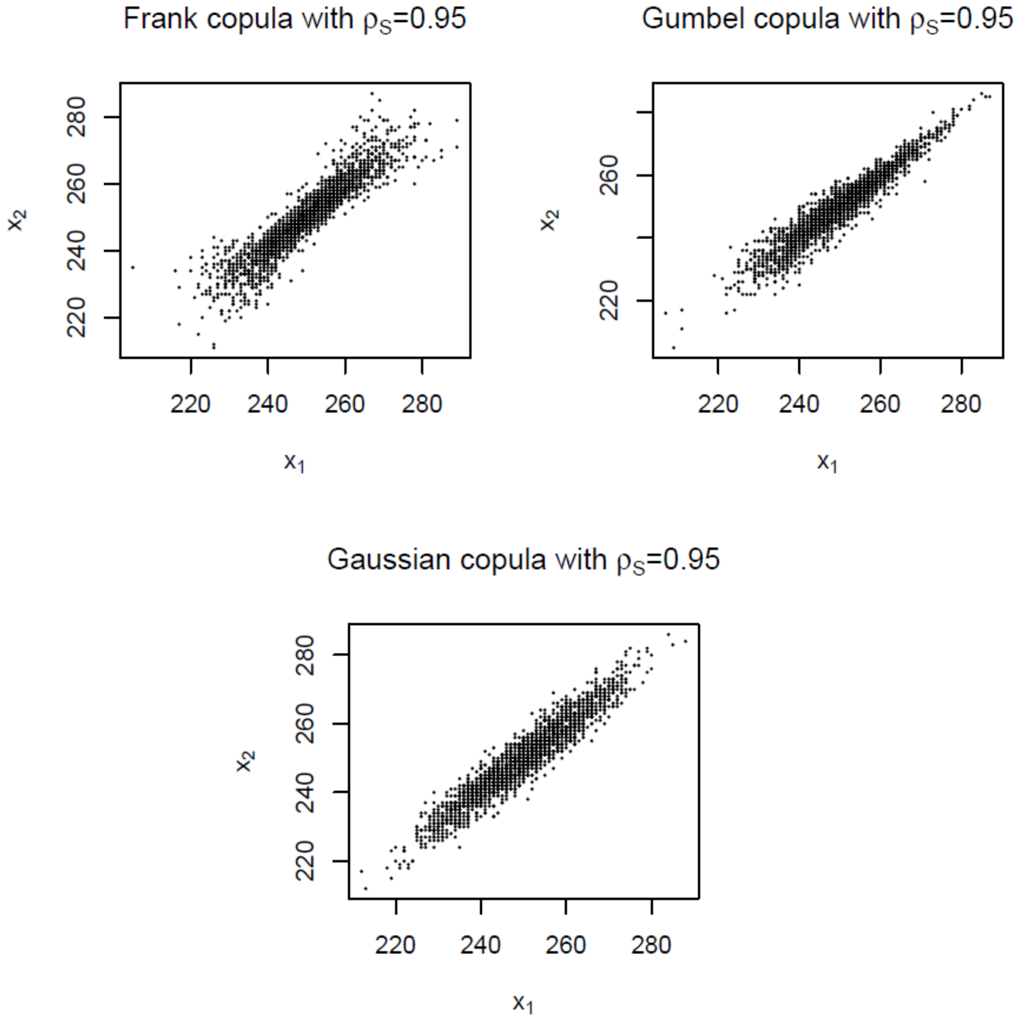
$$C_G(u_1, u_2, \theta) = exp\Big\{ - ((-log(u_1))^\theta + (-log(u_2))^\theta)^{\theta^{-1}} \Big\}, \quad \theta \in [1, +\infty) \qquad (5.16)$$

Gumbel copula interpolates from independence $\theta \to 1$ to the Frechet upper bound $\theta \to \infty$ (perfect positive dependence). Negative dependence in Gumbel copula can be introduced by rotating the copula function by 90° or 270°. For instance, the 90° rotated Gumbel copula is given by :

$$C_{90^\circ G}(u_1, u_2, \theta) = u_2 - C(1 - u_1, u_2, \theta) \qquad (5.17)$$

Figure 5.1 illustrates the dependence structure of each bivariate copula used in this thesis by simulating binomial data .

Figure 5.1: 2000 simulated samples of, a) Frank b) Gumbel and c) Gaussian copulas with Binomial marginal distributions ($x_{1,2} \sim Bin(p = 0.5, N = 500)$) and Spearman's correlation $\rho_s = 0.95$



5.2.4 Bivariate random effects meta-analysis with bivariate copulas (BRMA-BC)

BRMA-IB model assumes independence of the numbers of events across arms and outcomes and accounts only for correlation in the between-studies model. However, when modeling correlated binary outcomes (surrogate endpoint and final outcome) this assumption is too strong. As highlighted previously, at the within-study level, the numbers of events in each arm on the first and the second outcome are obtained from the same patients and are therefore correlated. Additionally, as discussed by Riley et al. [137], the heterogeneity of the treatment effects on both outcomes across studies generates the between-studies correlation. Hence, two sources of association exist in the data: at the within-study level and at between-studies level.

To account for the within-study association on the original binomial scale (without transforming the data to log odds ratios), the numbers of events on both outcomes should be modeled jointly, assuming association between them. This can be achieved by using a bivariate density function with binomial marginal distributions constructed with a bivariate copula representation, as copulas account for the dependence between marginal distributions and allow for modeling various dependence structures, providing a flexible representation of the bivariate distribution. Therefore, a joint density function constructed with copulas can be much more flexible compared to the bivariate normal distribution which only allows for normal marginals and a linear dependence structure.

$$\begin{pmatrix} r_{1Ai} \\ r_{2Ai} \end{pmatrix} \sim h(p_{1Ai}, p_{2Ai}, N_{Ai}, \theta_{Ai}) \quad \begin{pmatrix} r_{1Bi} \\ r_{2Bi} \end{pmatrix} \sim h(p_{1Bi}, p_{2Bi}, N_{Bi}, \theta_{Bi}) \tag{5.18}$$

$$g(p_{1Ai}) = \mu_{1i}, \quad g(p_{1Bi}) = \mu_{1i} + \delta_{1i} \quad g(p_{2Ai}) = \mu_{2i}, \quad g(p_{2Bi}) = \mu_{2i} + \delta_{2i}$$

$$\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_b \\ \tau_1\tau_2\rho_b & \tau_2^2 \end{pmatrix}\right) \tag{5.19}$$

where

$$\begin{aligned} h(r_{1\cdot}, r_{2\cdot}|p_{1\cdot}, p_{2\cdot}, N_{\cdot}, \theta_{\cdot}) = {} & C(F_1(r_{1\cdot}), F_2(r_{2\cdot}), \theta_{\cdot}) - C(F_1(r_{1\cdot} - 1), F_2(r_{2\cdot}), \theta_{\cdot}) \\ & - C(F_1(r_{1\cdot}), F_2(r_{2\cdot} - 1), \theta_{\cdot}) + C(F_1(r_{1\cdot} - 1), F_2(r_{2\cdot} - 1), \theta_{\cdot}), \end{aligned} \tag{5.20}$$

At the within-study level (eq. 5.18), we assume that the numbers of events in each arm on both outcomes follow bivariate distributions $h(p_{1i}, p_{2i}, N_i, \theta_i)$ with binomial marginal distributions. The parameters p_{1Ai}, p_{2Ai}, p_{1Bi}, p_{2Bi} denote the true probabilities of the numbers of events in each arm on the first and the second outcome, N_{Ai} and N_{Bi} are the number of patients in the control arm A and experimental arm B in trial i. Additionally, θ_{Ai}, θ_{Bi} are the dependence parameters in each arm respectively and they can be estimated when IPD are available. We assume that within-study dependencies are different across studies and hence, each study has a different dependence parameter. However, when IPD are not available across all studies, we assume the same dependence across them. In the absence of IPD, we can construct informative prior distributions, for example by combining evidence from external sources such as observational studies.

$F_1(r_{1\cdot})$ and $F_2(r_{2\cdot})$ are the cdfs of the binomial marginal distributions on the surrogate and the final outcome and $C(\cdot,\cdot)$ is the bivariate copula.

The between-studies model (eq. 5.19) is exactly the same as in BRMA-IB. The true probabilities of events p_{1Ai}, p_{2Ai}, p_{1Bi}, p_{2Bi} are transformed using a link function $g(\cdot)$ and the true treatment effects on both outcomes are normally distributed. This model was implemented in the Bayesian framework assuming the same prior distributions as for BRMA-IB. The Stan code of the model can be found in the Appendix in Section C.3.

Overall, bivariate random-effects meta-analysis with bivariate copulas (BRMA-BC) is less restrictive compared to BRMA and BRMA-IB, as it accounts for the within-study association and models the data on the original binomial scale, avoiding the a potentially inappropriate normal approximation.

5.3 Simulation study

We carried out a simulation study to assess the performance of BRMA model and the two proposed methods and in particular to investigate the impact of the assumptions made at the within-study level on estimates of the parameters at the between-studies level. Subsection 5.3.1 presents the data generation process and the simulation scenarios. The main estimands of the simulation study are reported in subsection 5.3.2. The section concludes reporting detailed results across the scenarios and discussing the key finding of the simulation study.

5.3.1 Simulation scenarios and generation process

We simulated data under 12 scenarios generating 1000 replications for each of them and varying the within-study association, the proportions of events and the numbers of participants. The proposed models were developed to model binomial aggregate data which can be a number of events out of a number of patients in each study i, in both arms. As within-study correlations ρ_{wi} and within-study dependence parameters θ_{Ai} and θ_{Bi} for the association between two outcomes in each study i are needed to populate the BRMA and BRMA-BC models respectively, we simulated data at the individual level (zeros and ones), as these parameters cannot be estimated from the aggregate data. All the models were fitted to the binomial aggregate data obtained from the IPD.

When investigating the impact of different modeling assumptions about the within-study variability on the model performance (in terms of estimating the between-studies parameters), we anticipated that such impact may depend on the strength of the within-study association. To explore this, we varied the strength of association by assuming weak, moderate and strong within-study associations (see details in step 6 of the generation process below). To test the effect of the magnitude of the proportions of events on the performance of the models, we considered two sets of scenarios - one with medium proportions and one with high proportions of events. This was implemented by varying the mean baseline treatment effects. In particular, baseline effects $\mu_{1i,2i}$ were drawn from the bivariate normal distribution (see details in step 3 of the data generation below). As the baseline effects were simulated on the *logit* scale, setting the mean baseline effects $\eta_{1,2} = 0$ corresponds to proportion of events equal to 0.5 (as $logit^{-1}(0) = 0.5$), and similarly, $\eta_{1,2} = 3$ corresponds to proportion of events equal to 0.95. Lastly, we considered two settings for study sizes. The number of patients in both arms of each study were drawn from the following normal distribution: $n_{Ai,Bi} \sim N(m, 5)$ where $i = 1, ..., N$ and rounded off to the nearest integer. Setting $m = 300$ (large study size) and $m = 80$ (small study size) covers the typical sizes of phase 3 and phase 2 trials in CML.

As mentioned in the previous paragraph, we generated IPD (zeros and ones) for each study and used bootstrapping to estimate the within-study correlations and association parameters for the copulas. However, in the scenarios with high proportions of events (95%) and small study size ($n_{Ai,Bi} \sim N(80, 5)$), it is likely to that some studies are generated without any non-events (zeros) both on the first and the second outcome. In such cases, the bootstrap method was unable to estimate the within-study association as the variability in the IPD is zero. To address this, we simulated studies with at least one 'zero value' either on the first or the second outcome.

The generation process is the following:

1. Set the number of studies to thirty ($N = 30$).

2. Simulate the heterogeneous arm sizes n_i of each study i from the following normal distribution ($n_i \sim N(m, 5)$) and then round them to the nearest integer.

3. Simulate the baseline treatment effects μ_{1i}, μ_{2i} from the following bivariate normal distribution $(\mu_{1i}, \mu_{2i})^T \sim BVN\left(\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix}, \begin{pmatrix}s_1^2 & s_1 s_2 \rho\\ s_1 s_{2i} \rho & s_2^2\end{pmatrix}\right)$, with $s_1 = s_2 = 0.1$, $\rho = 0.8$ and $\eta_{1,2} = 0$ (proportions of events equal to 0.5) or $\eta_{1,2} = 3$ (proportions of events equal to 0.95).

4. Simulate the true relative treatment effects from: $(\delta_{1i}, \delta_{2i})^T \sim BVN\left(\begin{pmatrix}d_1\\ d_2\end{pmatrix}, \begin{pmatrix}\tau_1^2 & \tau_1 \tau_2 \rho_b\\ \tau_1 \tau_{2i} \rho_b & \tau_2^2\end{pmatrix}\right)$, with $d_1 = 0.4$, $d_2 = 0.2$, $\tau_1 = 0.5$, $\tau_2 = 0.5$, $\rho_b = 0.8$.

5. Calculate the proportions of events from $p_{1Ai} = logit^{-1}(\mu_{1i})$, $p_{2Ai} = logit^{-1}(\mu_{2i})$, $p_{1Bi} = logit^{-1}(\mu_{1i} + \delta_{1i})$, $p_{2Bi} = logit^{-1}(\mu_{2i} + \delta_{2i})$ in each arm across outcomes.

6. To simulate (weakly, moderately and highly) correlated binary IPD, we used a joint density with Bernoulli marginal distributions constructed with Frank copula in both arms. For each set of proportions of events (0.5, 0.95) we varied the dependence parameters to reflect low, moderate and high within-study association. The true values of dependence parameters θ_A and θ_B along with the approximate value of corresponding Spearman's correlation were presented across all the simulated scenarios in the Appendix C.4.

7. Summarise the numbers of events in each arm and outcome by taking the sum of the binary responses and record the number of individuals in each study arm.

This process gives us a data-set with correlated numbers of events on the first and the second outcome in each arm.

To investigate the performance of the methods, we fitted BRMA, BRMA-IB and two versions of BMRA-BC to the simulated scenarios. Specifically, to assess effect of misspecifying the copula distribution on the estimates of the between-studies parameters, two versions of BRMA-BC model were applied. The first version (BRMA-BC($Frank$)) modeled the within-study variability using Frank copula - the same copula was used in the generation process (see step 6). The second version of the model, (BRMA-BC($Gauss$)) misspecified the dependence structure, modeling the within-study variability with Gaussian copula.

5.3.2 Estimands and performance measures

The primary estimand of the simulation study was the parameter of the between-studies correlation ρ_b. The second group of estimands of the simulation study were the heterogeneity parameters τ_1, τ_2, the pooled effects d_1, d_2 and the true treatment effects. These parameters could indirectly affect a trial-level surrogacy pattern, as the intercept λ_0 which is the second rule of the surrogacy criteria (see 5.2.1.1), was expressed in terms of the heterogeneity parameters and the pooled effects (eq. 5.5).

To evaluate the performance of the aforementioned models, in each simulation replication, we estimated the posterior median of the between-studies correlation ρ_b; 95% credible interval (CrI) of ρ_b; coverage probability of 95% CrIs of ρ_b and then we obtained values of bias of ρ_b averaged over 1000 simulation replications; RMSE of ρ_b across 1000 simulation replications. We also measured coverage, average bias and RMSE of the heterogeneity parameters τ_1 and τ_2, and the pooled treatment effects d_1 and d_2.

5.3.3 Results

This section presents the results of data analysis of the simulation study. Firstly, the estimated values of ρ_{wi}, $\theta_{Ai(Frank)}$, $\theta_{Bi(Frank)}$ and $\theta_{Ai(Gauss)}$, $\theta_{Bi(Gauss)}$ obtained by bootstrapping are presented across the scenarios. The second part of the section reports detailed results for the between-studies correlation ρ_b, the heterogeneity parameter for the final outcome τ_2, the pooled effect on the final outcome d_2. The section concludes discussing the key finding of the simulation study.

5.3.3.1 Within-study correlations ρ_{wi} and within-study dependence parameters θ_{Ai} and θ_{Bi}

As discussed, within-study correlations ρ_{wi} and within-study dependence parameters θ_{Ai} and θ_{Bi} for each study i were needed to populate the BRMA and BRMA-BC models respectively. Therefore, we simulated data at the individual level in order to estimate them. Table 5.1 displays the empirical distributions of ρ_{wi}, $\theta_{Ai(Frank)}$, $\theta_{Bi(Frank)}$ and $\theta_{Ai(Gauss)}$, $\theta_{Bi(Gauss)}$ of 30000 samples, obtained by bootstrapping the simulated IPD for each study across 1000 replications (30 studies$\times$ 1000 replications). The code and a short description of the bootstrap methods can be found in the Appendix C.5 and C.6.

Table 5.1: Medians, 2.5% and 97.5% quantiles of ρ_{wi}, $\theta_{Bi(Frank)}$, $\theta_{Bi(Frank)}$, $\theta_{Ai(Gauss)}$ and $\theta_{Ai(Gauss)}$ estimated by bootstrapping simulated IPD from all the studies (30 studies) and across 1000 simulation iterations

		Large study size		Small study size	
		Average Proportion of events = 0.5	Average Proportion of events = 0.95	Average Proportion of events = 0.5	Average Proportion of events = 0.95
Strength of association	Parameter	Median (2.5%, 97.5%)	Median (2.5%, 97.5%)	Median (2.5%, 97.5%)	Median (2.5%, 97.5%)
Low within-study association	ρ_{wi}	0.14 (0.02, 0.26)	0.16 (-0.00, 0.41)	0.14 (-0.08, 0.36)	0.16 (-0.14, 0.56)
	$\theta_{Ai(Frank)}$	0.87 (0.20, 1.55)	1.18 (-0.03, 2.66)	0.92 (-0.35, 2.30)	1.45 (-0.03, 5.29)
	$\theta_{Bi(Frank)}$	0.82 (0.12, 1.52)	1.02 (-0.02, 2.67)	0.86 (-0.42, 2.25)	1.27 (0.03, 5.37)
	$\theta_{Ai(Gauss)}$	0.15 (0.04, 0.27)	0.19 (-0.00, 0.41)	0.16 (-0.07, 0.36)	0.18 (-0.00, 0.55)
	$\theta_{Bi(Gauss)}$	0.14 (0.03, 0.26)	0.16 (-0.00, 0.40)	0.14 (-0.08, 0.36)	0.16 (0.01, 0.55)
Moderate within-study association	ρ_{wi}	0.44 (0.32, 0.54)	0.46 (0.11, 0.67)	0.44 (0.22, 0.62)	0.45 (-0.01, 0.82)
	$\theta_{Ai(Frank)}$	2.88 (2.10, 3.76)	3.62 (1.61, 6.32)	2.93 (1.48, 7.55)	3.73 (0.37, 29.66)
	$\theta_{Bi(Frank)}$	2.70 (1.78, 3.64)	3.06 (0.35, 6.28)	2.73 (1.22, 7.52)	3.14 (0.35, 30.97)
	$\theta_{Ai(Gauss)}$	0.45 (0.34, 0.55)	0.52 (0.26, 0.73)	0.45 (0.25, 0.64)	0.45 (0.05, 0.98)
	$\theta_{Bi(Gauss)}$	0.43 (0.29, 0.54)	0.45 (0.08, 0.73)	0.43 (0.20, 0.63)	0.38 (0.04, 0.98)
Strong within-study association	ρ_{wi}	0.78 (0.68, 0.84)	0.78 (0.45, 0.92)	0.78 (0.63, 0.88)	0.76 (0.27, 1.00)
	$\theta_{Ai(Frank)}$	7.61 (6.03, 9.81)	9.20 (4.99, 22.26)	7.72 (5.00, 13.77)	9.26 (1.90, 31.00)
	$\theta_{Bi(Frank)}$	6.61 (4.40, 9.01)	6.81 (2.34, 17.89)	6.61 (3.91, 11.73)	6.64 (0.93, 31.00)
	$\theta_{Ai(Gauss)}$	0.80 (0.73, 0.87)	0.85 (0.65, 0.97)	0.80 (0.66, 0.92)	0.86 (0.33, 0.99)
	$\theta_{Bi(Gauss)}$	0.75 (0.61, 0.84)	0.74 (0.24, 0.95)	0.75 (0.56, 0.89)	0.68 (0.11, 0.99)

5.3.3.2 Between-studies correlation ρ_b

The between-studies correlation was the main parameter of interest as it quantifies the trial-level association between the treatment effects on the surrogate endpoint and the final outcome.

Figure 5.2 displays posterior medians and 95% CrIs of ρ_b averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line). The plot on the left hand side (LHS), presents the results of the scenarios with large study size (the numbers of patients in both arms were simulated from $n_{Ai,Bi} \sim N(300, 5)$), whereas the plot on the right hand side (RHS) illustrates the results of the scenarios with small study size (the numbers of patients in both arms were simulated from $n_{Ai,Bi} \sim N(80, 5)$).

Starting from the scenarios where the proportions of events were on average 0.5 and the study size was large (LHS plot, first column), BRMA, BRMA-BC($Frank$) and BRMA-BC($Gauss$) models performed very similarly in terms of precision and accuracy regardless of the strength of the within-study association. They resulted in precise 95% CrIs and posterior medians very close to the true value (0.8). On the other hand, when the within-study association was moderate or strong, BRMA-IB model was the least accurate method overestimating between-studies correlation ρ_b.

The next set of scenarios (LHS plot, second column) include 0.95 average proportions of events and large study size. The BRMA-IB, BRMA-BC($Frank$) and BMRA-BC($Gauss$) models outperformed BRMA model in terms of precision. In particular when the within-study association was moderate and strong BRMA failed to estimate ρ_b with good precision (its 95% CrIs contained positive and negative values). However, BRMA-IB model was very sensitive to the effect of within-study association. The higher was the strength of the within-study association the more precise and less accurate the method was, resulting in accurate posterior medians only in the scenario with weak within-study association.

To investigate the effect of study size we repeated the same analysis reducing the number of patients in each study. The second plot in Figure 5.2 presents the results of the scenarios with small study size. Starting from the scenarios with 0.5 average proportions of events and the study size was small (RHS plot, first column), BRMA,

BRMA-BC(*Frank*) and BRMA-BC(*Gauss*) were less precise but equally accurate compared to the scenarios with large study size (LHS plot, first column) resulting in very similar posterior medians, but wider 95% CrI. On the other hand, BRMA-IB was more susceptible to the effect of study size in terms of accuracy compared to the other two methods. Specifically, when the within-study association was either moderate or strong in scenarios presented on the RHS plot, the method overestimated ρ_b resulting in larger posterior medians compared to the corresponding scenarios of the LHS plot and the true value.

The last set of scenarios (RHS plot, second column) corresponds to the proportions of events of 0.95 and small studies in terms of their size. In this extreme set of scenarios, all methods performed poorly in terms of estimating ρ_b. This was mainly due to the small study size combined with the high proportions of events. BRMA resulted in the least accurate posterior medians and the widest 95% CrIs. On the other hand, BRMA-IB was the most accurate and precise method.

Figure 5.2: Posterior medians (black dot) and 95% CrIs (solid bars) of ρ_b averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) across the 12 scenarios

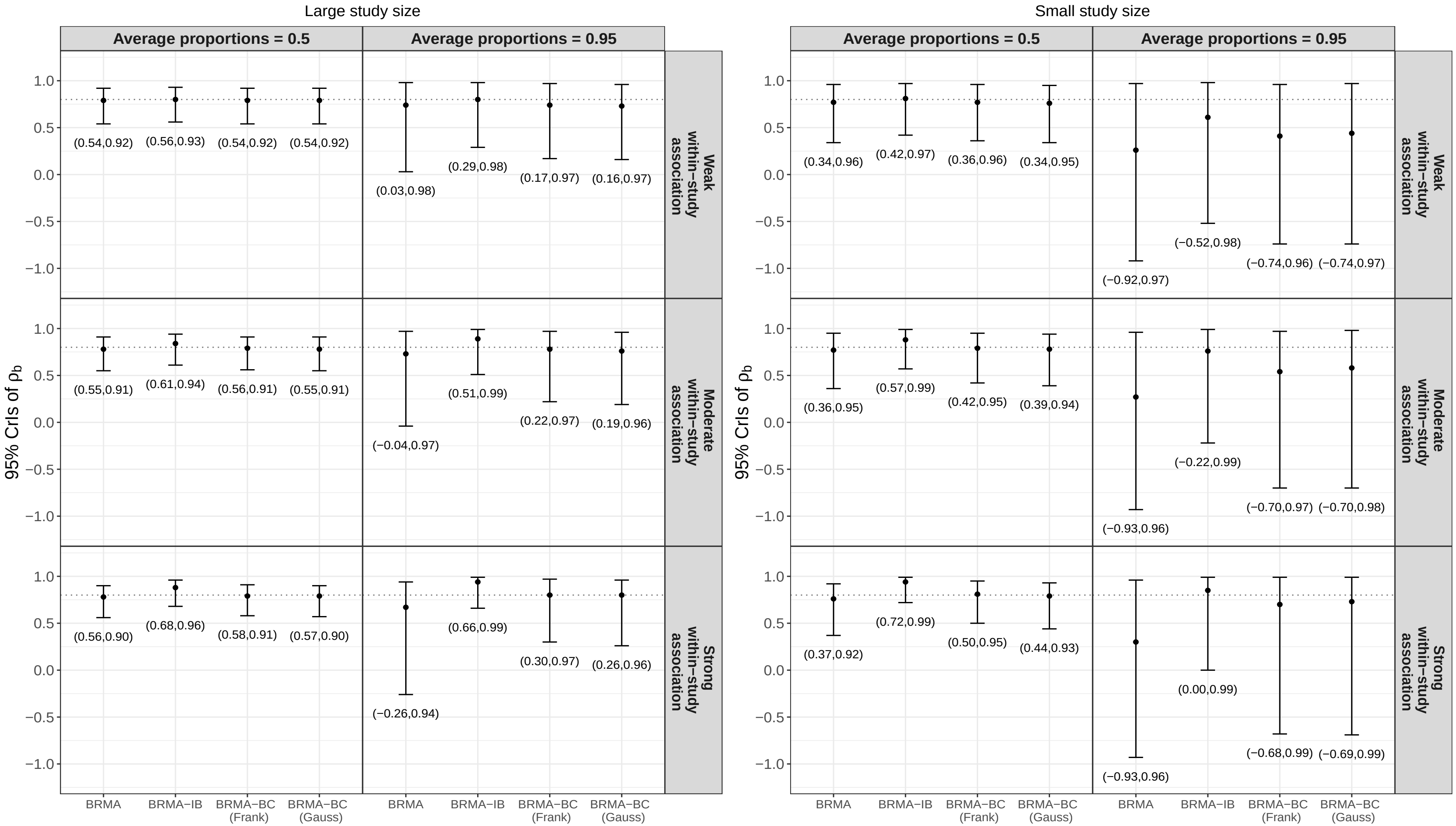


Figure 5.3 presents the bias of $\widehat{\rho}_b$ averaged over the 1000 replications along with the coverage probabilities of the 95% CrIs of ρ_b and RMSE across the 12 scenarios. In the scenarios with average proportions of events equal to 0.5 (first column of the LHS and RHS plots), BRMA, BRMA-IB, BRMA-BC($Frank$) and BRMA-BC($Gauss$) models performed very similarly in terms of bias, coverage and RMSEs regardless of the sample size. Specifically, there was no difference in their performance across the different strengths of within-study associations. On the other hand, when within-study association was moderate or strong, BRMA-IB resulted in upwardly biased estimates, slightly higher RMSEs and under-coverage. Concerning the effect of study size, the smaller was the study size the higher were the biases and RMSEs were across all methods.

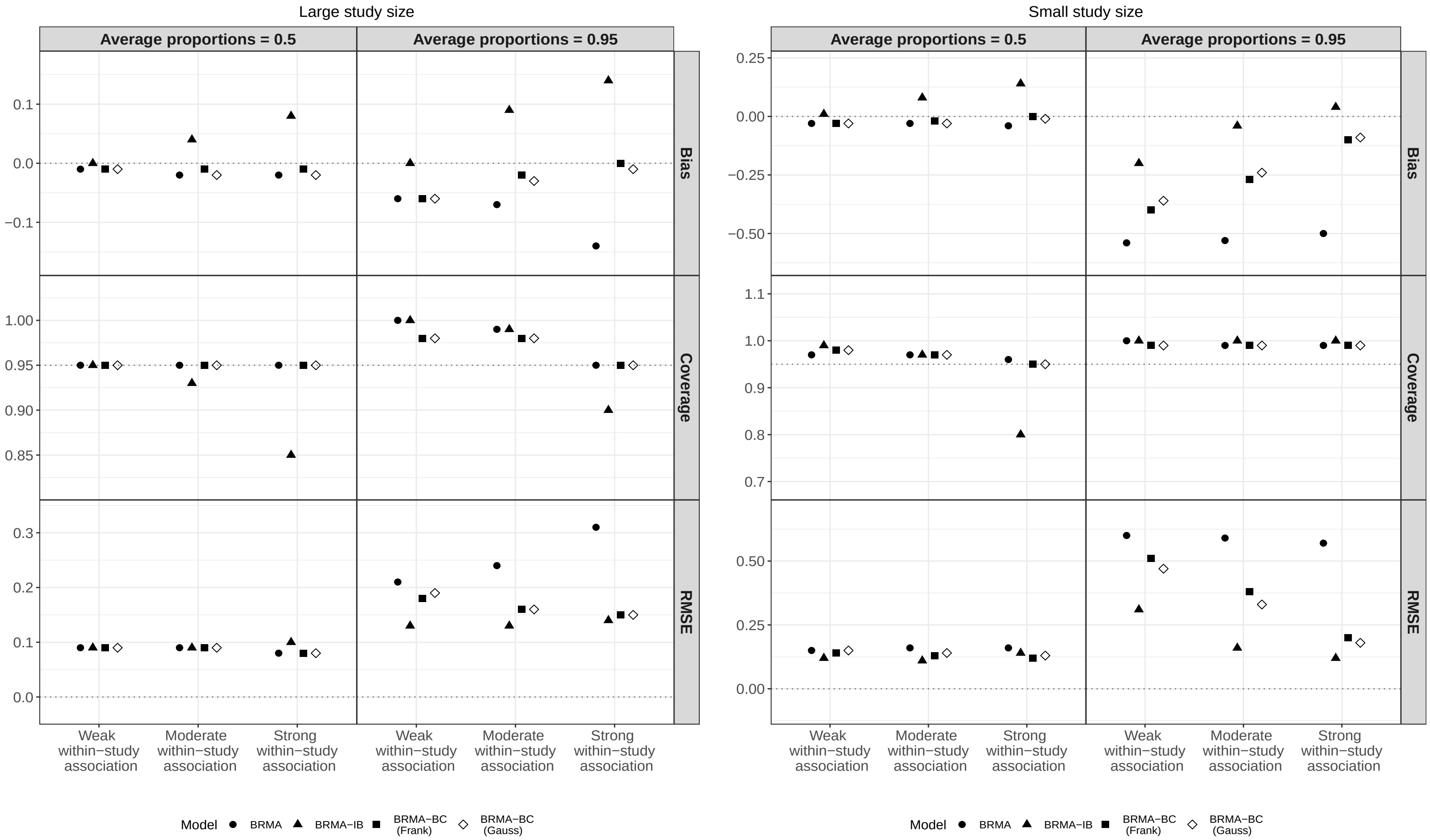
When the average proportions of events were 0.95 (second column of the LHS and RHS plots), BRMA-BC($Frank$), BRMA-BC($Gauss$) and BRMA-IB methods outperformed BRMA model across all scenarios regardless of the size of the studies. BRMA model underestimated substantially ρ_b in particular when the study size was small.

In the set of scenarios where the study size was large and the within-study association was moderate or strong (second column of the LHS), BRMA-BC($Frank$) and BRMA-BC($Gauss$) were less biased compared to BRMA-IB model resulting also in coverage probabilities closer to 95%. On the other hand the RMSEs of BRMA-BC($Frank$) and BRMA-BC($Gauss$) were slightly higher than RMSEs of BRMA-IB. This implies that the standard error of the estimates of BRMA-BC was larger compared to BRMA-IB despite being on average less biased across the 1000 replications (i.e. posterior medians were more dispersed around the true value). The posterior median of the between-studies correlation of BRMA-IB was upwardly biased when the study size was large and some under-coverage was also observed when the within-study association was strong. The second column of RHS plot presents the results of the scenarios with 0.95 average proportions of events and small study size. In this last set of scenarios, BRMA-IB was the best method in terms of bias and RMSE. The other three methods substantially underestimated the between-studies association, resulting in downwardly biased estimates of ρ_b and very conservative 95% CrIs as the coverage probabilities were higher than 95%.

However, BRMA-BC($Frank$) and BRMA-BC($Gauss$) models were always less biased compared to BRMA. BRMA model failed to estimate ρ_b as the assumption of normality was unreasonable in these scenarios.

The effect of misspecifying the copula density was minimal for the between-studies correlation ρ_b, since BRMA-BC($Gauss$) achieved similar performance as BRMA-BC($Frank$) across all the scenarios.

Figure 5.3: Bias of $\widehat{\rho}_b$ averaged over the 1000 replications along with the coverage probabilities and RMSE across the 12 scenarios



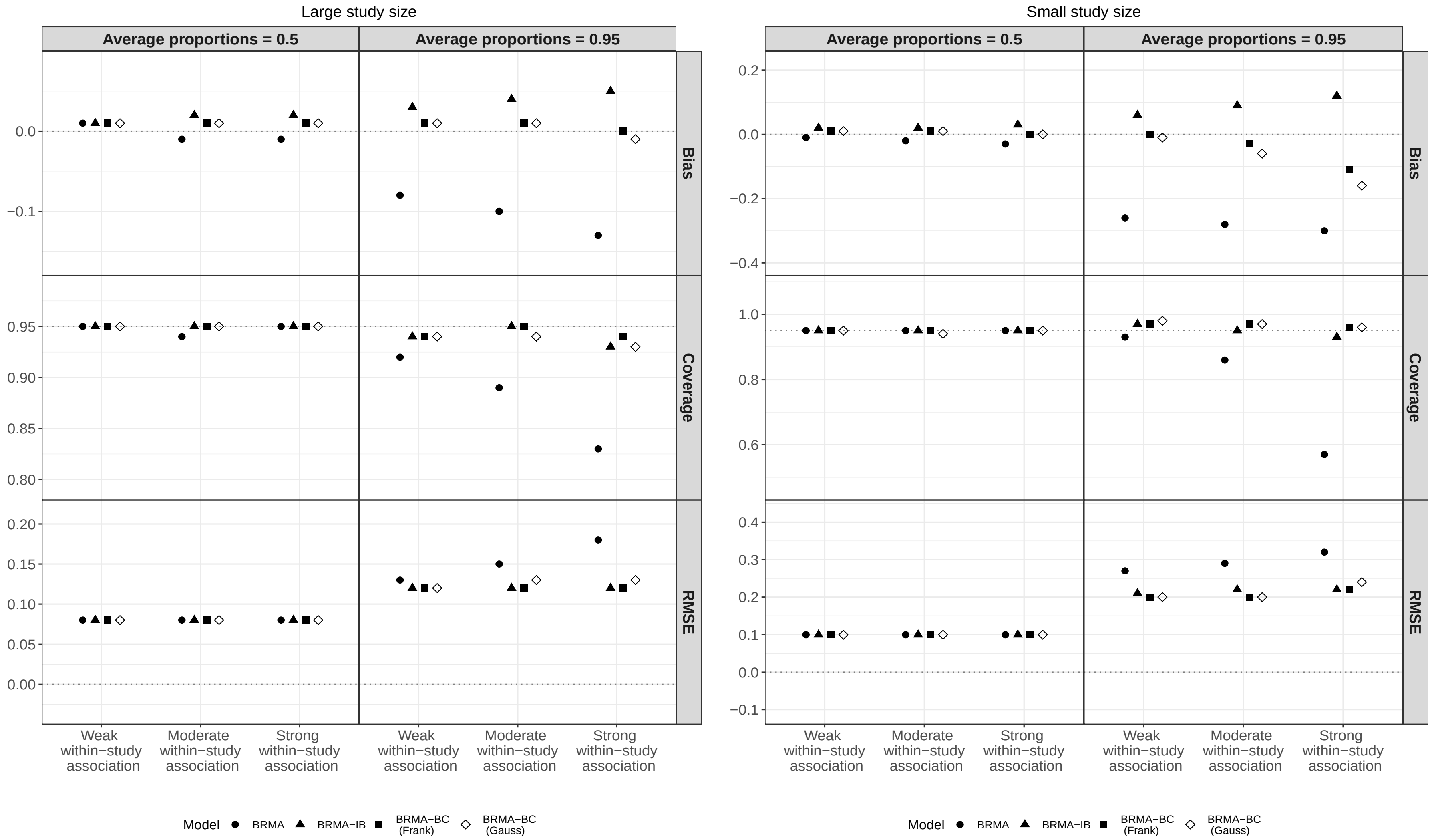
5.3.3.3 Between-studies standard deviation τ_2

In this section, we estimated the between-studies heterogeneity parameters τ_1, τ_2. We report only the results of $\widehat{\tau}_2$ as $\widehat{\tau}_1$ performed in a very similar way. Figure 5.4 presents the bias of $\widehat{\tau}_2$ averaged over the 1000 replications along with the coverage probabilities of the 95% CrIs of τ_2 and RMSE, across the 12 scenarios.

When the proportions of events were close to 0.5 (first column of the LHS and the RHS plots) all the methods were on average unbiased, with coverage probabilities equal to 0.95 and small RMSEs regardless of the study size. Only when within-study association was high, BRMA-IB slightly overestimated τ_2 resulting in higher on average biased estimates compared to BRMA-BC($Frank$), BRMA-BC($Gauss$) and BRMA models.

When the proportions of events were approximately 0.95 (second column of the LHS and the RHS plots), BRMA model substantially underestimated τ_2 across all strengths of within-study association regardless of the study size. Furthermore, substantial under-coverage was observed from BRMA model when the within-study association was moderate or strong regardless the sample size. Note that under- or overestimation of the heterogeneity parameters will affect the estimates of the between-studies correlation and vise-versa, which explains why ρ_b estimated from BRMA was downwardly biased. BRMA-IB overestimated the heterogeneity parameter τ_2 mainly when the within-study association was moderate or strong. This explains the upwardly biased estimates and the increased precision of the estimates of ρ_b from this method in these scenarios. The two versions of BRMA-BC were the most accurate methods across these scenarios, resulting in biases closer to zero, smaller RMSEs and acceptable coverage probabilities. Specifically, in the scenarios with moderate or strong within-study association, the estimates of τ_2 obtained from these two models, were slightly more biased compared to the scenario with weak within-study association. This effect was stronger for BRMA-BC($Gauss$) model, as this version of the model misspecified the copula function at the within-study level - it used the Gaussian instead of Frank which was used for data simulation.

Figure 5.4: Bias of $\widehat{\tau}_2$ averaged over the 1000 replications along with the coverage probabilities and RMSE across the 12 scenarios



5.3.3.4 Pooled treatment effects d_2

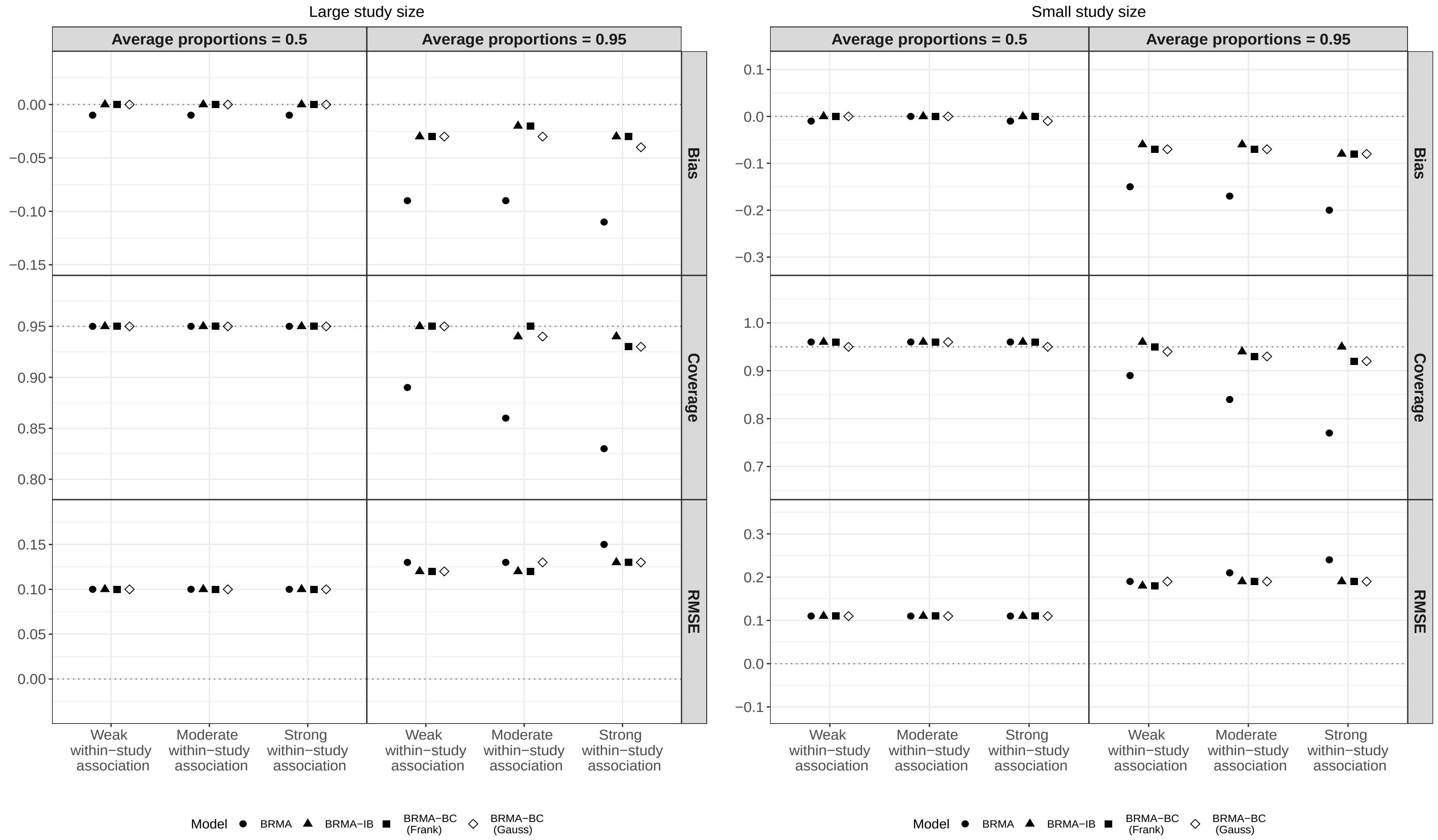
The last set of results, illustrates the performance of the methods in terms of the estimate of the pooled treatment effect on the final outcome d_2. Figure 5.5 presents the bias of $\widehat{d_2}$ averaged over the 1000 replications along with the coverage probabilities and RMSE across the 12 scenarios. Similarly as in the previous section, we decided to present only results of $\widehat{d_2}$, as the estimates of the treatment effect on the first outcome performed in a very similar way.

When the average proportions of events were 0.5 (first column of the plots on the LHS and the RHS) all the methods performed very well and in a very similar way achieving zero bias, 95% coverage probabilities and low RMSE regardless of the strength of the within-study association and the number of patients in each study.

When then the proportions of events were approximately 0.95 (second column of the plots on the LHS and the RHS), BRMA model gave downward biased estimates of d_2 and reduced coverage probabilities and marginally higher RMSEs compared to the proposed methods. Another interesting finding was the effect of within-study association on the estimates of d_2 of BRMA model. In the scenarios with small study size, the stronger was the within-study association the more downward biased were the estimates from BRMA. On the other hand, BRMA-IB and the two versions of BRMA-BC performed equally well across all the scenarios resulting in quite accurate estimates and acceptable coverage probabilities.

Overall, the effect of misspecifying the copula density was minimal for the pooled treatment effects on the final outcome d_2, since BRMA-BC($Gauss$) achieved similar performance as BRMA-BC($Frank$) across all the scenarios.

Figure 5.5: Bias of $\widehat{d}_2$ averaged over the 1000 replications along with the coverage probabilities and RMSE across the 12 scenarios



5.3.4 Key findings

A short summary of the key findings from the simulation study is given below:

- The simulation study showed that the normal approximation failed for binary outcomes when the proportions of events were simulated close to one. This confirmed findings by Hamza *et al.* [17] for univariate single-arm data and extends their finding to the bivariate setting for Binomial RCT data on two outcomes and two treatment arms. In our simulation study we focused on the performance of the parameters describing the between-studies variability: the between-studies correlation ρ_b and heterogeneity parameters τ_1, τ_2 and pooled treatment effects d_1 and d_2. When the average proportions of events were close to 0.5, there was no clear difference between BRMA model and the two version of BRMA-BC as they performed very similarly and sufficiently well in terms of the precision of ρ_b resulting also in very similar results across the performance measures (bias, coverage probability and RMSE). However, when the average proportions of events were increased to 0.95, BRMA model was not appropriate to investigate trial-level surrogate relationships for binary outcomes. This was reflected to the performance of the between-studies estimates of BRMA. Overall, the model resulted in inflated 95% CrIs of ρ_b, poor coverage probabilities, large RMSEs and downward biased estimates of ρ_b, $\tau_{1,2}$ and $d_{1,2}$.
- The main aim of the simulation study was to explore the impact of the within-study association on the estimation of the between-studies correlation ρ_b, as this parameter establishes a trial-level association pattern. As discussed above, BRMA model accounts for within-study association between the treatment effects on two outcomes. However, it is a suitable method for investigating a trial-level association pattern between treatment effects on binary outcomes, only when the proportions of events are close to 0.5, but will fail when the proportions are close to 0 or 1. On the other hand, BRMA-IB model was the most sensitive method to the effect of within-study association by far. This model assumes that the binomially distributed numbers of events are independent across outcomes. As a result, within-study associations are not taken into account and the ”excess” of the association manifests itself in

the upwardly biased estimate of the between-studies correlation. In the simulation study, higher within-study associations led to more upwardly biased estimates of ρ_b. Although, BRMA-IB resulted in substantial reduction in the uncertainty around the estimates of ρ_b, this improvement affected the coverage probabilities of the 95% CrIs of ρ_b causing major under-coverage in the scenarios with moderate or strong within-study association. In the extreme scenario with high proportions of events and small study size, BRMA-IB estimated ρ_b with better precision and accuracy compared to BRMA-BC due to the fact that overestimated the heterogeneity parameters. Overall, BRMA-IB model is quite robust when modeling data with no/small within-study association, but inappropriate to estimate a trial-level association pattern when the within-study association is moderate or strong.

- The simulation study also investigated the effect of the study size by having two sets of scenarios (one with small study size and one with large). Overall, in the scenarios with small study size, all the methods resulted in lower precision around the posterior medians of ρ_b, higher biases and larger RMSEs of the posterior medians of ρ_b compared to the scenarios with large study size. Furthermore, it highlighted the importance of study size in the scenarios with high proportions of events. Specifically, in the scenarios with average proportions of events equal to 0.95 and small study sizes, both versions of BRMA-BC failed to estimate the trial-level association with reasonable precision despite modeling the within-study variability on the original binomial scale and accounting for within-study associations. This indicates that, the study size is rather important and can substantially affect the precision and the accuracy of the estimates of the between-studies correlation, when investigating binary outcomes with very high/low proportions of events.
- Overall, BRMA-BC models were the most appropriate method to investigate the trial-level association patterns between treatment effects on two binary outcomes, in particular when the within-study association was strong. Both versions of the method achieved similar performance in most of the scenarios without substantially over/underestimating ρ_b, $\tau_{1,2}$ and $d_{1,2}$. This suggests that, the impact of misspecifying the copula density on the estimates of the

between-studies parameters was minimal. Additionally, there were scenarios where both versions of BRMA-BC failed to estimate ρ_b as accurately and precisely as BRMA-IB. As explained in the previous paragraph, this was due to the small size of the studies combined with the high proportions of events. In practice, investigating between-studies association between treatment effects on correlated binary outcomes with proportions of events close to one or zero requires studies with sufficiently large number of patients.

5.4 Data example

In this Section, we investigate whether the developed methods (BRMA-IB, BRMA-BC) improve the trial-level validation of surrogate endpoints compared to the standard method (BRMA) in a data example. Specifically, we applied BRMA, BRMA-IB and BRMA-BC, investigating the trial-level association pattern between the treatment effects on a surrogate endpoint and on a final outcome in CML. Another area of investigation was whether the dependence structure of the marginal distributions had an impact on the parameters describing the surrogate relationship (between-studies correlation and intercept). To investigate this, we applied three versions of BRMA-BC model. In the first version of BRMA-BC, the within-study variability was modeled with bivariate joint densities constructed with bivariate Frank copula which is symmetric and assumes no tail dependence. In the second one, the bivariate joint densities were constructed with Gaussian copula which is also symmetric and assumes weak tail dependence. In the last version, we used Gumbel copula which is asymmetric and assumes upper tail-dependence. We also investigated the predictions of the true treatment effects on the final outcome by carrying out a cross-validation procedure for each of the three models (BRMA, BRMA-IB and BRMA-BC).

5.4.1 Chronic myeloid leukemia and surrogate endpoints

CML is a myeloproliferative neoplasm of hematopoietic stem cells associated with the presence of a BCR-ABL fusion gene called the Philadelphia chromosome, which is the result of reciprocal translocation between chromosomes 9 and 22 [151]. The main characteristic is of the disease is that CML is regarded as a slow progressive disease [18]. Before the molecular pathogenesis of the disease was well understood, Philadelphia-positive CML was mainly treated with hydroxyurea, interferon alfa (INF-α) and allogeneic hematopoietic stem-cell transplantation. The median OS was 6 years, with a predicted 5-year OS of 47.2% [141]. The last 2 decades researchers established that BCR-ABL gene is the causal to the pathogenesis of CML and that tyrosine kinase activity is central to transform hematopoietic cells. TKIs treatments specifically target this activity [152], thus the introduction of first generation TKIs (imatinib) has led to dramatically improved long-term survival rates since 2001, resulting in high response rates of CCyR or major molecular response (MMR) at 1 year and very few events such as loss of response (e.g CCyR, MMR etc.), progression to accelerated phase or blast crisis and death from any cause. More recently, other drugs, most of them classifiable as second generation TKIs were developed [18, 19, 153]. When these agents are used as a first line treatment, they are capable of achieving faster and more durable CCyR and MMR [18, 19, 153, 154].

Most of the studies use biomarkers such as CCyR and MMR at 1 year as primary endpoints since, they are considered valid surrogate endpoints[4, 6, 18, 19, 19, 153, 155–157]. OS and EFS at 2 or 3 years are considered as secondary endpoints and are used as final outcomes in most of the RCTs. A systematic review and meta-analysis by Ciani et al. [154] confirmed the adoption of CCyR at 1 year as a surrogate endpoint of OS. On the other hand, they inferred that MMR did not fully qualify as a surrogate endpoint for OS as it provides a measure of success rather than a measure of failure, e.g. patients who do not achieve a deep molecular response (DMR) do not necessarily have a poor outcome. However, the surrogate relationship between treatment effects on CCyR at 1 year and on OS at 2 years was investigated by using a small number of RCTs which only compared first generation TKIs (imatinib) with standard treatments before the introduction of TKIs (e.g. INF-α). Lately, a variety of RCTs have compared first generation TKIs with second generation TKIs reporting high rates of CCyR at 1 year and very few events at 2 year OS or EFS in both arms.

5.4.2 Data extraction

To illustrate the proposed methods and compare them with BRMA model we identified 10 studies comparing first generation TKI therapies (e.g 400mg imatinib) with second generation TKIs (e.g. dasatinib, nilotinib,busotinib) or different doses of first generation TKIs (600mg or 800mg imatinib). We investigated whether CCyR at 1 year could be considered as a valid surrogate endpoint for EFS at 2 years or OS at 2 years. We chose CCyR at 1 year as candidate endpoint, as it has been extensively used in the literature as a gold standard for a good measure of response. EFS at 2 years as it is very significant in view of the dismal prognosis of the patients proceeding to advanced stages or losing response and OS at 2 years, since it is considered as the main long-term clinical outcome. Table 5.2 presents the summarised responses in the experimental and the control arm on both outcomes along with the sample size per arm and outcome. To apply BRMA, we transformed the treatment effects and their corresponding variances on the log odds ratio scale (using eq. 5.1-5.4). To work with positive correlations we also recorded the numbers of patients who were event-free on EFS and OS.

Table 5.2: Aggregate data in CML

	CCyR at 1 year				EFS at 2 years				OS at 2 years			
	Arm A		Arm B		Arm A		Arm B		Arm A		Arm B	
Study name	N_{1Ai}	r_{1Ai}	N_{1Bi}	r_{1Bi}	N_{2Ai}	r_{2Ai}	N_{2Bi}	r_{2Bi}	N_{2Ai}	r_{2Ai}	N_{2Bi}	r_{2Bi}
Cortes 2012 [158]	252	171	250	175	252	222	250	230	252	239	250	243
Kantarjian 2010 [19]	260	189	259	216	260	239	259	243	260	248	259	247
Radich 2012 [159]	61	42	70	59	123	117	123	118	123	121	123	119
Saglio 2010 [153]	243	184	236	219	283	267	281	276	283	272	281	275
Baccarani 2009 [160]	108	63	108	69	108	74	108	77	108	106	108	104
Preudhomme 2010[157]	158	92	160	104	159	149	160	149				
Hehlmann 2011 [161]	303	150	311	206	324	308	338	317	324	315	338	327
Cortes 2010 [162]	157	103	319	223	157	149	319	311	157	155	319	313
Deininger 2013 [156]	49	33	41	35	72	60	73	68	72	64	73	69
Wang 2015 [163]	133	107	134	104	133	125	134	124	133	131	134	132

5.4.3 Data Synthesis

Although BRMA and BRMA-BC models make different assumptions about the within-study variability, both of them account for the within-study association. As discussed above, within-study association between treatment effect on two outcomes on the log OR scale can be estimated using a bootstrap method from IPD. However, in this data-set IPD were not available for any of these studies, hence we were unable to estimate the dependence parameters θ_A and θ_B of each version of BRMA-BC and Pearson's within-study correlations ρ_w of BRMA. Instead, we constructed informative prior distributions for each of the parameters using external evidence obtained from three observational cohort studies [6, 155, 164] (these observational studies list three cohorts reporting data of treatments in the control arm and a cohort reporting data of a treatment in the experimental arm). The aim of these studies was to measure the impact of achieving a CCyR at 1 year on EFS or OS. They reported rates of CCyR at 1 year and the rates of EFS/OS at 2 years for the patients who either did or did not achieve CCyR at 1 year. Having this information, pseudo IPD could be generated for each of the studies, and hence the within-study associations could be estimated in each arm. To construct a unique informative prior distributions for each of the parameters (ρ_w, θ_A and θ_B) we performed the following steps:

1 we extracted the rates from each of the cohort studies and calculated an average rate in the control arm as there were three cohort studies reporting 3 different rates in this arm.

2 binary pseudo IPD were generated in each arm on the CCyR at 1 year and on EFS/OS at 2 years (vectors of "ones" and "zeros") using the extracted rates.

3 a double bootstrap method were applied to the binary pseudo IPD, to obtain an empirical distribution of the within-study association parameters (ρ_w, θ_A and θ_B)

4 unique informative prior $U(a, b)$ were constructed for each parameter, using the 2.5% and the 97.5% quantiles of the empirical distributions as boundaries.

Table 5.3 displays the median, the 2.5% and the 97.5% quantiles of the empirical distributions derived from the external evidence using the double bootstrap method (columns 3-5) and the constructed informative prior distribution placed on the parameters (column 6).

Table 5.3: Medians, 2.5% and 97.5% quantiles of the densities of ρ_w, θ_A and θ_B, estimated by a double bootstrap method on CCyR-EFS and CCyR-OS

		Empirical distributions			
Pairs of outcomes	Parameter	2.5%	Median	97.5%	Informative prior distributions
CCyR-EFS	ρ_w	0.20	0.33	0.45	$U(0.20, 0.45)$
	$\theta_{A(Frank)}$	0.99	1.71	2.38	$U(0.99, 2.38)$
	$\theta_{B(Frank)}$	1.11	1.92	2.78	$U(1.11, 2.78)$
	$\theta_{A(Gauss)}$	0.18	0.29	0.38	$U(0.18, 0.38)$
	$\theta_{B(Gauss)}$	0.19	0.32	0.42	$U(0.19, 0.42)$
	$\theta_{A(Gumbel)}$	1.09	1.17	1.27	$U(1.09, 1.27)$
	$\theta_{B(Gumbel)}$	1.09	1.19	1.31	$U(1.09, 1.31)$
CCyR-OS	ρ_w	0.00	0.12	0.23	$U(\ \ 0.00, 0.23)$
	$\theta_{A(Frank)}$	-0.15	0.61	1.27	$U(-0.15, 1.27)$
	$\theta_{B(Frank)}$	-0.10	0.74	1.50	$U(-0.10, 1.50)$
	$\theta_{A(Gauss)}$	-0.03	0.10	0.22	$U(-0.03, 0.22)$
	$\theta_{B(Gauss)}$	0.00	0.13	0.25	$U(\ \ 0.00, 0.25)$
	$\theta_{A(Gumbel)}$	1.00	1.03	1.11	$U(\ \ 1.00, 1.10)$
	$\theta_{B(Gumbel)}$	1.00	1.04	1.14	$U(\ \ 1.00, 1.14)$

5.4.4 Data analysis

We applied BRMA, BRMA-IB and three versions of BRMA-BC to investigate the trial-level association patterns between treatment effects on CCyR at 1 year and EFS at 2 years and between treatment effects on CCyR at 1 year and OS at 2 years in the CML data-set. To assess the performance of each method, we monitored the between-studies parameters and the intercept across models by estimating 95% CrIs, posterior medians and posterior mean for each parameter across the 5 modeling options. There primary parameters of interest were the between-studies correlation ρ_b and the λ_0 as they formed the surrogacy criteria in this chapter (see section 5.2.1.1). Additionally, we performed model comparison using the Watanabe-Akaike or widely applicable information criterion (WAIC) [165]. WAIC can be considered as an improvement on the deviance information criterion (DIC) [166] for Bayesian models and it was calculated across models using an R function developed by Vehtari et al. [167].

Table 5.4 displays estimates (posterior median/mean and 95% CrIs) of the between-studies parameters across models for CCyR-EFS pairs of outcomes. It can be seen that BRMA model resulted in a posterior distribution of ρ_b with significantly smaller posterior median (0.37) and marginally lower precision, compared to the estimates of BRMA-IB, BRMA-BC(*Frank*), BRMA-BC(*Gauss*) and BRMA-BC(*Gumbel*) models. In the simulation study, BRMA underestimated ρ_b resulting also in larger uncertainty around the estimate of ρ_b, compared to BRMA-IB and BRMA-BC when the proportions were close to 1. This is likely to occur in this data example, as the proportions of events on EFS where close to 1. On the other hand, BRMA-IB resulted in the highest posterior median of ρ_b. This can potentially mean that BRMA-IB slightly overestimated ρ_b as the model did not account for the two sources of association existed in the data (within and between-studies associations). Similar behaviour was also observed in the simulation study where the posterior medians of ρ_b of BRMA-IB were upwardly biased in the scenarios with moderate and high within study association. Moreover, BRMA model resulted in the smallest posterior means/medians of the heterogeneity parameters τ_1, τ_2 and of the pooled effects d_1, d_2, whereas BRMA-IB gave estimates with the largest values for these parameters. The three versions of BRMA-BC produced very similar posterior means/medians and 95% CrIs indicating that the dependence structure had negligible impact on the estimates of the parameters. Overall, we drew the same inferences about the trial-level association pattern between the treatment effects on CCyR at 1 year and EFS at 2 years regardless of the model we used. The between-studies correlation ρ_b was not very high and the parameter was estimated with considerable uncertainty across all methods (the 95% CrI spanned almost from -1 to 1). The intercept was also obtained with poor precision across all methods despite the 95% contained zero. Therefore, CCyR at 1 year could not be validated as surrogate endpoint of EFS at 2 years at the trial level.

Table 5.4: Between-studies estimates across models for CCyR-EFS pair of outcomes

Models	BRMA		BRMA-IB	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters				
ρ_b	0.23(0.37)	(-0.94, 0.98)	0.45(0.61)	(-0.79, 0.98)
λ_0	0.11(0.14)	(-0.40, 0.74)	0.10(0.13)	(-0.63, 0.63)
τ_1	0.40(0.38)	(0.11, 0.83)	0.46(0.43)	(0.15, 0.94)
τ_2	0.24(0.21)	(0.01, 0.74)	0.33(0.29)	(0.01, 0.85)
d_1	0.47(0.45)	(0.15, 0.83)	0.49(0.48)	(0.13, 0.86)
d_2	0.27(0.27)	(-0.05, 0.61)	0.30(0.30)	(-0.05, 0.69)

Models	BRMA-BC(*Frank*)		BRMA-BC(*Gauss*)		BRMA-BC(*Gumbel*)	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters						
ρ_b	0.35(0.51)	(-0.87, 0.98)	0.31(0.49)	(-0.91, 0.97)	0.31(0.48)	(-0.91, 0.98)
λ_0	0.17(0.18)	(-0.43, 0.70)	0.20(0.18)	(-0.43, 0.71)	0.14(0.18)	(-0.46, 0.71)
τ_1	0.43(0.41)	(0.14, 0.86)	0.43(0.42)	(0.12, 0.86)	0.43(0.40)	(0.11, 0.89)
τ_2	0.28(0.25)	(0.01, 0.78)	0.28(0.24)	(0.01, 0.80)	0.27(0.23)	(0.01, 0.80)
d_1	0.48(0.48)	(0.14, 0.83)	0.48(0.48)	(0.15, 0.85)	0.48(0.47)	(0.15, 0.83)
d_2	0.30(0.29)	(-0.04, 0.63)	0.30(0.29)	(-0.03, 0.65)	0.29(0.29)	(-0.03, 0.64)

Table 5.5 presents the same format of results as Table 5.4, obtained for CCyR-OS pair of outcomes. It can be seen that the same pattern also repeats in this analysis. BRMA resulted in the lowest posterior means/medians of the between-studies correlation ρ_b and the heterogeneity parameters τ_1 and τ_2. and BRMA-IB gave the highest estimates of these parameters, however, the differences between the estimates were less pronounced. The three versions of BRMA-BC produced very similar results to each other, indicating that modeling the within-study variability with different dependence structures did not affected the estimates of the between-studies parameters on CCyR-OS pair of outcomes. The estimates of the pooled treatment effects (d_2) on the second outcome were obtained with large uncertainty and their 95% CrIs included positive and negative values across all the methods. This is potentially due to the data being not sufficiently mature [19, 153, 158]. Overall, the trial-level association was very weak as the posterior means/medians were close to 0 implying poor trial-level surrogacy for this pair of outcomes in this data-set.

Table 5.6 list the values of the WAIC across models for both pairs of outcomes. The WAIC values of BRMA model is not included in the table as the model was fitted to the transformed data (numbers of events were transformed on the log odds ratio scale) and therefore, it could not be compared with the methods which were fitted to the binomial aggregate data directly.

Starting from CCyS-EFS pair of outcomes, the differences in the performance of the models were not that pronounced as the values of WAIC were not substantially different. BRMA-BC($Frank$) was the best fit to the data (smallest WAIC), whereas BRMA-IB was relatively poorer resulted in slightly higher value of WAIC. This suggest that BRMA-BC($Frank$) model performed better in this data-set as it accounted for both sources of association which existed in the data. This was also reflected to the estimates of the between-studies parameters which were very similar across models. Focusing on the values of WAIC across the three versions of BRMA-BC model, we could infer that symmetric dependence structures with no tail or weak tail dependence were slightly more appropriate modeling choices. When Gumbel copula (asymmetric copula with upper tail dependence) was used (BRMA-BC(*Gumbel*)), it resulted in higher WAIC compared to the other two versions of BRMA-BC (BRMA-BC($Frank$) and BRMA-BC($Gauss$)).

The results on CCyR-OS pair of outcomes suggested that all the models achieved similar performance and as the differences in the WAIC values were minimal. This suggest that BRMA-IB and the three versions of BRMA-BC were equally appropriate to model this pair of outcomes as the within-study association was very weak.

Table 5.5: Between-studies estimates across models for CCyR-OS pair of outcomes

Models	BRMA		BRMA-IB	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters				
ρ_b	0.09(0.15)	(-0.95, 0.96)	0.18(0.28)	(-0.92, 0.97)
λ_0	0.08(0.09)	(-0.69, 0.92)	0.03(0.06)	(-0.98, 0.80)
τ_1	0.49(0.45)	(0.16, 1.03)	0.52(0.48)	(0.17, 1.05)
τ_2	0.30(0.23)	(0.01, 1.05)	0.35(0.27)	(0.01, 1.12)
d_1	0.49(0.48)	(0.09, 0.92)	0.51(0.51)	(0.10, 0.96)
d_2	0.11(0.11)	(-0.38, 0.61)	0.13(0.13)	(-0.43, 0.60)

Models	BRMA-BC(*Frank*)		BRMA-BC(*Gauss*)		BRMA-CB(*Gumbel*)	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters						
ρ_b	0.15(0.24)	(-0.93, 0.96)	0.13(0.22)	(-0.94, 0.97)	0.14(0.22)	(-0.94, 0.97)
λ_0	0.06(0.08)	(-0.79, 0.82)	0.06(0.09)	(-0.82, 0.87)	0.05(0.09)	(-0.83, 0.82)
τ_1	0.50(0.46)	(0.19, 1.04)	0.51(0.47)	(0.17, 1.05)	0.52(0.47)	(0.17, 1.11)
τ_2	0.32(0.26)	(0.01, 1.03)	0.33(0.26)	(0.01, 1.05)	0.33(0.27)	(0.01, 1.03)
d_1	0.52(0.51)	(0.11, 0.95)	0.52(0.51)	(0.10, 0.96)	0.51(0.51)	(0.12, 0.94)
d_2	0.12(0.13)	(-0.34, 0.61)	0.13(0.14)	(-0.37, 0.59)	0.12(0.13)	(-0.39, 0.60)

Table 5.6: values of WAIC across models for each pair of outcomes

Pairs of outcomes	Model	WAIC
CCyR-EFS	BRMA	-
	BRMA-IB	255.9
	BRMA-BC(*Frank*)	252.4
	BRMA-BC(*Gauss*)	253.9
	BRMA-BC(*Gumbel*)	255.1
CCyR-OS	BRMA	-
	BRMA-IB	213.3
	BRMA-BC(*Frank*)	212.9
	BRMA-BC(*Gauss*)	212.4
	BRMA-BC(*Gumbel*)	213.8

5.4.4.1 Cross-validation procedure

Once a strong trial-level association between treatment effects on the surrogate endpoint and treatment effects on the final outcome is confirmed, the surrogate endpoint has to be assessed for its predictive value. To achieve this, a leave-one-out cross-validation procedure can be carried out (for details see section 4.2.4). In the CML data-set, the validation of CCyR at 1 year as a surrogate endpoint of OS at 2 years or EFS at 2 years was unsuccessful, as the trial-level association was weak and obtained with considerable uncertainty. However, we performed a cross-validation procedure to compare the performance of the models in terms of their predictions. Table 5.7 presents the performance of the predictions of the true treatment effect on the final outcome (EFS/OS) across models by reporting the following measures: the mean absolute error, which is defined as the absolute value of the difference between the observed mean treatment effect (measured on the log odds ratio scale) and the predicted treatment effect (estimated on the log odds ratio scale) averaged over the studies of the data-set, the performance of the predictive intervals, by checking whether the observed treatment effect of each study was contained in its corresponding predictive interval, and the mean ratio of the width of the 95% predictive intervals of BRMA-IB or BRMA-BCs models and the 95% predictive intervals of BRMA model averaged over the studies of the data-set.

The results of the cross-validation procedure on CCyR-EFS pair of outcomes showed that all methods performed equally well in terms of the performance of the 95% predictive intervals achieving perfect performance, as all the predictive intervals

Table 5.7: Performance of predictions across models

Pairs of outcomes	Models	Performance of 95% predictive intervals	Mean absolute error	Width ratio
CCyR-EFS	BRMA	1.00	0.43	1.00
	BRMA-IB	1.00	0.45	1.05
	BRMA-BC(*Frank*)	1.00	0.43	1.02
	BRMA-BC(*Gauss*)	1.00	0.42	1.03
	BRMA-BC(*Gumbel*)	1.00	0.43	1.03
CCyR-OS	BRMA	1.00	0.48	1.00
	BRMA-IB	1.00	0.50	1.08
	BRMA-BC(*Frank*)	1.00	0.50	1.08
	BRMA-BC(*Gauss*)	1.00	0.49	1.08
	BRMA-BC(*Gumbel*)	1.00	0.50	1.08

contained the observed treatment effect on the final outcome across methods. The three versions of BRMA-BC were slightly superior compared to BRMA in terms of mean absolute error, however they resulted in on average marginally wider 95% predictive intervals of the true effects on the final outcome compared to BRMA model.

Similarly, minor differences were observed in the performance of the models on CCyR-OS pair of outcome. However, modeling the within-study variability on the binomial scale (BRMA-IB and the three versions of BRMA-BC) resulted in on average 8% more uncertainty around the predictions of the true treatment effects on the final outcome compared to BRMA.

5.5 Discussion

We have introduced a new bivariate meta-analytic method (BRMA-BC) and modified an existing method (BRMA-IB) to investigate trial-level validation of surrogate endpoints when such validation is based on binomial aggregate data with high proportions of events. The proposed models improve the trial-level surrogate endpoint evaluation of binary outcomes. This can be particularly useful in diseases where the increased effectiveness of targeted treatments often leads to high numbers of responses and reduced numbers of events. The proposed models estimate trial-level association patterns with improved precision compared to the standard methodology (BRMA), as they allow for modeling the within-study variability on the original binomial scale, avoiding the use of an unreliable approximation of normality for log odds ratios. However, each method makes different assumptions about the within-study variability. BRMA-IB ignores potential within-study associations assuming that the aggregate data in both arms, on the surrogate endpoint and the final outcome follow independent binomial distributions. On the other hand, BRMA-BC accounts for within-study associations on the original binomial scale, modeling the aggregate data on each outcome jointly. This can be done by using bivariate distribution with binomial marginal distributions constructed with a bivariate copula. This model is very flexible as it can account for different dependence structures between the marginal distributions using different copulas.

BRMA-IB model performs well only when the within-study association is weak regardless of the size of the studies. In such scenarios, it can offer substantial gains in precision of the estimates of the parameters describing the surrogate relationship (in particular when the proportions of events are close to one or zero), resulting also in less biased estimates and smaller RMSEs compared to BRMA model. For instance, in the scenarios of the simulation study with weak within-study association, BRMA-IB was superior compared to BRMA in terms of precision of ρ_b. However, as the strength of the within-study association increases, the performance of the model becomes problematic. BRMA-IB ignores the within-study association and the "excess" of the association manifests itself in the upwardly biased estimate of the between-studies correlation. For example, in the scenarios where the within-study association was moderate or strong the model failed to estimate well the between-studies variability, giving upwardly biased estimates and low coverage probabilities of the between-studies correlation ρ_b and standard deviations $\tau_{1,2}$.

BRMA-BC is the most robust model to quantify the trial-level association regardless of the strength of within-study associations. In particular in the scenarios with 0.95 proportions of events, the model resulted in reduced uncertainty around the estimates of ρ_b compared to BRMA model. Furthermore, the fact that in the majority of the scenarios it did not over/underestimate the heterogeneity parameters $\tau_{1,2}$ led to more reasonable estimates of the between-studies correlation ρ_b compared to BRMA-IB. Although BRMA-IB estimated the between-studies correlation better in terms of precision, it resulted in poor coverage probabilities of the 95% CrIs of ρ_b and upwardly biased estimates of $\tau_{1,2}$ (in particular in the scenarios where the within-study association was strong). Additionally, in some scenarios, the performance of BRMA-BC highlighted the importance of the study size. Investigating trial-level association patterns in a data-set with small (in terms of the number of patients) studies and very high/low proportions of events, makes the estimation of the association pattern (surrogate relationship) extremely difficult. Therefore, when the proportions of events are very close to one or zero, large studies (in terms of size) are required to estimate trial-level association patterns accurately and precisely. An important aspect of the BRMA-BC is the choice of the copula function. In the simulation study, we investigated the effect of misspecifying the copula density on the parameter of between-studies correlation (correlated IPD were generated using Frank copula). Overall, the effect was minimal for the between-studies correlation ρ_b, as BRMA-BC($Gauss$) (where the within-study variability is modeled with the Gaussian copula) and BRMA-BC($Frank$) (where the within-study variability is modeled with Frank copula) achieved very similar performance across all the scenarios. This was potentially due to the sparsity of the data. In general, when data are sparse (as is often the case in meta-analysis of aggregate data) it is challenging to capture the exact dependence structure, and therefore different copula functions may have little impact on the performance of the between-studies parameters. However, it can always be useful to plot or perform diagnostics to the data in order to detect tail dependencies or asymmetries.

In the data example, all the methods found suboptimal trial-level association between the treatment effects on CCyR at 1 year and EFS at 2 years as the posterior median of between-studies correlation was not very high and the 95% CrI of ρ_b was extremely wide, spanning almost from -1 to 1. This suggests that CCyR cannot be considered as a valid surrogate endpoint of EFS at 2 years, at the trial level. However, this example can still illustrate the benefits of modeling the within-study variability on the original binomial scale. BRMA-IB and all versions of BRMA-BC model gave larger estimates of the median between-studies correlation ρ_b with slightly reduced uncertainty. Additionally, the median between-studies standard deviations τ_1, τ_2 were also higher compared to BRMA model. This behaviour is in agreement with the findings of the simulation study, where BRMA resulted in higher uncertainty around the estimates of the between-studies correlation and downwardly biased estimates of the heterogeneity parameters when the proportions of events were close to 1. Furthermore, by comparing the values of WAIC across models, we can infer that BRMA-BC with Frank copula was the most appropriate modeling techniques for this data-set. The three versions of BRMA-BC model resulted also in slightly more uncertainty around predictions of the true treatment effect on the final outcome compared to BRMA model. This was potentially due to the different assumptions at the within-study level, and the different scale of data (BRMA models the within-study level on the log odds ratio scale, whereas BRMA-BC at the original binomial scale).

A very weak trial-level association pattern between the treatment effects on CCyR at 1 year and the treatment effects on OS at 2 year was found regardless of the method we applied. The between-studies correlation and the intercept were estimated with very large uncertainty. Therefore, CCyR cannot be considered as a valid surrogate endpoint of OS at 2 years, at the trial level. However, this pair of outcomes was a good example to illustrate the performance of BRMA-IB and BRMA-BC models when the within-study association is very weak (this was reflected in the values of the informative priors placed on the within-study association parameters). In this case, BRMA-IB and BRMA-BC models resulted in very similar estimates of the between-studies parameters and almost identical values of WAIC. This suggest that, BRMA-BC model was able to perform equally well as BRMA-IB in a data-set with very weak association, as BRMA-BC was able to account for it.

Although BRMA-BC model provide robust results in a variety of scenarios, potential limitations should always be kept in mind. First, in order to perform Bayesian inference, we run HMC with RStan [168]. The model was very sensitive to initial values, making the initiation of the HMC process difficult. We solved this problem by fitting BRMA-IB or BRMA models first and then we used their estimates as initial values for BRMA-BC model.

A limitation of the data example was the lack of IPD. We informed the prior distributions of within-study association parameters using 3 cohort studies. We constructed binary pseudo IPD from external evidence and calculated the within-study association between the numbers responses on the surrogate endpoint and the numbers of events on the final outcome by using a double bootstrap method to account for uncertainty. Another limitation of the data example is the slightly inconsistent definition of EFS across these studies. For instance, some studies presented it as PFS, some others included more types of events in their definition than others.

BRMA-BC can be extended in a number of ways. For instance, it can be extended by using also a copula at the between-studies level in a similar way as Chu and Nikolopoulos have proposed [169, 170]. This will allow to model the trial-level association on the true scale (proportions of events) with beta marginal distributions avoiding the *logit* transformation. Furthermore, taking advantage of the setting proposed by Bujkiewicz et al. [87], BRMA-BC can be extended to allow for modeling multiple surrogate endpoints (or the same surrogate endpoint but reported at multiple time points) via a vine-copula.

In summary, we developed a new Bayesian hierarchical meta-analytic method and modified an existing method to perform bivariate meta-analysis of binary outcomes and particularly, to quantify the trial-level surrogate relationships between the treatment effect on binary outcomes. In our view, BRMA-BC is a preferred model for modeling binary outcomes in the context of surrogate endpoints. The model can improve the process of the validation of surrogate endpoints based on data from modern trials of personalised therapies where the increased effectiveness of targeted treatments such as TKIs often leads to high numbers of responses and reduced numbers of events.

Chapter 6

Improving the validation of surrogate endpoints by incorporating data from cohort studies

6.1 Introduction

This Chapter discusses about the methodological challenges in the trial-level validation of surrogate endpoints, when external evidence, such as non-comparative observational cohort studies, (OBs) need to be incorporated in the analysis for such validation.

Traditionally, when meta-analysing data from multiple studies, whether for purpose of obtaining pooled effects or for trial-level surrogate endpoint evaluation, the analysis has been based on data from RCTs. When treatment efficacy is of interest, RCTs are used as a gold standard as they achieve high internal validity due to randomisation and blinding [171–173] . However, very often RCTs exclude groups of population such as children, patients with comorbidities, making an assessment of clinical practice difficult [174–178]. On the other hand, observational cohort studies often have less restrictive inclusion criteria focusing mainly on the external validity, while their limited internal validity results in unreliable effectiveness estimates due to confounding factors and various biases, such as selection bias. Traditionally, researchers have been very sceptical about synthesis of evidence from different study designs, such as OBs, arguing that there may be strong dependence on assumptions and there is concern that including studies of "poor" designs will make the analysis weaker. Many authors commended caution when observational evidence is included in a meta-analysis, suggesting that a careful sensitivity analysis is always necessary regarding the plausibility of introducing observational evidence into the analysis [36].

However, more recently there has been an increasing interest in methods for inclusion of data from observational studies in evidence synthesis. Researchers are motivated by a number of factors when combining RCTs with OBs. For instance, inclusion of OBs can help to increase the power to detect a treatment effect when data from RCTs are too limited. This can be particularly important when policy decisions need to be made and further experimentation may be unfeasible due to time or budget constrains. A very detailed review of the most of popular methods for combing evidence from multiple sources was carried out by Verde et al. [179].

The confidence profile method (CMP) was the first statistical framework to combine evidence from different sources proposed by Eddy [180] and was used to in a series of clinical guidelines and clinical applications [180–184]. Cross-design synthesis was introduced in 1992 by the General Accounting office and described by Droitcour et al. [185] and Chelimsky [186]. However, the reliability of the methods was criticised by Begg [187], who pointed out that the authors have underestimated the problem of harmonizing results from medical registries and RCTs [179]. Begg and Pilote proposed a method to perform meta-analysis combining data from RCTs with historical controls [188]. It assumes that the baseline effect of each study (RCTs and observational cohorts) is random and the treatment effect is assumed to be constant. Verde et al. [189] presented a unified modeling framework (hierarchical meta-regression) to combine aggregated data from RCTs with IPD from observational studies. This framework allows for exploring treatment effects in specific patient populations reflected by the IPD and can potentially gain new insights from RCTs' results, which cannot be seen using only a meta-analysis of RCTs. Additionally, Verde et al. [190] evaluated the hierarchical meta-regression approach further using simulated data examples. They also presented a new R package called jarbes (just a rather Bayesian evidence synthesis), which implements their proposed framework in R. More recently, Verde et al. [191] presented a new Bayesian hierarchical model, called bias corrected meta-analysis model which combines different study types in a single meta-analysis but also accounts for the multiple biases that exist in such meta-analysis.

In this Chapter a new method for combining evidence from different sources is proposed, to improve the trial-level validation of surrogate endpoints in circumstances where RCT data offer limited evidence. Meta-analysing sparse RCT data may affect the validation of candidate endpoints as the estimates of the parameters describing the trial-level association between the surrogate endpoint and the final outcome are obtained with considerable uncertainty and poor accuracy. Furthermore, RCTs usually report observed treatment effects up to 2 years and very few provide long-term follow-up such as OS at 4 years or 5 years. Reporting short-term observed treatment effects can potentially affect the trial-level validation of surrogate endpoints as the estimates of the treatment effects on the final outcome are obtained with large uncertainty due to the data being not sufficiently mature. Immature data on the treatment effects can also affect the shape of the trial-level surrogate relationship between the treatment effects on the surrogate endpoint and the final outcome leading to poor inferences about the parameters describing the association. In this chapter, we extended the model proposed by Begg and Pilote [188], discussed in detail in section 6.2, to a bivariate case. This hierarchical method is designed to investigate trial-level surrogate relationships between the treatment effects on binary outcomes, whilst combining RCT data and observational evidence in a single model. The method models the observed treatment effects obtained from correlated binary outcomes with a Bivariate normal distribution and does not allow for adjusting for systematic biases across different types of designs. To account for such biases, a generalisation of the method was also developed. Similarly as in the extended model by Begg and Pilote, the generalised version of the model allows for adjusting for potential biases, i.e. systematic differences in the effectiveness estimates between the RCTs and OBs. Finally, an additional version of the method was introduced, presented in section 6.3.3, modeling the within-study variability on the original binomial scale as suggested in Chapter 5. In section 6.4, a simulation study was carried out to assess the performance of the method and in section 6.5 two data examples were presented to illustrate the method. The Chapter concludes with a discussion in section 6.6.

6.2 Begg and Pilote method

One of the first methods to incorporate historical controls into a meta-analysis was proposed by Begg and Pilote in 1991 [188]. The method was designed to model continuous and normally distributed outcomes, however, it can be used for binomial data when they are transformed to the log odds scale (data measured on this scale are assumed to be approximately normally distributed). Specifically, the observed treatment effects y_{Ai}, y_{Bi}, $i = 1, \ldots, n$ obtained from RCTs in arm A and B follow a bivariate normal distribution. In addition, to incorporate data from k single arm OBs reporting either arm A or B, the authors assumed that the observed treatment effects y_{Ai}, $i = n+1, \ldots, n+r$ and y_{Bi}, $i = n+r+1, \ldots, n+r+k$ are also normally distributed. The within-study model is given by:

RCTs:

$$y_{Ai} \sim N(\mu_i, \sigma^2_{Ai)}) \quad i = 1, \ldots, n \tag{6.1}$$

$$y_{Bi} \sim N(\mu_i + \delta, \sigma^2_{Bi}) \quad i = 1, \ldots, n \tag{6.2}$$

OBs$_A$:

$$y_{Ai} \sim N(\mu_i, \sigma^2_{Ai}) \quad i = n+1, \ldots, n+r \tag{6.3}$$

OBs$_B$:

$$y_{Bi} \sim N(\mu_i + \delta, \sigma^2_{Bi}) \quad i = n+r+1, \ldots, n+r+k \tag{6.4}$$

Where μ_i are the baseline effects, δ is a constant treatment effect and σ^2_{Ai}, σ^2_{Bi} are the corresponding within-study variances of the observed treatment effect of RCTs and OBs respectively. These variances are known and the can be calculated from the aggregate data.

The between-studies heterogeneity is reflected in the baseline effect assuming that

μ_i are random, normally distributed (equation 6.5) and account for differences in the baseline characteristics across RCTs and OBs.

$$\mu_i \sim N(\mu, s), \quad i = 1, \dots, n, n+1, \dots, n+r, n+r+1, \dots, n+r+k \tag{6.5}$$

On the other hand, the treatment effect δ is assumed to be constant across studies and study designs. Although this method was developed under the frequentist framework it can easily adapted to the Bayesian. Vague prior distributions can placed on the unknown parameters such as: $\delta \sim N(0, a)$, $\mu \sim N(0, a)$ and $s \sim U(0, b)$, where the constants a and b depend on the scale of the parameter and are considered sufficiently large.

A generalisation of the model was proposed to tackle potential limitations of this method with respect to the systematic biases in OBs. OBs do not have as strict inclusion criteria as RCTs and very often are prone to different kind of biases. When bias is present in OBs, it leads to biased the estimates of the baseline effect μ_i and the treatment effect δ, as OBs contribute to both parameters. To account for such these systematic differences between the RCTs and OBs, Begg and Pilote proposed adding bias terms at the within-study level of the model to account for them. Consequently, the within-study level of the model describing OB data becomes:

$$y_{Ai} \sim N(\mu_i + \eta, \;\; \sigma^2_{Ai}) \quad i = n+1, \dots, n+r \tag{6.6}$$

$$y_{Bi} \sim N(\mu_i + \xi + \delta, \;\; \sigma^2_{Bi}) \quad i = n+r+1, \dots, n+r+k \tag{6.7}$$

The terms η and ξ represent the biases in the OBs in arm A and B respectively. To implement this generalisation in the Bayesian framework a vague prior distributions can be placed on the bias parameters: $\eta \sim N(0, 100)$ and $\xi \sim N(0, 100)$.

A second extension of the this model was also briefly discussed. It allowed the treatment effects to vary across studies, assuming that they were exchangeable and normally distributed (eq. 6.8). The parameter d corresponded to the pooled effect and τ to the between-studies heterogeneity.

$$\delta_i \sim N(d, \tau) \tag{6.8}$$

6.3 Methods for trial-level surrogate endpoint evaluation when combining OBs with RCTs

This section presents the proposed methods to evaluate trial-level surrogacy patterns of potential surrogate endpoints when external evidence from observational cohort studies (OBs) need to be incorporated in a meta-analysis. Firstly, it introduces the extension of the method proposed by Begg and Pilote to the bivariate case. Secondly, it presents a generalisation of the new method which allows for adjusting for systematic biases across different types of designs and concludes with an additional version of the method which models the within-study variability on the original binomial scale as suggested in Chapter 5.

6.3.1 Extending Begg et al. method to the bivariate case (M1)

Bivariate meta-analytic methods provide a natural framework for combining evidence obtained from two outcomes and for modeling the trial-level surrogacy between the treatment effects on the surrogate endpoint and the final outcome. In this section, we extend to the bivariate case the method described in section 6.2, to allow for modeling two binary outcomes (surrogate endpoint and final outcome). This method can be applied to combine RCTs and OBs on the surrogate endpoint and the final outcome in order to improve the trial-level surrogacy when evidence from RCTs are limited. Firstly, y_{1Ai}, y_{2Ai}, y_{1Bi}, y_{2Bi} represent the observed treatment effects obtained from n RCTs in the control arm A, the experimental arm B, on the surrogate endpoint (1^{st} outcome) and the final outcome (2^{nd} outcome). These observed treatment effects in each study i, each arm (A or B) on the surrogate endpoint and the final outcome follow a bivariate normal distribution and are measured on an absolute scale such as log odds of an event The same setting and assumptions also apply to the observed effects obtained from OBs reporting data either in arm A or B on the surrogate endpoint (1^{st} outcome) and the final outcome (2^{nd} outcome). The within-study variability of the model is described in the following three parts:

RCTs:

$$\begin{pmatrix} y_{1Ai} \\ y_{2Ai} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{1Ai} & \sigma_{1Ai}\sigma_{2Ai}\rho_{wAi} \\ \sigma_{1Ai}\sigma_{2Ai}\rho_{wAi} & \sigma^2_{2Ai} \end{pmatrix}\right) \tag{6.9}$$
$$i = 1, \ldots, n$$

$$\begin{pmatrix} y_{1Bi} \\ y_{2Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} + \delta_{1i} \\ \mu_{2i} + \delta_{2i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{1Bi} & \sigma_{1Bi}\sigma_{2Bi}\rho_{wBi} \\ \sigma_{1Bi}\sigma_{2Bi}\rho_{wBi} & \sigma^2_{2Bi} \end{pmatrix}\right) \tag{6.10}$$
$$i = 1, \ldots, n$$

OBs$_A$:

$$\begin{pmatrix} y_{1Ai} \\ y_{2Ai} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{1Ai} & \sigma_{1Ai}\sigma_{2Ai}\rho_{wAi} \\ \sigma_{1Ai}\sigma_{2Ai}\rho_{wAi} & \sigma^2_{2Ai} \end{pmatrix}\right) \tag{6.11}$$
$$i = n+1, \ldots, n+r$$

OBs$_B$:

$$\begin{pmatrix} y_{1Bi} \\ y_{2Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} + \delta_{1i} \\ \mu_{2i} + \delta_{2i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{1Bi} & \sigma_{1Bi}\sigma_{2Bi}\rho_{wBi} \\ \sigma_{1Bi}\sigma_{2Bi}\rho_{wBi} & \sigma^2_{2Bi} \end{pmatrix}\right) \tag{6.12}$$
$$i = n+r+1, \ldots, n+r+k$$

where σ^2_{1Ai}, σ^2_{2Ai}, σ^2_{1Bi}, σ^2_{2Bi}, $i = 1, \ldots, n$ are the corresponding variances of the observed treatment effects and ρ_{wAi}, ρ_{wBi}, $i = 1, \ldots, n$ are the within-study correlations obtained from RCTs on the two outcomes in arm A and B. Similarly, σ^2_{1Ai}, σ^2_{2Ai}, $i = n+1, \ldots, n+r$ and σ^2_{1Bi}, σ^2_{2Bi}, $i = n+r+1, \ldots, n+r+k$, are the corresponding variances obtained from OBs on the two outcomes either in arm A or B and ρ_{wAi} $i = n+1, \ldots, n+r$, ρ_{wBi} $i = n+r+1, \ldots, n+r+k$, are the within-study correlations obtained from OBs.

The parameters μ_{1i}, μ_{2i}, $i = 1, \ldots, n$ correspond to the baseline effects estimated from RCT data, μ_{1i}, μ_{2i}, $i = n+1, \ldots, n+r$ are the baseline effects estimated from OBs reporting data in arm A and μ_{1i}, μ_{2i}, $i = n+r+1, \ldots, n+r+k$ are the baseline effects estimated from OBs reporting data in arm B. Similarly δ_{1i}, δ_{2i}, $i = 1, \ldots, n$ are the true treatment effects estimated from RCT data and δ_{1i}, δ_{2i}, $i = n+r+1, \ldots, n+r+k$ are the true treatment effects estimated from OBs reporting data in arm B. It is important to highlight that under this modeling approach, OBs for arm A, directly contribute only to the baseline effects $\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}$.

On the other hand, OBs for arm B, directly contribute both to the baseline effects and to the true treatment effects. $\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}$, $\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}$.

Here, as Begg and Pilote briefly discussed [188], we let the true treatment effects across RCTs and OBs to vary randomly from study to study as they assumed to be exchangeable. The random effects $\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix}$, $i = 1, \ldots, n, n+1, \ldots, n+k$, are modeled jointly following a bivariate normal distribution with the same mean and the same between-studies variance-covariance matrix. Additionally, the random baseline effects $\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}$, $i = 1, \ldots, n, n+1, \ldots, n+r, n+r+1, \ldots, n+r+k$ estimated across RCTs and OBs are also assumed to be exchangeable, normally distributed and uncorrelated with the true treatment effects. The between-studies level of the model is given by:

$$\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} s_1^2 & s_1 s_2 \rho \\ s_1 s_2 \rho & s_2^2 \end{pmatrix}\right) \tag{6.13}$$

$$\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1 \tau_2 \rho_b \\ \tau_1 \tau_2 \rho_b & \tau_2^2 \end{pmatrix}\right) \tag{6.14}$$

where ρ is the correlation between the baseline effects on the two outcomes, s_1^2, s_2^2 are the variances of the baseline effects, ρ_b is the between-studies correlation between the true treatment effects on the two outcomes and τ_1^2, τ_2^2 are the between-studies variances of the true treatment effect. To implement this model in the Bayesian framework non-informative prior distributions can be placed on the unknown parameters:$s_{1,2} \sim U(0,5)$, $\tau_{1,2}, \sim U(0,5)$, $\mu_{1,2} \sim N(0,100)$, $d_{1,2} \sim N(0,100)$, $\rho = \rho_b = tanh(z)$, $z \sim N(0,1)$. The Stan code of the model can be found in the Appendix in Section D.1.

Overall, this model can be used to perform bivariate meta-analysis based on RCT and OB data as it combines different study designs into a single meta-analytic method. It assumes that the baseline effects and the true treatment effects obtained from RCTs and OBs are exchangeable and normally distributed. To investigate the trial-level surrogacy patterns, we applied the surrogacy criteria, discussed in section 5.2.1.1. In the absence of OBs the method can be used to perform bivariate meta-analysis and to investigate trial-level surrogate relationship based only on RCTs, by using only the fist part (eq. 6.9, 6.10) of the within-study variability.

6.3.2 Accounting for bias in OBs (M2)

As discussed in sections 6.1, OBs are prone to various biases and suffer from many confounding factors, as their internal validity is very often poor. Therefore, it is crucial that models which combined evidence from different study designs, such as RCTs and OBs, account for bias. The model described in section 6.3.1 can account for bias in the OB data in the same way as Begg and Pilote proposed in their generalisation. Specifically, the part of M1 that describes the within-study variability of the observed treatment effects obtained from OBs becomes:

OBs$_A$:

$$\begin{pmatrix} y_{1Ai} \\ y_{2Ai} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} + \eta_1 \\ \mu_{2i} + \eta_2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{1Ai} & \sigma_{1Ai}\sigma_{2Ai}\rho_{wAi} \\ \sigma_{1Ai}\sigma_{2Ai}\rho_{wAi} & \sigma^2_{2Ai} \end{pmatrix}\right) \tag{6.15}$$
$$i = n+1, \dots, n+r$$

OBs$_B$:

$$\begin{pmatrix} y_{1Bi} \\ y_{2Bi} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} + \delta_{1i} + \xi_1 \\ \mu_{2i} + \delta_{2i} + \xi_2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{1Bi} & \sigma_{1Bi}\sigma_{2Bi}\rho_{wBi} \\ \sigma_{1Bi}\sigma_{2Bi}\rho_{wBi} & \sigma^2_{2Bi} \end{pmatrix}\right) \tag{6.16}$$
$$i = n+r1, \dots, n+r+k$$

where $\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$ and $\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$ represent the biases in OBs reporting arm A and B respectively, on the surrogate endpoint and the final outcome. These bias terms are assumed to be constant across studies in each arm and can be assigned non-informative prior distributions, such as: $\eta_{1,2} \sim N(0, 100)$, $\xi_{1,2} \sim N(0, 100)$. The between-studies level of the method remains exactly the same as presented in section 6.3.1 (eq. 6.13 and eq. 6.14). The implementation of the model in Stan can be found in the Appendix in Section D.2.

6.3.3 Modeling correlated binomial data using copulas (M3)

The proposed methodology described in sections 6.3.1, 6.3.2 was developed to model the observed treatment effects on the first and the second outcome jointly, using bivariate normal distributions. When M1 is applied binomial data at the aggregate level, the observed treatment effects can be represented at log of odds scale, assuming that they are approximately normally distributed. As discussed in chapter 5, the normal approximation used for binomial data often leads to biased results and underestimates the between-studies correlation, in particular when the proportions of events are close to one or zero. Therefore, a more appropriate way is to model the within-study variability (when the proportions are high/low) is by using the joint densities with binomial marginals constructed with copulas as described in section 5.2.4. Adapting model M1 to include such bivariate densities give the within-study model in the following form:

RCTs:

$$\begin{pmatrix} r_{1Ai} \\ r_{2Ai} \end{pmatrix} \sim h(p_{1Ai}, p_{2Ai}, N_{Ai}, \theta_{Ai}) \quad \begin{pmatrix} r_{1Bi} \\ r_{2Bi} \end{pmatrix} \sim h(p_{1Bi}, p_{2Bi}, N_{Bi}, \theta_{Bi}) \tag{6.17}$$

$$g(p_{1Ai}) = \mu_{1i}, \quad g(p_{2Ai}) = \mu_{2i} \quad g(p_{1Bi}) = \mu_{1i} + \delta_{1i}, \quad g(p_{2Bi}) = \mu_{2i} + \delta_{2i}$$

$$i = 1, \ldots, n$$

OBs$_A$:

$$\begin{pmatrix} r_{1Ai} \\ r_{2Ai} \end{pmatrix} \sim h(p_{1Ai}, p_{2Ai}, N_{Ai}, \theta_{Ai}) \tag{6.18}$$

$$g(p_{1Ai}) = \mu_{1i}, \quad g(p_{2Ai}) = \mu_{2i}, \quad i = n + 1, \ldots, n + r$$

OBs$_B$:

$$\begin{pmatrix} r_{1Bi} \\ r_{2Bi} \end{pmatrix} \sim h(p_{1Bi}, p_{2Bi}, N_{Bi}, \theta_{Bi}) \tag{6.19}$$

$$g(p_{1Bi}) = \mu_{1i} + \delta_{1i} \quad g(p_{2Bi}) = \mu_{2i} + \delta_{2i}, \quad i = n + r + 1, \ldots, n + r + k$$

$$\begin{aligned} h(r_{1\cdot}, r_{2\cdot}|p_{1\cdot}, p_{2\cdot}, N_{\cdot}, \theta_{\cdot}) = {} & C(F_1(r_{1\cdot}), F_2(r_{2\cdot}), \theta_{\cdot}) - C(F_1(r_{1\cdot} - 1), F_2(r_{2\cdot}), \theta_{\cdot}) \\ & - C(F_1(r_{1\cdot}), F_2(r_{2\cdot} - 1), \theta_{\cdot}) + C(F_1(r_{1\cdot} - 1), F_2(r_{2\cdot} - 1), \theta_{\cdot}). \end{aligned}$$

where, $F_1(r_{1\cdot})$, $F_2(r_{2\cdot})$ are the cdfs of the binomial marginal distributions on the surrogate endpoint (1st outcome) and the final outcome (2nd outcome), $C(\cdot, \cdot)$ is the bivariate copula, r_{1Ai}, r_{2Ai}, r_{1Bi}, r_{2Bi} are the numbers of events in each arm on the two outcomes, N_{1Ai}, N_{2Ai}, N_{1Bi}, N_{2Bi} are the number of patients for the two outcomes and each arm and study. $g(\cdot)$ is a link function and it is used to transform the true probabilities to the normal line scale. The between-studies model remains the same as in M1 (eq. 6.13, 6.14) and it can be implemented in the Bayesian framework by using the same prior distributions as in section 6.3.1. M3 can be generalised in the way as M1, to account for biases in the OBs. This can be achieved by adding bias terms in the following equations:

$$g(p_{1Ai}) = \mu_{1i} + \eta_1 \quad i = n + 1, \ldots, n + r \tag{6.20}$$

$$g(p_{2Ai}) = \mu_{2i} + \eta_2 \quad i = n + 1, \ldots, n + r \tag{6.21}$$

$$g(p_{1Bi}) = \mu_{1i} + \delta_{1i} + \xi_1 \quad i = n + r + 1, \ldots, n + r + k \tag{6.22}$$

$$g(p_{2Bi}) = \mu_{2i} + \delta_{2i} + \xi_2. \quad i = n + r + 1, \ldots, n + r + k. \tag{6.23}$$

The Stan code of the model can be found in the Appendix in section D.3.

6.4 Simulation study

A simulation study was carried out to assess the performance of models M1 and M2 and, in particular, the impact of observational data and bias on the estimates of the parameters describing the trial-level surrogacy patterns. To achieve this, throughout the simulation study, we compare results from meta-analyses where RCTs were combined with OBs with results from meta-analyses based only on RCT data, using either M1 or M2. Subsection 6.4.1 presents the generation process and the simulation scenarios. The main estimands of the simulation study are reported in subsection 6.4.2. The section concludes reporting detailed results across the scenarios and discussing the key finding of the simulation study

6.4.1 Simulation scenarios and generation process

We simulated data under 50 scenarios generating 1000 replications for each one. Firstly, we measure the effect of the number of studies on the trial-level surrogacy patterns, considering 4 sets of scenarios varying the number of RCTs and the number of OBs. In the first two scenarios, the number of RCTs was set to 5 and the number of OBs was either 4 or 10, whereas in the other two, the number of RCTs was 10 and the number of OBs was either 4 or 10. The first two sets of scenarios illustrate the situation where OBs supplement RCTs (RCT data are very sparse), aiming to validate a candidate endpoint as a surrogate endpoints at the trial level. The other two scenarios cover the case where RCT data offer sufficient information to validate an endpoint as a surrogate. In this situation, the incorporation of OBs aims to improve the precision and the accuracy of the estimates.

Secondly, to investigate whether the reporting arm in OBs (whether it is the control or the experimental arm or both) affects the estimation of the between-studies parameters, 3 sets of scenarios were constructed. The first scenario, includes RCTs and OBs reporting data only in arm A . The second one consists of RCTs and OBs reporting data only in arm B and the final one consists of RCTs and equal number of OBs reporting data in arm A and arm B. (for example, in the scenario with 5 RCTs and 10 OBs, 5 the OBs reported data in arm A, and the other 5 in arm B).

To measure the effect of the study size, the number of patients in each observational study was generated from a normal distribution: $n_i \sim N(m, 5)$ and rounded off to the nearest integer. The scenario with large OBs was generated by setting the mean of the normal distribution to $m = 400$ and the scenario with the small ones by simulating from $n_i \sim N(100, 5)$. Lastly, to test the effect of bias, we generated a set of scenarios no bias was present in the OBs and another one where systematic differences in the magnitude of the effects between RCTs and OBs were present.

In addition to the scenarios of data from mixed study designs, two scenarios were constructed including only RCT data. These scenarios were used as a reference for all the scenarios of the simulation study, allowing us to measure the added value of OBs in the analysis.

In total, 50 scenarios ($4 \times 3 \times 2 \times 2 + 2 = 50$) were generated.

The generation process of the RCTs was as follows:

1 Set the number of RCTs ($N_{RCTs} = 5$ or $N_{RCTs} = 10$).

2 Simulate heterogeneous arm sizes for each study i from: $n_{Ai} \sim N(100, 5)$, $n_{Bi} \sim N(100, 5)$ and then round them off to the nearest integer

3 Simulate the baseline treatment effects μ_{1i}, μ_{2i}, from a bivariate normal distribution: $\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} s_1^2 & s_1 s_2 \rho \\ s_1 s_2 \rho & s_2^2 \end{pmatrix}\right)$, with $\mu_1 = \mu_2 = 0$ (the effects are on the log odds scale, which corresponds to proportions of events equal to 0.5), $s_1 = s_2 = 0.1$, and $\rho = 0.8$

4 Simulate the true treatment effects δ_{1i}, δ_{2i}, from a bivariate normal distribution: $\begin{pmatrix} \delta_{1i} \\ \delta_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1 \tau_2 \rho_b \\ \tau_1 \tau_2 \rho_b & \tau_2^2 \end{pmatrix}\right)$, with $d_1 = 0.4$, $d_2 = 0.2$, $\tau_1 = \tau_2 = 0.5$ and $\rho_b = 0.8$

5 Calculate the proportions of events in each arm on the surrogate endpoint and the final outcome in study i: $p_{1Ai} = logit^{-1}(\mu_{1i})$, $p_{2Ai} = logit^{-1}(\mu_{2i})$, $p_{1Bi} = logit^{-1}(\mu_{1i} + \delta_{1i})$, $p_{2Bi} = logit^{-1}(\mu_{2i} + \delta_{2i})$

6 Simulate correlated binary IPD on the surrogate endpoint and the final outcome using a joint density (made with Gaussian copula) with Bernoulli marginals in both arms with dependence parameters $\theta_A = \theta_B = 0.6$.

7 Summarise the numbers of events in each arm, outcome and study by taking the sum the binary responses and then calculate the observed treatment effects (on the log odds scale) using: $y_{1Ai} = log(\frac{r_{1Ai}}{n_{1Ai} - r_{1Ai}})$, $y_{2Ai} = log(\frac{r_{2Ai}}{n_{2Ai} - r_{2Ai}})$, $y_{1Bi} = log(\frac{r_{1Bi}}{n_{1Bi} - r_{1Bi}})$, $y_{2Bi} = log(\frac{r_{2Bi}}{n_{2Bi} - r_{2Bi}})$

The generation process of the OBs is:

1 Set the number of OBs ($N_{OBs} = 4$ or $N_{OBs} = 10$).

2 Simulate heterogeneous OBs from $n_i \sim N(m, 5)$ and then round them off to the nearest integer ($m = 100$ or $m = 400$).

3 Simulate the baseline treatment effects μ_{1i}, μ_{2i}, from a bivariate normal distribution: $\begin{pmatrix}\mu_{1i}\\ \mu_{2i}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}s_1^2 & s_1 s_2 \rho\\ s_1 s_2 \rho & s_2^2\end{pmatrix}\right)$, with $\mu_1 = \mu_2 = 0$, $s_1 = s_2 = 0.1$ (corresponding to proportions of events equal to 0.5), and $\rho = 0.8$ (this step was used across all OBs regardless of which arm they reported)

4 Simulate the true treatment effects δ_{1i}, δ_{2i}, from a bivariate normal distribution: $\begin{pmatrix}\delta_{1i}\\ \delta_{2i}\end{pmatrix} \sim N\left(\begin{pmatrix}d_1\\ d_2\end{pmatrix}, \begin{pmatrix}\tau_1^2 & \tau_1\tau_2\rho_b\\ \tau_1\tau_2\rho_b & \tau_2^2\end{pmatrix}\right)$, with $d = 0.4$, $d_2 = 0.2$, $\tau_1 = \tau_2 = 0.5$ and $\rho_b = 0.8$ (this step was used only for OBs reporting arm B)

5 Calculate the proportions of events of OBs reporting arm A: $p_{1Ai} = logit^{-1}(\mu_{1i} + \eta)$, $p_{2Ai} = logit^{-1}(\mu_{2i} + \eta)$, calculate the proportions of events of OBs reporting arm B: $p_{1Bi} = logit^{-1}(\mu_{1i} + \delta_{1i} + \eta)$, $p_{2Bi} = logit^{-1}(\mu_{2i} + \delta_{2i} + \eta)$, with $\eta = 0$ for scenarios with unbiased OBs and with $\eta = 1$ for scenarios with biased OBs.

6 Simulate correlated binary IPD on the surrogate endpoint and the final outcome using a joint density (made with Gaussian copula) with Bernoulli marginals for OBs reporting either arm A or arm B with dependence parameters $\theta_A = \theta_B = 0.6$.

7 Summarise the numbers of events in each cohort study by taking the sum the binary responses and then calculate the observed treatment effects. The observed treatment effects from OBs reporting A are: $y_{1Ai} = log(\frac{r_{1Ai}}{n_{1Ai} - r_{1Ai}})$, $y_{2Ai} = log(\frac{r_{2Ai}}{n_{2Ai} - r_{2Ai}})$, the observed treatment effects from OBs reporting A are: $y_{1Bi} = log(\frac{r_{1Bi}}{n_{1Bi} - r_{1Bi}})$, $y_{2Bi} = log(\frac{r_{2Bi}}{n_{2Bi} - r_{2Bi}})$

As within-study associations for each study were needed to populate the models, we simulated data at the individual level (zeros and ones) for each RCT and OB. The parameters $\rho_w Ai$ and $\rho_w Bi$ were estimated by using a bootstrap method.(see details in the Appendix in section C.7).

As described in section 6.2, two parameterisations of the within-study level of the method were presented (M1, M3). M1 method, models the within-study variability on log odds scale assuming that the observed treatment effects are approximately normally distributed, whereas M3 method models the within-study variability on the exact binomial scale using joint densities with binomial marginals constructed with copulas. A generalisation of M1 and M3 were also presented (M2) which accounts for bias in OB data.

In the simulation study we assessed the performance of M1 and M2 and in particular, the impact of OBs and bias on the estimates of the parameters describing the surrogacy patterns. M1 model was applied to the data from scenarios where there was no bias present and M2 to the data from the scenarios assuming bias in OBs. In the reference scenarios, we fitted M1, using only the RCT part of the method (eq. 6.9, eq. 6.10, eq. 6.13).

Although M3 was not fitted to any of the data scenarios, if it had been used it would have achieved very similar performance as M1. This is because in the simulation study the proportions of events were generated close to 0.5 (see step 3), therefore, as discussed in chapter 5, both methods (M1, M3) could perform equally well in such scenarios. In practice, M1 was preferred over M3 for computational reasons - M1 runs approximately 10 times faster in Stan compared to M3.

6.4.2 Estimands and performance measures

The primary estimand of the simulation study was the parameter of the between-studies correlation ρ_b as it quantifies the trial-level association between the treatment effects on the surrogate endpoint and the final outcome. The second group of estimands of the simulations study were the pooled effects on the first and the second outcome. These parameters were rather important as they affect a trial-level surrogacy pattern since the intercept λ_0, which is the second rule of the surrogacy criteria (see 5.2.1.1), was expressed in terms of them and influenced the predictions of the true treatment effect on the final outcome.

To evaluate the performance of the aforementioned models, we calculated and monitored the following measures across all the simulated scenarios: posterior median of the between-studies correlation ρ_b in each simulation replication; 95% CrI of ρ_b in each simulation replication; and the absolute bias (i.e. $|\widehat{d_2}\text{-}d|$) was calculated in each replication.

6.4.3 Results

This section presents the results of the data analysis of the simulation study. It lists detailed results of the between-studies correlation ρ_b discussing two sets of scenarios - one without systematic bias across study designs and another one with OBs biased against the RCTs. Similar set of analyses are carried out for the performance of the methods on the relative treatment effect, which are discussed in section 6.4.3.2.

6.4.3.1 Between-studies correlation ρ_b

The between-studies correlation was the main parameter of interest as it quantified the trial-level surrogate relationship between the treatment effects on the surrogate endpoint and the final outcome.

Data scenarios without bias in the OBs

Figure 6.1 shows the posterior medians and 95% CrIs of ρ_b averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) in the scenarios where OBs were unbiased. On the LHS, the plot presents the same measures of ρ_b based on RCT data alone (5 RCTs) which were used as reference scenario.

Starting with the reference scenario, the estimate of between-studies correlation ρ_b was obtained with considerable uncertainty and poor accuracy as the RCT data were very sparse (only 5 RCTs were available). The average 95% CrI of ρ_b spanned from -0.64 to 0.98 and the average posterior median was $\hat{\rho}_b = 0.56$, substantially underestimating the true value ($\rho_b = 0.8$).

The scenarios on the RHS plot in Figure 6.1, included sparse RCT data and unbiased OBs. It can be seen that when OBs only report data in arm A, there was no improvement in the precision of the 95% CrIs of ρ_b and the accuracy of the posterior medians of ρ_b regardless of the size of OBs and the number of OBs included in the analysis. On the other hand, incorporating OBs reporting data only in arm B, resulted in reduced uncertainty and higher accuracy for the estimate of

ρ_b. Furthermore, the number of OBs included in the analysis had a considerable impact on precision and accuracy. When 10 OBs were available the average 95% CrI of ρ_b was considerably narrower compared to the scenarios where only four OBs were available. The size of the OBs was the least important factor in terms of improving the precision and the accuracy of the estimates. However, larger OBs resulted in, on average, slightly narrower 95% CrIs of ρ_b. For example, when 4 OBs where available and the mean size of the OBs was $n = 100$ the average 95% CrI was (-0.27,0.97), whilst when the mean size of the OBs was $n = 400$ the average 95% CrI was (-0.20,0.97).

The third column of the plot on the RHS of Figure 6.1, includes results of combining data from RCTs and OBs reporting data in both arms A and B. Specifically, in the scenario with 4 OBs, two of them reported data on arm A and two on arm B. Similarly, in the scenario with 10 OBs, five of them reported data in arm A and five in arm B. In this situation M1 yielded estimates of ρ_b with slightly larger uncertainty compared to the scenarios with OBs reporting data only in arm B, but substantially improved the precision of estimates of ρ_b compared to the reference scenario. The improvement was substantially higher compared to the scenario where OB data were available only for in A. Similarly as in the previous scenarios, the total number of OBs included in the analysis and the size of the OBs had an impact on the precision and the accuracy of the estimates of ρ_b. In general, when more and larger (in terms of the number of patients) OBs were generated, they resulted in narrower 95% CrIs of ρ_b and more accurate average posterior medians.

Figure 6.2 presents the same format of results including 10 RCTs and unbiased OBs. Therefore, the reference scenario in Figure 6.2, consisted of 10 RCTs alone. In this case the RCT data offered enough information, allowing M1 to estimate the between-studies correlation with relatively good precision and accuracy. The average 95% CrI of ρ_b spanned from -0.01 to 0.97, being much narrower compared to the 95% CrIs obtained from the previous reference scenario with 5 RCTs. The average median $\widehat{\rho_b} = 0.73$ was also more accurate compared to the case with 5 RCTs only.

A very similar pattern was seen when 10 RCTs were included in the analysis (Figure 6.2, RHS plot). Incorporating OBs reporting data only in arm A in the analysis did not improve the inferences about between-studies correlation ρ_b regardless of the number and the size of OBs. Specifically, in the first column of the RHS plot in Figure 6.2, the 95% CrIs of ρ_b were identical to the 95% CrIs obtained from the reference scenario where only RCT data were available (Figure 6.2, LHS plot). When OBs reported data either only in arm B or in both arms, it resulted in reduced uncertainty around the estimates. In both sets of scenarios the 95% CrIs of ρ_b were narrower compared to those in the scenario where OBs reported only arm A or the reference scenario. However, the 95% CrI of ρ_b obtained for the scenario where OBs consisted of only arm B, were the most precise.

The key difference between the set of scenarios presented in Figure 6.1 and the set of scenarios presented in Figure 6.2 was the number of available RCTs (5 RCTs available in Figure 6.1, 10 RCT in Figure 6.2). It can be seen that when RCT data were sparse the inclusion of OBs in the analysis had larger impact on the precision and the accuracy of the estimate of ρ_b compared to the case where sufficient number of RCTs was available.

Figure 6.1: Posterior medians (black dot) and 95% CrIs of ρ_b (solid bars) averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) in the scenarios with 5 RCTs and unbiased OBs

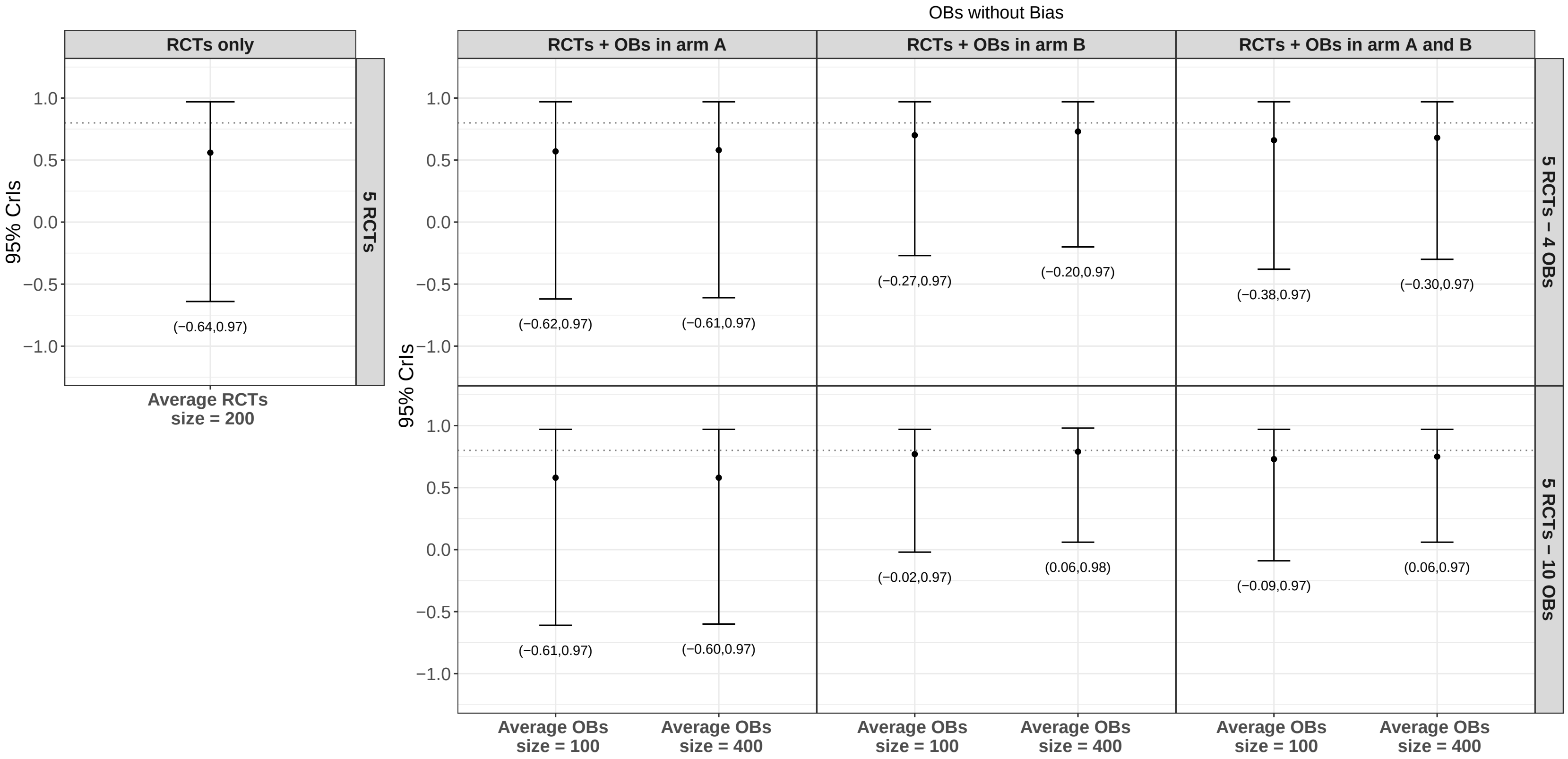
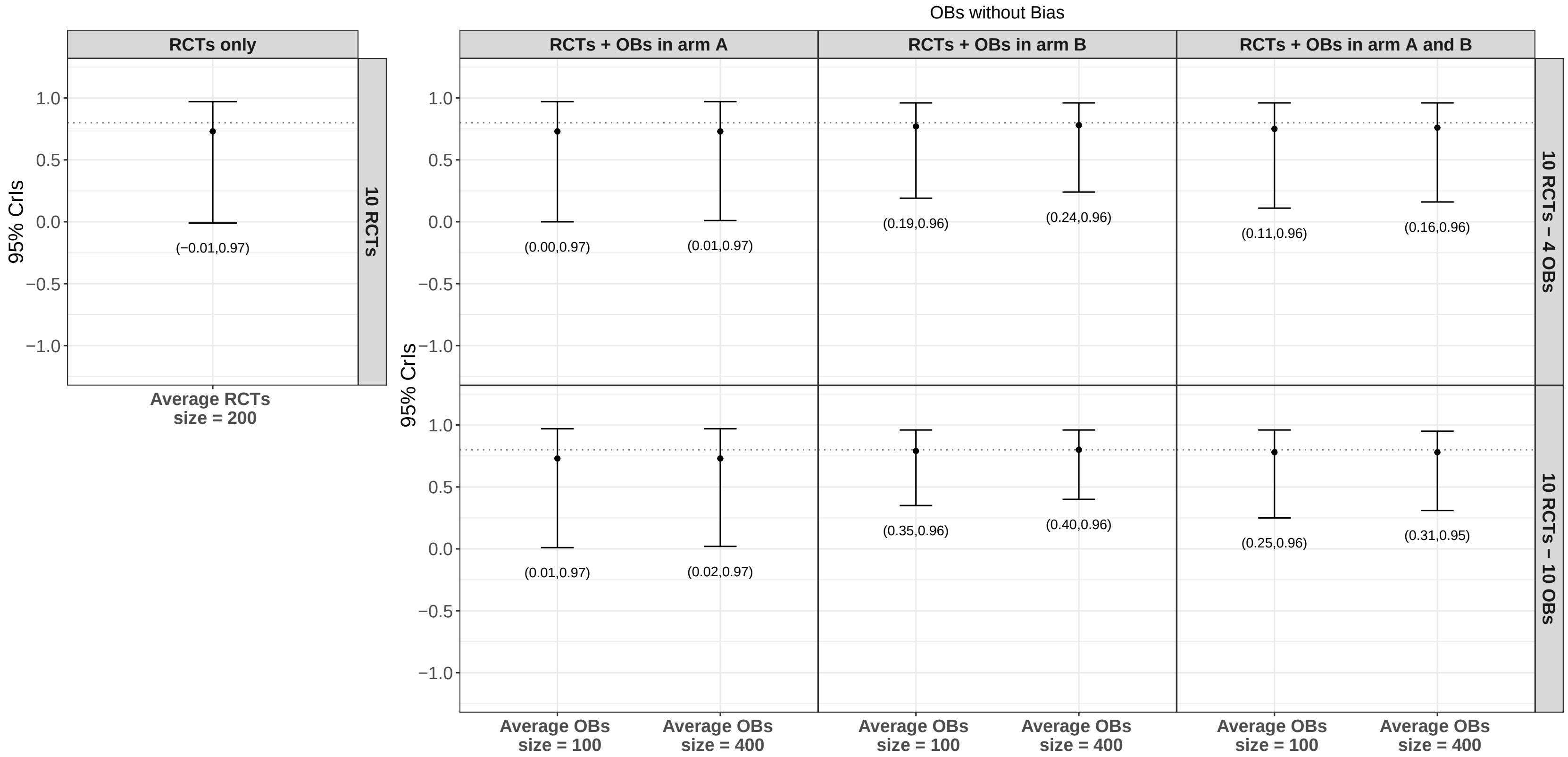


Figure 6.2: Posterior medians (black dot) and 95% CrIs of ρ_b (solid bars) averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) in the scenarios with 10 RCTs and unbiased OBs



Data scenarios with bias in the OBs

The same scenarios were regenerated, but this time introducing bias to the OBs, as described in step 5 of the generation process of OBs. To analyse these scenarios, we used M2 which accounts for such biases. Figures 6.3 and 6.4 present posterior medians and 95% CrIs of ρ_b averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) in the scenarios with 5 or 10 RCTs and biased OBs. Similarly as in Figures 6.1 and 6.2, the plots on the LHS illustrate the scenarios where the meta-analyses were based on RCTs alone and used as reference scenarios.

Moving to the RHS, it can be seen that incorporating OBs reporting data only in arm A, did not improve inferences about the between-studies correlation ρ_b. It resulted in exactly the same precision and accuracy of estimates of the between-studies correlation as in the reference scenarios. A minimal benefit was observed in the scenario where the size of OBs were large and RCT data were sparse (Figure 6.3, first column of the plot on RHS). Incorporating OBs reporting data either in arm B or in both arms, resulted in reduced uncertainty around the estimates of ρ_b. However, the benefit was smaller compared to the scenarios with unbiased OBs (for instance, in the scenario with 4 biased OBs reporting data in arm B and consisting of on average 100 patients, the average 95% CrI of ρ_b was (-0.37,0.97), while in the same scenario with unbiased OBs the average 95% CrI of ρ_b was (-0.27,0.97)). Furthermore, including e.g. 4 OBs reporting data only in arm B in the analysis, resulted in higher precision and slightly better accuracy of the estimates of ρ_b compared to the scenarios where two OBs reported data in arm A and two in arm B. The number of OBs played an important role on the performance of the estimates of ρ_b; the more OBs were incorporated in analysis the more precise were the average 95% CrIs of ρ_b and the average posterior medians of ρ_b were closer to the true value. The number of patients in OBs affects the estimates of ρ_b to a lesser extent compared to the number of OBs included in the analysis, resulting in very similar 95% CrIs of ρ_b. Typically, fitting M2 to data with large OBs (in terms of the number of their patients), yielded slightly narrower 95% CrIs of ρ_b compared to the scenarios where the size of OBs was small.

Figure 6.3: Posterior medians (black dot) and 95% CrIs of ρ_b (solid bars) averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) in the scenarios with 5 RCTs and biased OBs

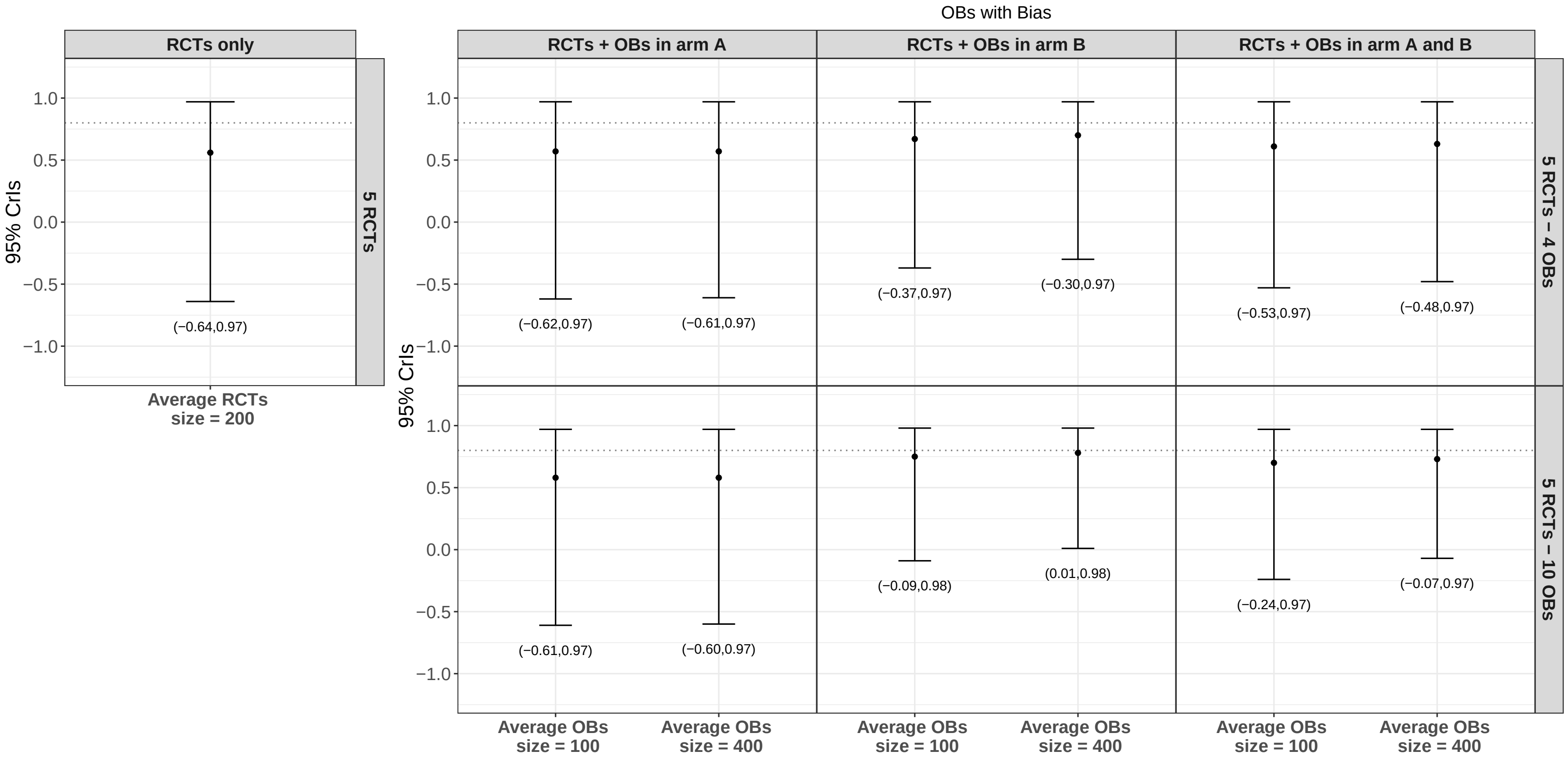
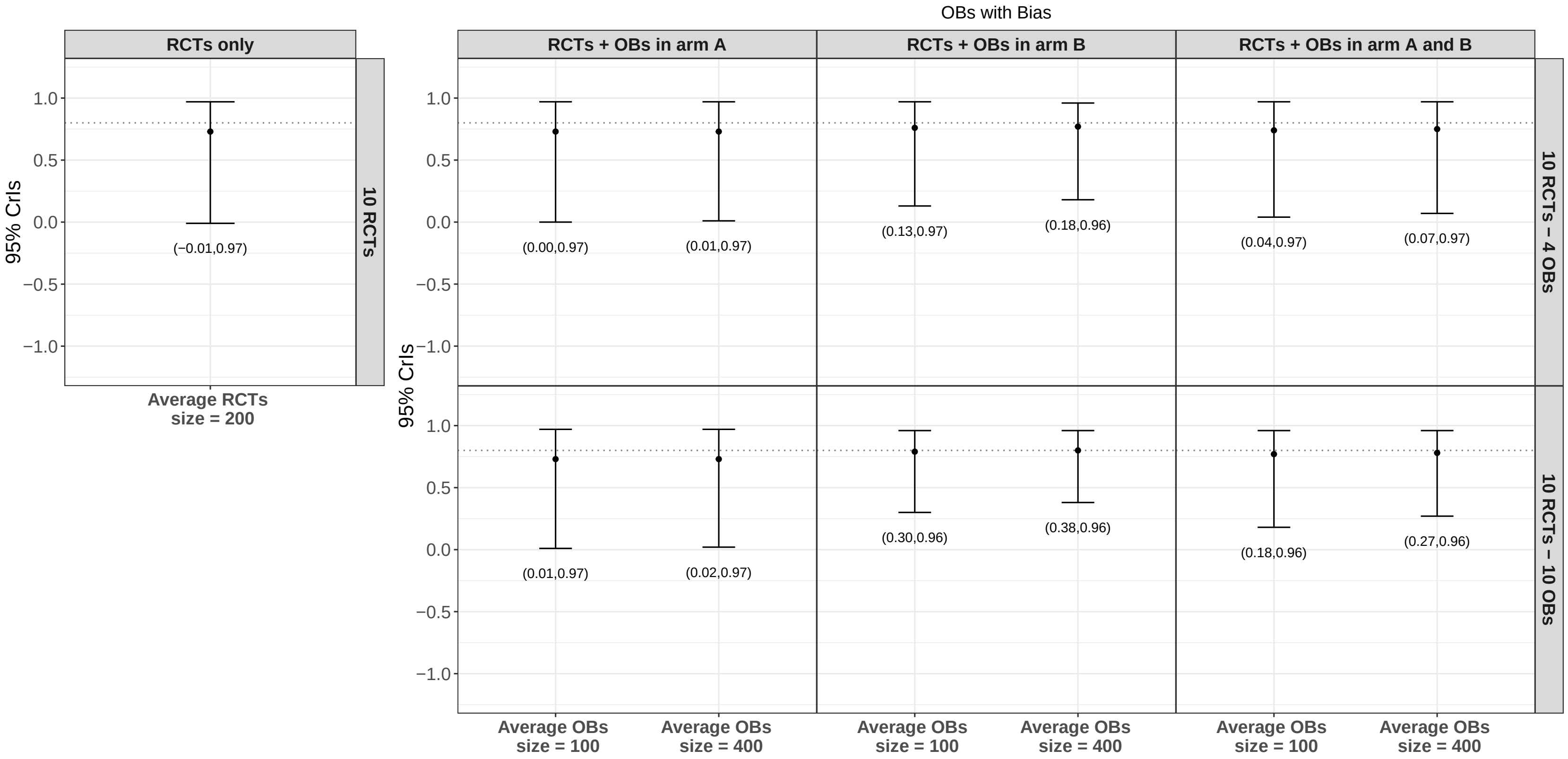


Figure 6.4: Posterior medians (black dot) and 95% CrIs of ρ_b (solid bars) averaged over the 1000 replications along with the true value of $\rho_b = 0.8$ (dotted line) in the scenarios with 10 RCTs and biased OBs



6.4.3.2 Pooled effect d_2

This section presents the results from the estimates of pooled effects. we focused on presenting the results of the pooled effect on the final outcome d_2. The pooled effect on the surrogate endpoint d_1 gave very similar results. To measure the performance of the estimates of d_2, the absolute bias was calculated in each replication.

Data scenarios without bias in the OBs

Figures 6.5 and 6.6 show the average absolute bias of $\widehat{d_2}$ across the 1000 replications in the reference scenarios and the scenarios with unbiased OBs. Here, as the OBs were unbiased, we fitted model M1.

On the LHS, the two plots in Figures 6.5 and 6.6 present the average absolute biases of $\widehat{d_2}$ based on RCT data alone (5 RCTs in Figure 6.5 and 10 RCTs in Figure 6.6 and are used as the reference scenarios in this section. The average absolute bias of $\widehat{d_2}$ was 0.21 when 5 RCTs where available and 0.14 when 10 RCT were included in the analysis.

The scenarios described in the plot on the RHS of Figure 6.5, consisted of 5 RCTs and 4 unbiased OBs (1st row) or 5 RCTs and 10 unbiased OBs (second row). It can be seen that when OBs reporting data only in arm A were included in the analysis, the average absolute bias was marginally improved (0.2 in both scenarios) compared to the reference scenarios. On the other hand, incorporating OBs reporting data either only in arm B (2nd column) or in both arms evenly (3rd column), resulted in lower absolute biases of $\widehat{d_2}$ compare to the reference scenarios. Additionally, including OBs reporting data only in arm B had the greatest impact on the estimates leading to estimates with the lowest average absolute biases across scenarios. The number of OBs included in the analysis substantially affected the values of the average absolute biases across scenarios. In practice, the more OBs were included in the analysis the lower the average absolute bias was. Similarly to the performance of 95% CrIs of ρ_b, the number of patients in OBs was least important of the factors, marginally influencing the performance of $\widehat{d_2}$. For example in the scenario with 10 RCTs and 10 OBs with arm B (2nd row, 2nd column), when the mean number of patients in OBs was 100, the average absolute bias was 0.14, whilst when the mean number of patients was 400, the average absolute bias was 0.13.

The scenarios described in the plot on the RHS of Figure 6.6, consisted of 10 RCTs and unbiased 4 OBs (1st row) or 10 RCTs and unbiased 10 OBs (second row). It can be seen that when sufficient number of RCTs were included in the analysis (10 RCTs), it resulted in sufficiently low average absolute biases (i.e. the average absolute bias was 0.14 in the reference scenario). Incorporating 4 OBs in the analysis did not have substantial impact across on the performance of $\widehat{d_2}$ regardless of the type of the arm (1st row) or the number of patients in OBs. Increasing the number of OBs to 10 resulted in lower average absolutes biases when OBs reported either arm B or both arms evenly. The lowest average bias was measured when the mean number of patients in OBs was 400 and 10 OBs reported only arm B (2nd row, 2nd column).

Data scenarios with bias in the OBs

As presented in the first part of this section, the same analysis was applied to the scenarios with biased OBs. Figures 6.7 and 6.8 present the average absolute bias of $\widehat{d_2}$ across the 1000 replications in the reference scenarios (LHS) and the scenarios where RCT data and biased OBs were combined in a single analysis (RHS). To obtain results across these scenarios, model M2 was used, which accounts for bias in OBs via bias terms. It is clear that when bias existed in OBs and it was also formally incorporated in the model via bias terms η_1, η_2, ξ_1, ξ_2 the estimation of the pooled effects d_1, d_2 is based mainly on evidence provided from the RCTs ignoring the OBs. The average absolute bias remained the same as in the reference scenarios (0.21 and 0.14 when 5 and 10 RCTs were included in the analysis respectively) across all the scenarios regardless of the number of OBs, the number of patients in OBs and the arm that OBs reported.

Figure 6.5: Absolute bias of $\widehat{d_2}$ averaged over 1000 replications in the scenarios with 5 RCTs and unbiased OBs

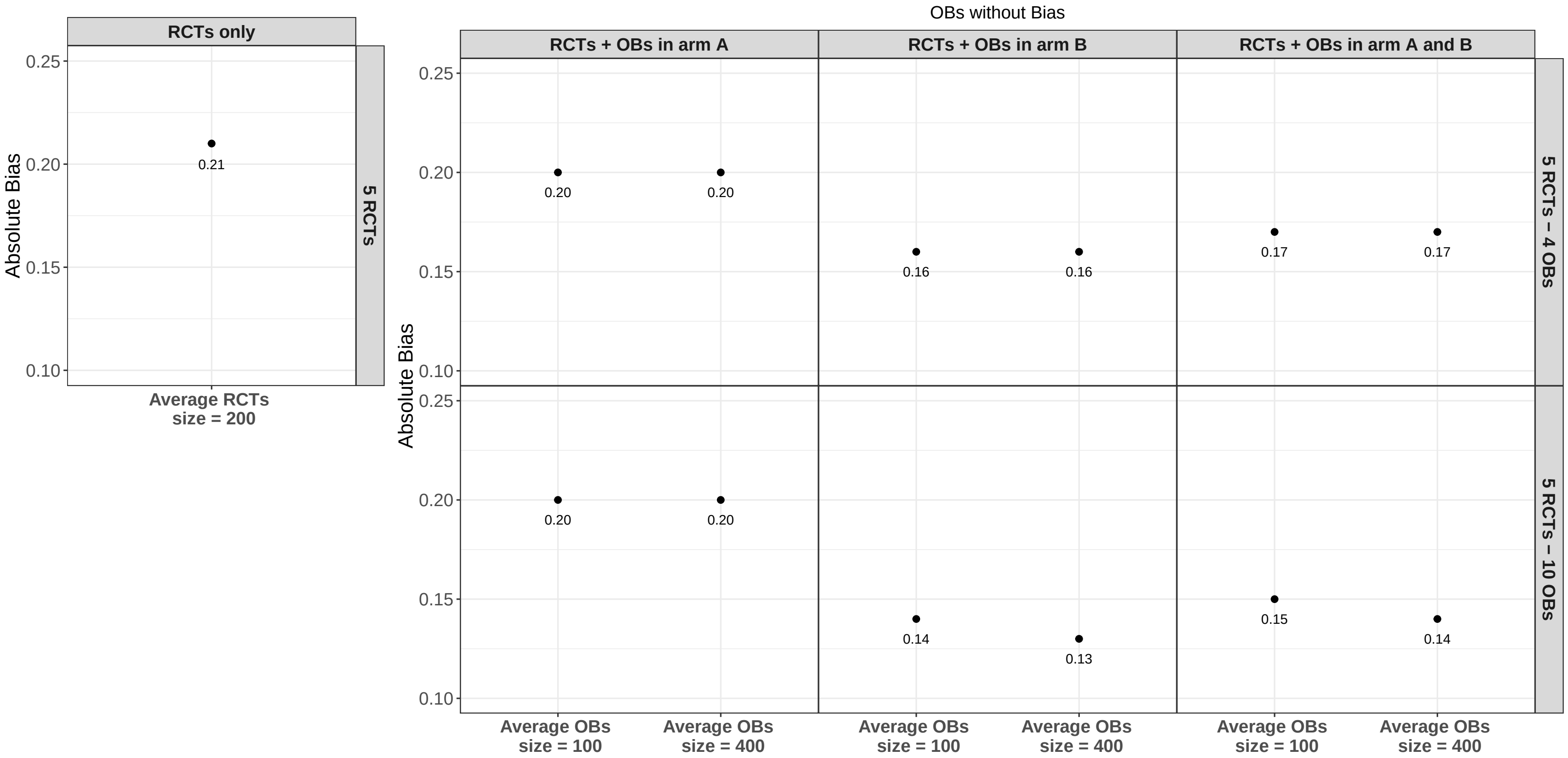


Figure 6.6: Absolute bias of $\widehat{d_2}$ averaged over 1000 replications in the scenarios with 10 RCTs and unbiased OBs

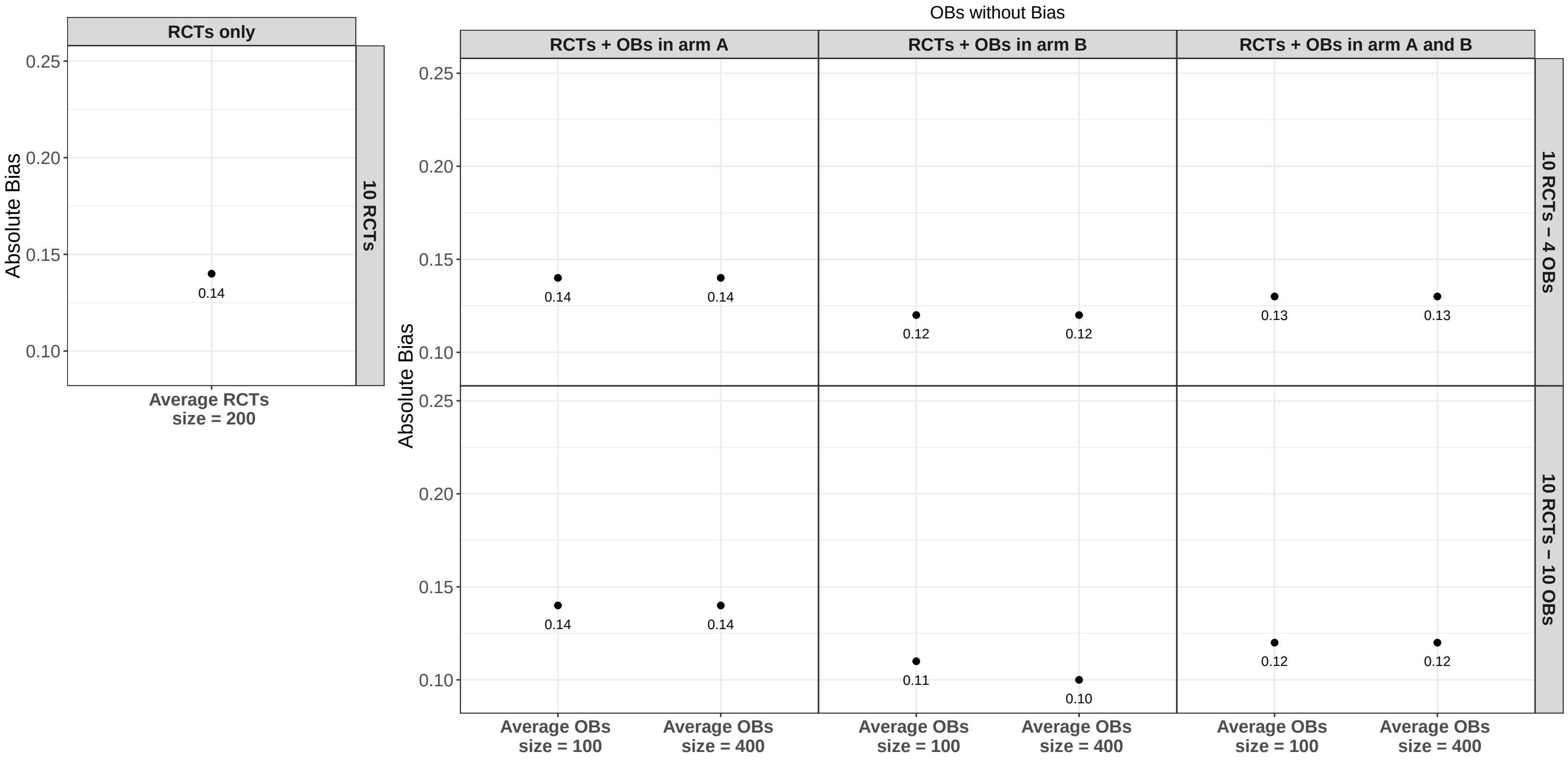


Figure 6.7: Average absolute bias of $\widehat{d_2}$ across replications in the scenarios with 5 RCTs and biased OBs

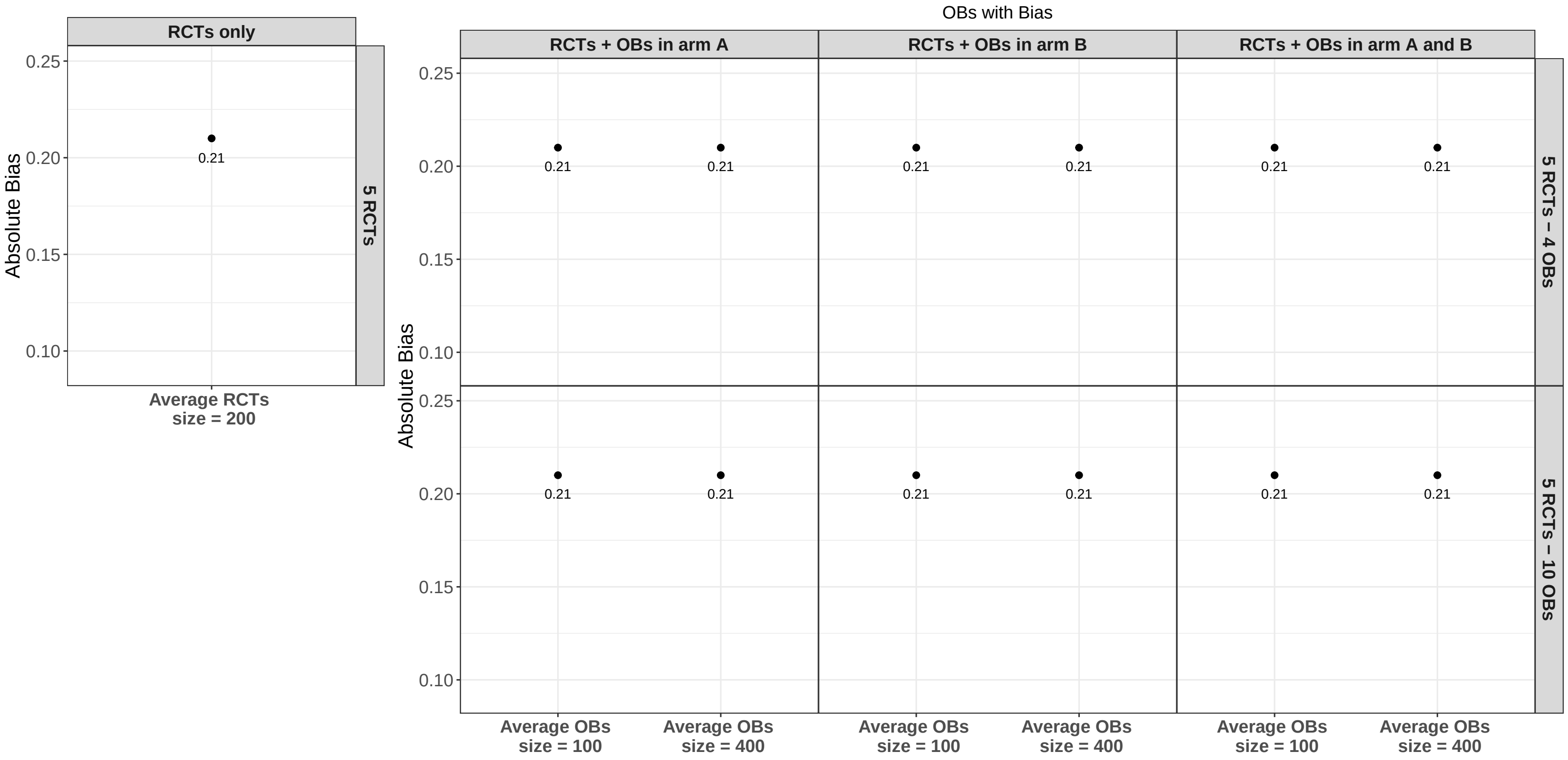
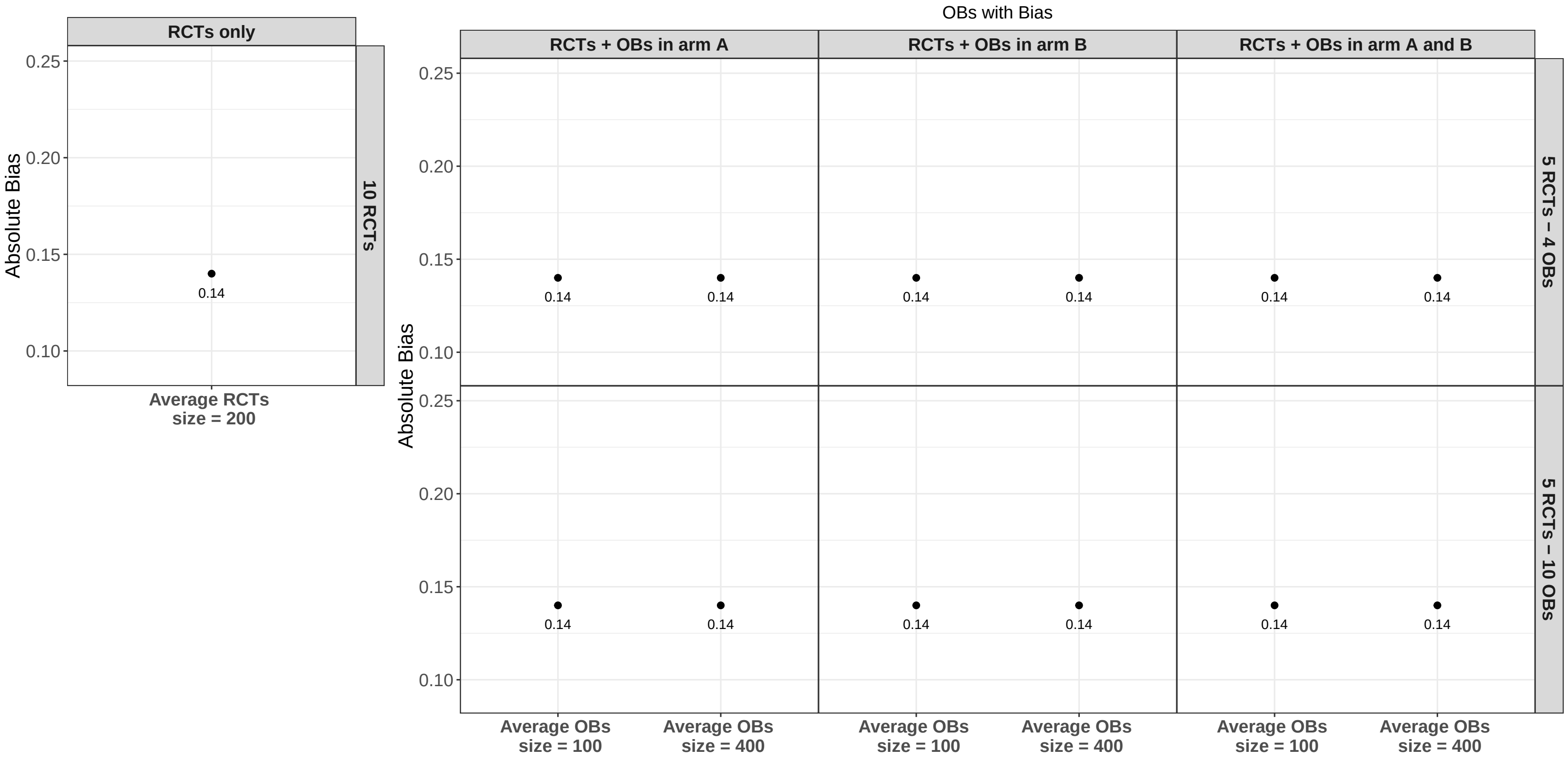


Figure 6.8: Average absolute bias of $\widehat{d}_2$ across replications in the scenarios with 10 RCTs and biased OBs



6.4.4 Key findings

This section presents a short summary of the key findings from the simulation study:

- Overall, the simulation study showed that including OBs in meta-analysis, improved the precision and the accuracy of the estimates of between study correlation. This indicates that when evidence obtained from RCTs are combined with evidence obtained from OBs in a single meta-analysis, it leads to improved inferences about the trial-level surrogacy patterns.
- The simulation study also investigated the effect of the number of OBs included in the analysis, simulating two set of scenarios (one with 4 OBs and a second one with 10 OBs). Overall, the more OBs were included in the analysis, the better was the precision and the accuracy of the estimates of the between-studies correlation. In contrast to this, the number of patients in OBs had the smallest impact on the estimates of the model. Two sets of scenarios were generated one with fewer patients in the studies ($n_i \sim N(100, 5)$) and one with a larger number ($n_i \sim N(400, 5)$). Typically, combining RCTs and OBs with large sample sizes, resulted in slightly more precise and accurate estimates of ρ_b compared to the reference scenarios and the scenarios where OBs were generated with small sample sizes.
- Another aim of the simulation study, was to investigate the performance of the pooled effects d_1 and d_2 when bias was present in the OBs. In a set of scenarios bias was introduced to the data (using step 5 if the generation process of OBs) and accounted for in the model (M2 was used) via the bias terms (η_1, η_2, ξ_1, ξ_2). In these scenarios, the performance of the model as exactly the same as in the reference scenarios, resulting in on average the same absolute biases of the pooled effect across all the scenarios. This means that the estimation of the posterior mean/median of the pooled effects d_1 and d_2 was based on the RCTs across all these scenarios regardless of the inclusion of biased OBs in the analysis. Therefore, accounting for systematic biases in the model via the bias terms (η_1, η_2, ξ_1, ξ_2), prevents the estimates of d_1, d_2 from being susceptible to bias. This is rather important in the context of surrogate endpoints, as biased estimates of the pooled effects (d_1, d_2) would also imply biased estimates of the

true treatment effects. This means that the predictions of the true treatment effect on the final outcome in a given study would also be biased.

In summary, it is clear that a preliminary analysis using M2 is always necessary to evaluate whether or not bias is present in the data, and whether it is small enough to be ignored.

- The simulation study highlighted the importance of which treatment arm the data in OBs represented. The findings confirm that when the RCT data were combined with OBs reporting data only in arm A, there was only marginal improvement in the estimates of between-studies correlation ρ_b in terms of precision and accuracy, compared to the reference scenarios where only RCTs where included in the analysis. Similar performance was also observed for the estimates of d_2. On the other hand, combining RCTs and OBs reporting data in arm B, led to the most precise and accurate estimates of ρ_b and on average the least biased estimates of d_2. Improvement (in terms of precision and accuracy) was also observed when RCTs were combined with OBs reporting data in both arms evenly, however, the impact was smaller compared to the scenarios with OBs reporting data only in arm B.

  This behaviour is due to lack of symmetry at the within-study level of the hierarchical model (described by eq. 6.11, eq. 6.12). As the observed treatment effects obtained from OBs for arm A, directly contribute only to the estimation of the baseline effects, whilst the observed treatment effects obtained from OBs for arm B, directly contribute both to the baseline effects and the true treatment effects. This lack of symmetry was reflected to the performance of the estimates across scenarios.

6.5 Data examples

We illustrate the proposed methodology with two data examples in disease areas where OBs were widely available. The first one focuses on the class of the anti-angiogenic treatments in aCRC and the second one in CML. We presented point estimates (posterior means and medians) and 95% CrIs of the between studies correlation ρ_b and the intercept λ_0 to evaluate whether combing evidence obtained from RCTs and OBs improve the trial-level validation of surrogate endpoints. For completeness, we also presented the results from the between-studies parameters τ_1, τ_2, d_1, d_2. To investigate the susceptibility of the estimates to different assumptions at the within-study level, we modelled the first data example using M2 (section 6.3.2, eq. 6.15 and eq. 6.16) and the generalisation of M3 (section 6.3.3, eq. 6.17-6.23). Additionally, two sets of sensitivity analysis were carried out to asses the impact of prior distributions of ρ_{wA} and ρ_{wB} on the results.

6.5.1 Anti-angiogenic treatments in aCRC

The surrogacy patterns in this particular treatment class of aCRC were also investigated in Chapter 4. Here, we aim to illustrate the benefits of including evidence obtained from OBs in a disease area with plethora of published OBs. The data-set consist of 12 RCTs and 16 OBs evaluating interventions from anti-angiogenic treatment class in aCRC. The RCT data compare anti-angiogenic treatments such as Bevacizumab, Valatinib and Cediranib combined with various types of chemotherapy against cytotoxic agents such as, FOLFOX (folinicacid, fluorouracil, oxaliplatin), FOLFIRI (folinic acid, fluorouracil, irinotecan) or XELOX (capecetabine and oxaliplatin). Ten RCTs were obtained from the literature review conducted by Ciani et al. [11] (for details see chapter 4) and additional two RCTs were added after a short review of PubMed database. As discussed in section 6.4.4, the proposed methodology is expected to improve the validation of trial-level surrogate relationships by combining RCT data and OBs, when OBs data are available mainly in arm B. Therefore in this example, we focused on identifying OBs which investigate the experimental arm. We extracted 16 OBs evaluating anti-angiogenic treatments.

Typically in comparative studies PFS and OS are reported and analysed on the hazard ratio scale, however, this is not possible for single-arm OBs. Therefore, to combine RCTs and OBs, we extracted data on specific time points using the binomial scale. The extracted binary outcomes were PFS at one year (candidate surrogate endpoint) and OS at two years (final outcome). The data on these two outcomes were either reported directly by the studies or were obtained from their Kaplan-Meier curves. Upper part of Table 6.1 shows the summarised data from RCTs presenting the number of events in each arm and on both outcomes (PFS at one year, OS at two years) and the lower part of Table 6.1 presents the data obtained from the 16 OBs on both outcomes.

Table 6.1: Summarised data

	PFS at 1 year				OS at 2 years			
	Arm A		Arm B		Arm A		Arm B	
Study name	N_{1Ai}	r_{1Ai}	N_{1Bi}	r_{1Bi}	N_{2Ai}	r_{2Ai}	N_{2Bi}	r_{2Bi}
RCTs								
Diaz 2012 [192]	238	74	238	93	238	93	238	114
Guan 2011 [193]	72	8	142	33	72	22	142	50
Hecht 2011 [194]	410	185	413	145	410	230	413	116
Hecht 2009 [195]	583	99	585	99	583	251	585	234
Hoff 2012 [196]	358	61	502	30	358	143	502	216
Hurwitz 2004 [124]	411	58	402	28	411	127	402	171
Kabinnavar 2005[197]	105	22	104	25	105	25	104	29
Schmoll 2012 [198]	713	257	709	227	713	309	709	330
Souglakos 2012 [199]	166	60	167	60	166	102	167	107
Tebbutt 2010 [200]	156	23	157	38	156	53	157	55
Van Cutsem 2012 [201]	614	86	612	101	614	108	612	173
Van Cutsem 2011 [202]	429	19	426	34	429	88	426	77
OBs								
Bendell 2012 (1) [203]			968	397			968	481
Bendell 2012 (2) [203]			243	100			243	188
Hurwitz 2014 [204]			482	116			482	188
Van Cutsem 2009 (1) [205]			300	93			300	104
Van Cutsem 2009 (2) [205]			503	236			503	247
Van Cutsem 2009 (3) [205]			552	248			552	293
Van Cutsem 2009 (4) [205]			346	145			346	159
Bennouna 2017 (1) [206]			521	224			521	271
Bennouna 2017 (2) [206]			154	42			154	61
Buchler 2014 (1) [207]			1218	585			1218	658
Buchler 2014 (2) [207]			973	467			973	589
Ocvirk 2011 (1) [208]			45	23			45	26
Ocvirk 2011 (2) [208]			94	43			94	57
Moriwaki 2012 (1) [209]			115	41			115	48
Moriwaki 2012 (2) [209]			45	12			45	15
Kotaka 2016 [210]			40	12			40	26

6.5.1.1 Data Synthesis

We conducted two meta-analyses to assess whether the inclusion of evidence obtained from OBs improved the validation of PFS at 1 year as a surrogate endpoint of OS at 2 years. The first meta-analysis consisted only of RCT data, whilst the second analysis incorporated also the 16 OBs. Within-study correlations between the observed treatment effects on the surrogate endpoint and the final outcome were available for the anti-angiogenic treatment class on the log hazard ratio scale in chapter 4 (see 4.4.3). Here, we assumed that the within-study association was approximately the same in each arm as in 4.4.3 despite modeling the treatment effects on a different scale (log of odds). To allow for some uncertainty, we placed informative prior distributions on, $\rho_{wA} \sim U(0.45, 60)$, $\rho_{wB} \sim U(0.45, 60)$ which were the same across all studies (RCTs and OBs).

6.5.1.2 Data analysis

Analyses of aCRC data using model M2

This section presents the results of both meta-analyses giving estimates of all the between-studies parameters. To investigate the presence of bias in the OBs, we fitted M2 which accounts for bias in the OBs via bias terms and models the within-study variability on the log odds scale using a normal approximation. Table 6.2 contains the results from both meta-analyses (of RCT data alone and combination and RCT and OB data) presenting the estimates of the between-studies parameters and the bias terms.

Table 6.2: Between-studies estimates across data-sets using M2

Models	RCTs alone		RCTs combined with OBs	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters				
ρ_b	0.50(0.55)	(-0.15, 0.87)	0.59(0.63)	(0.18, 0.85)
λ_0	-0.12(-0.13)	(-0.44, 0.22)	-0.13(-0.13)	(-0.36, 0.10)
τ_1	0.54(0.51)	(0.31, 0.91)	0.40(0.39)	(0.26, 0.60)
τ_2	0.51(0.49)	(0.30, 0.86)	0.39(0.38)	(0.26, 0.57)
d_1	0.33(0.32)	(0.01, 0.67)	0.31(0.31)	(0.04, 0.57)
d_2	0.04(0.05)	(-0.29, 0.35)	0.05(0.05)	(-0.20, 0.31)
η_1			0.60(0.60)	(0.03, 1.14)
η_2			0.48(0.47)	(-0.07, 0.93)

When only RCT data were included in the analysis, the method estimated the between-studies association with relatively large uncertainty. Although, the 95% CrI of the intercept contained zero and obtained with relatively small uncertainty, the median of ρ_b was 0.55 and the 95% CrI spanned from -0.15 to 0.87. This suggest that PFS at 1 year can be considered as a relatively weak surrogate endpoint of OS at 2 years.

On the other hand, when the RCT data were combined with OBs, there was a substantial improvement in the precision of the estimates of ρ_b and λ_0. In this case, the 95% CrI of ρ_b was considerably narrower compared to the analysis with only RCT data, spanning from 0.18 to 0.85. Similar effect was observed for the 95% CrI of the intercept. These findings allow us to draw more precise inferences about the surrogate relationship between PFS at 1 year and OS at 2 years indicating also a stronger surrogacy pattern between the treatment effects on this pair of outcomes.

Similar gain in precision were also observed across all the between-studies estimates. Furthermore, the method identified substantial bias in the OBs both on the surrogate (η_1) and the final outcome (η_2) making the choice of including the bias terms in the model justifiable. The point estimates of the pooled effects on the surrogate endpoint and the final outcome were very similar compared to the analysis with only RCT data. This is due to the inclusion of bias terms in the model as explained in Section 6.4.4.

Evaluating the assumption of normality at the within-study level

To evaluate the impact of the assumption of normality at the within-study level on the results,the same analysis was repeated fitting the M3 (see details in Section 6.3.3), which models the within-study variability on the exact binomial scale and does account for biases in OBs via bias terms. This version of the method avoids the assumption of normality when modeling the within-study variability by using instead joint densities with binomial marginal distributions constructed with copulas. To maintain the within-study correlations on the same scale as in the previous analysis, we used the Gaussian copula (see details in section 5.2.3). Table 6.3 illustrates the results from both meta-analyses (of RCT data alone and combination and RCT and OB data) presenting the estimates of the between-studies parameters and the bias terms.

Modeling the within-study variability on the original binomial scale gave relatively similar results about the between-studies correlation in both meta-analyses. Specifically, M3 resulted in slightly higher posterior medians of between-studies

Table 6.3: Between-studies estimates across data-sets using M3

Models	RCTs alone		RCTs combined with OBs	
Measures	Mean(Median)	95% CrI	Median(Mean)	95% CrI
Parameters				
ρ_b	0.55(0.60)	(-0.08, 0.89)	0.64(0.68)	(0.23, 0.90)
λ_0	-0.12(-0.11)	(-0.44, 0.20)	-0.11(-0.11)	(-0.35, 0.08)
τ_1	0.54(0.51)	(0.31, 0.92)	0.40(0.39)	(0.26, 0.59)
τ_2	0.52(0.49)	(0.31, 0.89)	0.39(0.37)	(0.26, 0.55)
d_1	0.33(0.33)	(0.01, 0.68)	0.32(0.32)	(0.05, 0.58)
d_2	0.07(0.06)	(-0.23, 0.38)	0.06(0.06)	(-0.19, 0.33)
η_1			0.61(0.61)	(0.04, 1.15)
η_2			0.47(0.48)	(-0.05, 0.95)

correlation ρ_b and narrower 95% CrI of ρ_b compared to M2 in both analyses (one with RCT data only and one with RCTs and OBs). Overall, modeling with M3, led to slightly more precise inferences about the trial-level surrogate relationship between PFS at 1 year and OS at 2 years compared to M2.

Minor differences were observed between the results obtained from models M2 and M3 for the remaining parameters. The point-estimates (posterior means/medians) and their corresponding 95% CrIs were very similar across the two versions of the method.

Sensitivity to the choice of the prior distributions for ρ_{wA} and ρ_{wB}

To evaluate how susceptible were the results of the between-studies parameters to choice of prior distributions for the within-study correlations ρ_{wA} and ρ_{wB}, we performed a sensitivity analysis placing non-informative prior distributions ($\rho_{wA}, \rho_{wB} \sim U(-1, 1)$) and weakly informative prior distributions ($\rho_{wA}, \rho_{wB} \sim U(0, 1)$) on these parameters respectively. Two sets of sensitivity analysis were conducted, one for each meta-analysis. In the first set of sensitivity analysis, we included only RCT data, whereas in second one, both RCTs and OBs were incorporated into the analysis. The results from both sets of the sensitivity analysis were also compared with the results from the analyses using the informative prior distributions for ρ_{wA} amd ρ_{wB} ($\rho_{wA}, \rho_{wB} \sim U(0.45, 0.60)$) as presented previously in this section. Throughout this sensitivity analysis we used M2. Table 6.4a displays the results from the first set of sensitivity analysis of between-study parameters, where RCT data alone were included in the analysis. Table 6.4b contains the results from the second set of sensitivity analysis, where 12 RCTs and 16 OBs were included in the analysis.

The results from the first set of sensitivity analysis (Table 6.4a) show that placing non-informative prior distributions on the within-study correlations resulted in the slightly higher point-estimates of between-studies correlation ρ_b and slightly narrower 95% CrI. Similarly, marginal differences were observed in the estimates and the 95% CrIs of between-studies standard deviations and the pooled effects on the surrogate endpoint and the final outcome across the different prior distributions. Overall, the estimates of between-studies parameters were not susceptible to the choice of prior distributions for ρ_{wA} and ρ_{wB}. Particularly, being less informative about the within-study correlations did not substantially affect the inferences of ρ_b and λ_0 and, consequently, the trial-level validation of PFS at 1 year as surrogate endpoint of OS at 2 years, when only RCT data were included in the analysis.

The results from the second set of sensitivity analysis (Table 6.4b) suggest that the impact of different prior distributions of ρ_{wA} on the estimates of between-studies parameters was relatively small when OBs were incorporated into the analysis. Specifically, being completely ignorant about ρ_{wA} and ρ_{wB} resulted in the highest point-estimates of between-studies correlation ρ_b compared to the scenarios with weakly informative and informative prior distributions. However, the choice of prior distribution about ρ_{wA} and ρ_{wB} did not substantially affected the inferences about ρ_b and consequently the trial-level validation of PFS at 1 year as surrogate endpoint of OS at 2 years.

Table 6.4: Estimates of between-studies parameters across different degrees of prior information of ρ_{wA} and ρ_{wB}

(a) RCTs alone

ρ_{wA}, ρ_{wB}	non-informative prior distributions U(-1,1)		weakly informative prior distributions U(0,1)		informative prior distributions U(0.45,0.60)	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters						
ρ_b	0.54(0.59)	(-0.08, 0.89)	0.51(0.56)	(-0.14, 0.87)	0.50(0.55)	(-0.15, 0.87)
λ_0	-0.12(-0.12)	(-0.43, 0.21)	-0.12(-0.12)	(-0.43, 0.21)	-0.12(-0.13)	(-0.44, 0.22)
τ_1	0.53(0.50)	(0.31, 0.89)	0.54(0.51)	(0.31, 0.91)	0.54(0.51)	(0.31, 0.90)
τ_2	0.51(0.49)	(0.31, 0.85)	0.51(0.49)	(0.30, 0.85)	0.51(0.49)	(0.30, 0.86)
d_1	0.33(0.33)	(-0.01, 0.69)	0.34(0.33)	(0.00, 0.68)	0.33(0.32)	(0.01, 0.67)
d_2	0.07(0.07)	(-0.26, 0.40)	0.06(0.06)	(-0.26, 0.38)	0.04(0.05)	(-0.29, 0.35)

(b) RCTs combined with OBs

ρ_{wA}, ρ_{wB}	non-informative prior distributions U(-1,1)		weakly informative prior distributions U(0,1)		informative prior distributions U(0.45,0.60)	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters						
ρ_b	0.69(0.72)	(0.23, 0.94)	0.59(0.65)	(0.18, 0.88)	0.59(0.63)	(0.18, 0.85)
λ_0	-0.13(-0.13)	(-0.36, 0.10)	-0.12(-0.12)	(-0.36, 0.09)	-0.11(-0.11)	(-0.35, 0.08)
τ_1	0.40(0.39)	(0.26, 0.59)	0.40(0.39)	(0.26, 0.59)	0.40(0.39)	(0.26, 0.59)
τ_2	0.39(0.38)	(0.26, 0.57)	0.39(0.38)	(0.26, 0.57)	0.39(0.37)	(0.26, 0.55)
d_1	0.31(0.30)	(0.05, 0.58)	0.31(0.30)	(0.06, 0.60)	0.31(0.31)	(0.04, 0.57)
d_2	0.07(0.07)	(-0.18, 0.32)	0.06(0.06)	(-0.20, 0.31)	0.05(0.05)	(-0.20, 0.31)
η_1	0.61(0.61)	(0.04, 1.16)	0.60(0.60)	(0.06, 1.17)	0.61(0.61)	(0.04, 1.15)
η_2	0.46(0.46)	(-0.04, 0.94)	0.48(0.47)	(0.01, 0.97)	0.46(0.48)	(-0.05, 0.95)

6.5.1.3 Cross-validation procedure

PFS at 1 year was deemed a relatively weak surrogate endpoint of OS at 2 years based on RCT data alone, whilst the inclusion of OBs strengthen its validation resulting in more precise inferences about the parameters describing the surrogate relationships. In this section two cross-validation procedures with M2 model were carried out to assess the effect of including OBs in the analysis of the predictions of the true treatment effect on the final outcome. In the first cross-validation, the true treatment effect on the final outcome of each RCT was predicted based on the treatment effects on the surrogate endpoint and the final outcome of the remaining RCTs and the treatment effect on the surrogate endpoint of the particular study, whilst in the second one the true treatment effect on the final outcome of each RCT was predicted based on the treatment effects on the surrogate endpoint and final outcome of the remaining RCTs and the OBs and the treatment effect on the surrogate endpoint of the particular study.

Table 6.5 presents the performance of the predictions of the true treatment effect on the final outcome (OS at 2 years) by reporting the following measures: The mean error, mean absolute error, the performance of the predictive intervals, the ratio of the width of the 95% predictive intervals obtained from the cross-validation in the data-set consisting of RCTs and OBs and the predictive intervals obtained from the cross-validation based on RCTs alone, averaged over the RCTs. The definitions of the measures can also be found in Section 5.4.4.1.

Table 6.5: Performance of predictions in the two data-sets

Measures	RCTs alone	RCTs combined with OBs
Performance of 95% predictive intervals	0.92	0.92
Mean error	0.03	0.01
Mean absolute error	0.31	0.30
Mean width ratio	1.00	0.68

The results of the cross-validation procedure on PFS-OS pair of outcomes showed that the inclusion of OBs in the data-set resulted in better predictions compared to the one with RCTs alone, resulting in lower mean error, mean absolute error and, on average, 32% narrower 95% predictive intervals of the true effects on the final outcome. The performance of the 95% predictive intervals was exactly the same, as the 11 out of the 12 predictive intervals contained the observed treatment effect on the final outcome in both data-sets.

6.5.2 CML

As described in Chapter 5 of this thesis, CML is a slow progressive disease. The introduction of TKI has revolutionized the management of CML and patients' prognosis [164]. In Section 5.4.4, we investigated the trial-level surrogacy patterns between CCyR at 1 year and EFS/OS at 2 years using 3 different modeling options. Overall, a sub-optimal surrogate relationship between CCyR at 1 year and EFS or OS at 2 years was found regardless of the modeling assumptions. This can potentially be due to the immature data on the treatment effects on the final outcome. The lack of maturity of the data on the treatment effects may affect the shape of the relationship as it typically results in treatment effects clustered around the mean effect, and hence very small between-studies heterogeneity. This leads to unsuccessful validation of surrogate endpoints. In this data example we use a longer term-outcome as final outcome (OS at 4 years). Unfortunately only five RCTs of the data example in Chapter 5 reported treatment effects on this final outcome.

To enrich our data-set with OBs, we carried out a literature review screening titles of relevant papers and abstracts. We identified 8 OBs, two of them reported data in arm A (standard dose of imatinib) and six reported data in arm B (second generation TKI, or high dose of imatinib). We also used the experimental arm of IRIS trial [211] (this trial compared imatinib against standard chemotherapy and it was excluded from the RCT data) as an additional OBs. Table 6.6 displays the number of survivors at 4 years (OS at 4 years) and the number of patients who achieved a CCyR at 1 year (potential surrogate endpoint).

Table 6.6: Summarised data

	CCyR at 1 year				OS at 4 years			
	Arm A		Arm B		Arm A		Arm B	
Study name	N_{1Ai}	r_{1Ai}	N_{1Bi}	r_{1Bi}	N_{2Ai}	r_{2Ai}	N_{2Bi}	r_{2Bi}
RCTs								
Cortes 2013 [212]	260	189	259	216	260	216	259	246
Hochhaus 2016 [213]	278	181	278	217	283	261	282	272
Baccarani 2009 [160]	108	63	108	69	108	91	108	98
Hehlmann 2011 [161]	303	150	311	206	324	295	338	314
Deininger 2014 [156]	49	33	41	35	72	65	73	69
OBs								
Lavadalle 2008 [6]	198	113			204	190		
Hochhaus 2017 [211]	553	421			553	503		
Kizaki 2019 (1) [214]	139	97			139	128		
Kizaki 2019 (2) [214]			169	136			169	166
Kizaki 2019 (3) [214]			144	119			144	137
Hoffmann 2017 (1) [215]			192	96			294	273
Hoffmann 2017 (2) [215]			52	39			78	72
Jabbour 2011(1) [164]			187	169			208	199
Jabbour 2011(2) [164]			118	116			154	154

6.5.2.1 Data Synthesis

Two meta-analyses were conducted, one using the RCTs alone and one including the OBs in the analysis. We applied the version of the method which models the within-study variability on the original binomial scale. This method was preferred due to the high number of events on the final outcome (OS at 4 years). To construct the joint density with binomial marginal distributions we used the Gaussian copula as dependence structure, placing the same prior distributions on the within-study associations as in Chapter 5 ($\theta_1 \sim U(-0.03, 0.22)$ and $\theta_2 \sim U(0, 0.25)$)).

6.5.2.2 Results

Table 6.7 displays the results from both meta-analyses presenting the estimates of between-studies parameters and the bias terms. Including only the RCTs in the meta-analysis resulted in relatively low between-studies association (the median $\hat{\rho}_b = 0.41$) and substantial uncertainty around the estimate. Specifically, the 95% CrI of the parameter was (-0.87,0.96) implying that CCyR at 1 year is not a valid a surrogate endpoint of OS at 4 years based on the RCT data. In the second

Table 6.7: Estimates of between-studies parameters across the two meta-analyses

Models	RCTs alone		RCTs combined with OBs	
Measures	Mean(Median)	95% CrI	Mean(Median)	95% CrI
Parameters				
ρ_b	0.25(0.41)	(-0.87, 0.96)	0.61(0.82)	(-0.53, 0.98)
λ_0	0.04(0.21)	(-1.87, 1.76)	0.21(0.38)	(-1.61, 1.40)
τ_1	0.30(0.20)	(0.02, 0.84)	0.64(0.59)	(0.05, 1.49)
τ_2	0.58(0.46)	(0.08, 1.45)	0.72(0.65)	(0.14, 1.51)
d_1	0.61(0.60)	(0.32, 0.94)	0.62(0.60)	(0.04, 1.25)
d_2	0.78(0.77)	(0.21, 1.35)	0.78(0.78)	(0.14, 1.44)
η_1			0.24(0.24)	(-0.67, 1.13)
η_2			0.39(0.40)	(-0.23, 1.03)
ξ_1			0.54(0.54)	(-0.57, 1.67)
ξ_2			0.58(0.55)	(-0.41, 1.64)

meta-analysis the RCTs were combined with the OBs. This had a direct impact on the estimate of between-studies correlation ρ_b, increasing its point estimate from 0.41 to 0.82. The precision of the 95% CrI was also increased, however, it still contained negative values. Additionally, the 95% CrI of the intercept contained zero and was narrower compared to the 95% CrI of the intercept obtained from the meta-analysis with RCT data alone. However, it still yielded with substantial uncertainty. All the above indicate that the inclusion of OBs in the meta-analysis allowed us to draw more precise inferences about the trial-level surrogacy pattern between CCyR at 1 year and OS at 4 years.

6.5.2.3 Cross-validation procedure

Similarly as in 5.4.4, the validation of CCyR at 1 year as a surrogate endpoint of OS at 4 years was unsuccessful, as the trial-level association was weak and obtained with considerable uncertainty however the use of a longer term final outcome (OS at 4 years instead of OS at 2 years) and the inclusion of OBs in the analysis resulted in higher trial-level association between the treatment effects on the first and the second outcome. Despite the unsuccessful validation of the candidate endpoint, two cross-validation procedures were carried out, one for each data-set, to assess the effect of including OBs in the analysis on the predictions of the true treatment effect on the final outcome. Specifically, in the first cross-validation, the true treatment effect on the final outcome of each RCT was predicted based on the treatment effect on the surrogate endpoint of the particular study and the true treatment effects on the surrogate endpoint and the final outcome of the remaining RCTs, whilst in the second one the true treatment effect on the final outcome of each RCT was predicted based on the true treatment effects on the surrogate endpoint and final outcome of the remaining RCTs and the OBs and the treatment effect on the surrogate endpoint of the particular study.

Table 6.5 presents the performance of the predictions of the true treatment effect on the final outcome (OS at 4 years) across models by reporting the same measures as Table 6.5.

Table 6.8: Performance of predictions in the two data-sets

Measures	RCTs alone	RCTs combined with OBs
Performance of		
95% predictive intervals	1.00	1.00
Mean error	0.09	0.12
Mean absolute error	0.37	0.41
Mean width ratio	1.00	0.92

The results of the cross-validation procedure on CCyR at 1 year - OS at 4 years pair of outcomes showed that both procedures gave similar predictions. In the data-set consisting of RCTs combined with OBS, the predictions of the true treatment effect on the final outcome had on average, 8% narrower predictive intervals compared to the predicted intervals obtained from the data-set consisting of RCTs alone. In contrast to this, the predictions of the data-set consisting of RCTs alone had slightly lower mean error and mean absolute error. The performance of the 95% predictive intervals was exactly the same in both data-sets as all the predictive intervals (5 out of 5) contained the observed treatment effect on the final outcome.

6.6 Discussion

In this chapter we proposed Bayesian bivariate meta-analytic methods for combining evidence from different data sources such as RCTs and single arm OBs. The proposed methodology offers a flexible framework for combining data from RCTs and OBs in a single bivariate meta-analysis with the aim of improving the trial-level validation of surrogate endpoints by drawing more accurate and precise inferences about the study-level association (ρ_b) between the treatment effects on the surrogate endpoint and the final outcome.

Our method extends and the model proposed by Begg and Pilote for combining RCT data with evidence from single arm studies [188] to a bivariate hierarchical method for combining such data sources for treatment effects on two outcomes. Similarly as Begg and Pilote briefly discussed in their second model generalisation, the proposed method (M1) assumes that the baseline effects and the true treatment effects obtained from RCTs and OBs are exchangeable. We further generalised model M1, in a similar way as Begg and Pilote, to account for bias in the OBs by introducing bias terms. This generalisation (M2) is important as OBs have very different inclusion criteria compared to OBs and they are prone to bias due to lack of randomisation. M2 should always be used as preliminary analysis in order to assess the presence of such biases (systematic differences between treatment effect across the study designs) and whether or not they could be ignored. Additionally, an alternative version of the method (M3) was proposed which models the within-study variability on its original binomial scale avoiding the inappropriate assumption of normality when the proportions of events in the binomial data are either very high or low as discussed in Chapter 5.

According to the findings of the simulation study, the inclusion of OBs in the analysis can result in estimates of the between-studies correlation obtained with improved precision and accuracy compared to the analysis based only on RCTs. This is desirable, as it leads to improved trial-level validation of surrogate endpoints. The improvement in the precision of the estimate of the between-studies correlation is more likely to be observed when there are more OBs included in the meta-analysis and when these OBs are available for the experimental treatment rather than the control. On the other hand, the number of patients in OBs has a very small impact on the precision and the accuracy of the estimate of the between-studies correlation. Typically, larger OBs can result in slightly more accurate and precise estimates of between-studies correlation. Furthermore, the inclusion of OBs in the analysis has larger impact on the precision and the accuracy of the estimates of between-studies correlation (when RCT data are sparse).

In data scenarios with noticeable systematic differences in the magnitude of the treatment effects between the RCT and OB data, the generalisation of the model which accounts for such systematic biases via bias terms, estimated the bias very accurately. Such, potential biases in the data should always be assessed. Therefore we recommend that model M2 should always be the first modeling option when meta-analysing data from different study designs. The method is capable of identifying potential biases in the OBs and without affecting the estimates of the pooled treatment effects, e.g in the scenarios of the simulation study with biased OBs, the model estimated the bias very accurately. Furthermore, the model did not introduce any additional bias to the pooled effect when comparing with the results of the RCT data alone. This is key when obtaining predictions of the true treatment effect of a study on the final outcome in a cross-validation procedure, as biased estimates of the pooled effects would imply biased estimates of the true treatment effects.

Overall, the proposed framework can be particularly useful when a new treatment needs to be approved quickly and such decision is based on the treatment effects measured on a surrogate endpoint obtained from only a few RCTs.

The results from the two data examples illustrate the benefits of combing RCTs and OBs in a single meta-analysis. In both examples, the inclusion of OBs resulted in substantial higher point-estimates of the between-studies correlation and narrower 95% CrIs, leading to improved inferences about the trial-level surrogacy patterns. Specifically in the first data example in aCRC, we found strong trial-level association between the treatment effect on PFS at 1 year and OS at 2 years when RCTs and OBs were included in a single meta-analysis; the 95% CrI of the between-studies correlation contained only positive values and the point-estimate was higher compared to the meta-analysis based on RCTs alone. Substantial improvement in the validation of CCyR at 1 year as a surrogate endpoint of OS at 4 years was observed in the second data example in CML. Specifically, the point-estimate (posterior median) of between-studies correlation was more than twice as high and the 95% CrI was considerably narrower when OBs were incorporated into the meta-analysis.

Benefits in the precision of the predicted true effects on the final outcome were also observed. When a cross-validation was carried out on the data-sets consisting of RCTs and OBs, the model resulted in narrower predictive intervals compared to the cross-validation based on RCTs alone. This confirms the improvement in the trial-level validation of the candidate endpoints in both data examples, as the strength (or weakness) of the surrogate relationship manifests itself in the width of the predicted interval of the true treatment effects.

Although the proposed methodology offers a flexible framework for combining evidence from different study designs into a single meta-analysis, the plausibility of the assumptions of the method should be investigated in depth and potential limitations should be identified. The proposed methods were implemented by assuming that the treatment effects and the baseline effects on the surrogate endpoint and the final outcome, obtained from RCTs and OBs, are exchangeable (random effects). As discussed by White et al. [216], models with random baseline effects have appealing properties and are extremely useful in solving otherwise impossible problems, but their main weakness is susceptibility to bias when there are systematic differences between data from studies of different designs. In these situations, the assumption of exchangeability of the baseline effects may be too strong, as the estimated baseline effects are shrunk toward the overall mean; therefore, the estimation of the treatment effect within-a study is influenced by information outside the study. This conflicts with the principle that treated individuals should only be compared with randomized controls [217] and as such may compromise the randomisation [216, 218]. However, many authors stated that, 'in practice little harm is likely to be done by this' [217, 219]. In an analysis based on hypothetical data, White et al.[216] found that compromising randomisation through random baseline effects can introduce important bias to the analysis, however, they concluded that more research was required to identify any situations where this could be of practical importance.

Another limitation of the method was highlighted in the simulation study. According to the results, the inclusion of OBs for the baseline arm, only marginally improve the precision and the accuracy of the estimates of the between-studies correlation. As explained above, this is due to lack of symmetry at the within-study level of the hierarchical model. The part of the method which models arm A does not contribute directly to the relative treatment effects and hence inclusion of OBs for the baseline arm does not substantially improve the inferences about the between-studies correlation. On the other hand, OBs for the experimental treatment arm offer significant gains in precision and accuracy of the estimates of the between-studies correlation. This problem can be resolved by introducing a symmetry in the model parameterisation to account for treatment effects in both arms. This will allow OBs reporting data in the baseline arm to contribute directly to the estimation of the treatment effects.

As discussed above, the proposed methods combine evidence from different study designs by using random baseline effects. An alternative way to incorporate single-arm OBs in a single meta-analysis, is via a matching technique. Complex methods such as propensity scores or matching adjusted indirect comparisons make the use of IPD to match single-arm OBs, whilst adjusting for covariates to reduce the impact of selection bias [220–224]. However in practice, IPD are rarely available. Matching techniques based solely on aggregate data have been also discussed by many authors [225–227], however, their results need to be interpreted with caution as they tend to underestimate the uncertainty and consequently are prone to bias [227].

In summary, we extended the method proposed by Begg and Pilote for combining data from RCTs and single-arm studies into a bivariate method in the Bayesian framework. The method allows for inclusion of evidence from different study designs enhancing the inferences about the parameters describing the trial-level surrogacy patterns. The method allow us to efficiently identify biases in the OBs and predict the true treatment effects with reduced uncertainty. Overall, the proposed method can improve the trial-level validation of surrogate endpoints, and in particular in the era of precision medicine where the quick approval of new promising therapies warrants the inclusion of all available evidence in the analysis.

Chapter 7

Discussion

In this concluding chapter, we give brief overview of the thesis, summarising the main findings of the undertaken work, discussing the strengths and the limitations. The Chapter concludes by highlighting opportunities for further work.

7.1 Summary of the thesis

This thesis considers a range of methodological challenges related to the trial-level validation of surrogate endpoints in disease areas where targeted treatments have been used. It discusses novel methodology developed to address these challenges. The methodological solutions were proposed to achieve the following three aims:

- To improve the trial-level validation of surrogate endpoints within a specific class of treatment in disease areas where trial-level surrogacy patterns vary across treatment classes.
- To improve the trial-level validation of surrogate endpoints, when such validation is based on correlated binomial aggregate data within high or low proportions of events.
- To strengthen the trial-level validation of surrogate endpoints when Randomised controlled trials (RCTs) offer limited information and external evidence from different study designs are needed for such validation.

Chapter 1 outlined the aims and the structure of thesis. It also provided a brief introduction to the background of the thesis and discussed the concepts of surrogate endpoints. Chapter 2 focused on making a brief review of the Bayesian statistics and the meta-analytic methods. In addition to this, it discussed about the Markov chain Monte Carlo methods and the statistical software used in Chapters 4,5 and 6. The third Chapter reviewed existing methodology developed for evaluation of surrogacy patterns.

Chapters 4, 5, and 6 consisted of the main body of the thesis and proposed novel methodology, addressing a number of methodological challenges to achieve the above three aims. Each Chapter started with an introduction section reviewing the background of each methodological challenge and presented a motivating case study to illustrate and the proposed methods. Then a methods section followed where the existing and the proposed modeling methods were described in detail. Each Chapter contained a simulation study, to evaluate the performance of the proposed modeling approaches and compare them against standard methodology. The same structure was used across the three simulation studies, defining the aims of each simulation study, the generation process of the data, the estimands and the performance measures, and using, similar graphical presentations of the results. Each simulation study concluded with a discussion of the key findings. Furthermore, data examples were presenting across all the Chapters illustrating the performance of the methods in real data. The Chapters concluded with a discussion section, which was framed in the form of conclusions, recommendations, limitations of the proposed methods and suggestions for further work.

7.2 Strengths of the thesis

This sections reviews the strengths of the methodology for trial-level surrogate endpoint evaluation, developed in this thesis.

In Chapter 4, we aimed to improve the trial-level validation of surrogate endpoints within a specific class of treatment in disease areas where trial-level surrogacy patterns vary across treatment classes. Two extensions (F-EX, P-EX models) of a model proposed by Daniels and Hughes [13] were developed accounting for differences in surrogate relationships between treatment classes and assuming some level of similarity between them. The first extension (F-EX model) allowed for full borrowing of information for surrogate relationships across treatment classes assuming exchangeability for the parameters describing the surrogate relationships whilst, the second method (P-EX) relaxed this assumption, allowing for partial borrowing of information. The proposed methods showed a lot of potential in terms of improving the trial-level validation of surrogate endpoints within classes, as they resulted in substantial reduction in the uncertainty of the parameters describing a surrogate relationship within a treatment class compared to subgroup analysis in the simulated data scenarios. P-EX model was the best method in data scenarios where there was a treatment class with distinctly different surrogacy pattern (slope), as it was able to estimate the correct degree of borrowing of information for this parameter. The proposed methods outperformed subgroup analysis also in a data example in aCRC (consisting of three classes of treatment), where the trial-level surrogacy patterns differ across treatment classes.

Chapter 5, introduced novel methodology to enhance the trial-level validation of surrogate endpoints, when such validation was based on correlated binomial aggregate data with high or low proportions of events. As discussed, such data typically include two sources of association - one at the individual-level and one at the study-level. The proposed method (BRMA-BC) accounted for within-study associations and modeled the within-study variability on the original binomial scale avoiding the controversial normal approximation (which was used by the standard methodology (model BRMA)). This was implemented by modeling the aggregate data on each outcome jointly, using a bivariate distribution with binomial marginal distributions constructed with copula. To highlight the importance of accounting for within-study associations we also presented another approach (BRMA-IB), which modeled the within-study variability on the original binomial scale, but ignored within-study associations. Overall, the proposed method (BRMA-BC) was able to improve the inferences about the trial-level validation of surrogate endpoints in terms of precision and bias compared to the standard methodology (BRMA), as it resulted in more precise and less biased estimates of the parameters describing a surrogate relationship in a series of simulated data scenarios with high proportions of events. The proposed method (BRMA-BC) resulted also in higher and slightly more precise estimates of the parameters describing a surrogate relationship compared to BRMA in a data example in CML, where the high effectiveness of targeted therapies led to large proportions of treatment responders and very small proportions of disease progressions or deaths. However, the difference was not that pronounced as in the simulation study. This may have occurred due to the fact that the data on the final outcome were not mature, resulting in the treatment effects on the final outcome obtained with large uncertainty.

In chapter 6, we aimed to strengthen the trial-level validation of surrogate endpoints, in situations where RCTs offer limited evidence and external evidence from different study designs were required for such validation. The proposed methodology offered a flexible framework for incorporating OBs in bivariate meta-analysis. It extended the model proposed by Begg and Pilote into a bivariate hierarchical method (model M1) combining RCT data and evidence from single arm observational studies in a single analysis. Two alternative versions of the method were also introduced. The first one accounted for bias in the OBs by introducing bias terms (model M2), whilst the second version modeled the within-study variability on the binomial scale using bivariate densities with binomial marginals constructed with copulas. This allowed as to avoid the controversial normal approximation, when modeling binomial data with high proportions of events (model M3). Based on the results from the simulation study, we inferred that the inclusion OBs in the analysis could lead to improved trial-level validation of surrogate endpoints, as it resulted in estimates of the between-studies correlation obtained with improved precision and accuracy compared to the analysis based on RCTs alone in the most of the scenarios of the simulation study. Similar behaviour was also observed in the two data examples (aCRC and CML), used to illustrate the method. Furthermore, the results of the cross-validation procedure, performed in the two data examples, suggested similar benefits in terms of the improved precision of the predictions of the treatment effect on the final outcome.

7.3 Discussion of the limitations

This sections discusses the key limitations of the proposed methodology in this thesis.

In Chapter 4, the proposed methodology illustrated the benefits of borrowing of information across classes of treatment for the parameters describing the surrogate relationship. Particularly, the more classes we have in the data the easier it is for the models to borrow information across them. However in practice, it can be challenging to find data sets with sufficient number of treatment classes. A small number of treatment classes in the data can affect the performance of the proposed methods substantially reducing the impact of borrowing of information [133]. Additionally, applying P-EX model to data sets with only three treatment classes may lead to a situation where only one class is deemed exchangeable by the model during the estimation process (in some of the MCMC iterations). However, this is not possible as there is no other class to exchange information with. Therefore, this should always be investigated when P-EX is applied to data sets with only three treatment classes. In the data example of Chapter 4, this issue did not affect the performance of P-EX model as it occurred only in the 0.5% of the MCMC iterations. Furthermore, the evaluation framework proposed by Daniels and Hughes assesses the strength of trial-level surrogate relationships by examining whether zero is contained in the CrIs of the parameters describing the surrogate relationships (λ_1 and λ_0). This is very restrictive, as the width of the CrIs depends on the number of studies included in the analysis, hence they may lead to increased uncertainty around the intercept and slope, invalidating one or some the surrogacy criteria. Therefore, alternative evaluation frameworks should always be taken into account. For instance, other authors have been more flexible emphasising on the balance between the actual need for a surrogate endpoint and the strength of the trial-level surrogacy pattern [21]. In addition to this, an alternative evaluation framework could focus largely on the predictions of the true treatment effect of the final outcome, as the strength or weakness of a surrogate relationship will be evident in the uncertainty around the predicted treatment effect on the final outcome [136].

In Chapter 5, we implemented the proposed methodology (BRMA-BC model) using RStan [56]. A limitation of the method was the fact that BRMA-BC model was very sensitive to initial values. Therefore, the initiation of the estimation process (HMC) was very difficult without setting "sensible" initial values. This was tackled by fitting BRMA-IB prior to BRMA-BC and then converting the estimates of BRMA-IB to initial values for BRMA-BC. However, this issue makes the use of BRMA-BC model quite restrictive, as it requires either another method to be fitted prior to BRMA-BC model or a detailed understanding of the data set in order to provide the model with "sensible" initial values. The data example in CML consisted of RCTs reporting data on CCyR at 1 year (surrogate endpoint) and OS or EFS at two years. Unfortunately, the definition of EFS was the relatively inconsistent across studies. For example, in some RCTs, the definition of EFS overlapped with the definition of PFS and in some of others the definition was more vague including various events. Additionally, another limitation of the CML data example was the lack of IPD across all the RCTs. As a result, the accurate estimation of the within-study associations (within-study correlations for BRMA and dependence parameters for BRMA-BC) was not possible. To address this, we used evidence from three observational cohort studies to inform prior distributions of the within-study association parameters.

In Chapter 6, we proposed a hierarchical method to combine evidence from different study designs into a single meta-analysis. The model combined RCTs and OBs assuming that the baseline effects were exchangeable across all studies. This meant that the baseline effect estimated within a particular study was influenced by information outside that study. Many authors have criticised this assumption, arguing that it compromises the randomisation [216, 218]. However, it is not clear to what extent this is a problem in practice. Senn et al. [217] stated that "in practice little harm is likely to be done" and other authors such as Achana et al. [219] found little bias in their analysis when this assumption was used. A simulation study conducted by White et al. [216], showed that compromising randomisation through random baseline effects can introduce bias to the analysis, however, the authors concluded that more research was required to identify any situations where this could be of practical importance. Another limitation of the proposed hierarchical method was highlighted in our simulation study. The inclusion of OBs for the control arm, had only a marginal effect on the inferences of the trial-level surrogacy patterns. On the other hand, including OBs reporting data in the experimental arm resulted in substantial improvement of the inferences of the trial-level surrogacy patterns. This was due to lack of symmetry at the within-study level of the hierarchical model and data reported in arm A did not contribute directly to the estimation of relative treatment effects, whilst data reported in arm B contributed both to the estimation of the baseline effects and the estimation of relative effects thus improving evidence base for surrogate endpoint evaluation.

7.4 Future work

This section outlines potential methodological extensions of the work presented in this thesis highlighting the opportunities for further work in the area of the trial-level validation of surrogate endpoints.

In Chapter 4, the proposed methodology improved the trial-level validation of surrogate endpoints within a specific class of treatment in a disease area, borrowing information for the parameters describing the surrogate relationship from other treatment classes in that disease area. Further methodological work can be undertaken by extending the methods to account for differences in lines of treatments or to account for different treatments within a treatment class. This can be done by adding more layers of hierarchy in the model. However, a relatively large number of studies for each line of treatment (or each treatment) will be required to fit such model and obtain estimates of the parameters describing the surrogate relationships without considerable uncertainty.

P-EX model was the most flexible approach achieving superior performances in data scenarios where there was a treatment class with distinctly different slope. This was due to the assumption of partial exchangeability which allowed the model to regulate the degree of borrowing of information for the parameter of the slopes. This assumption can easily be applied to the intercepts and the conditional variances (the other two parameters describing a surrogate relationship in the evaluation framework proposed by Daniels & Hughes).

In Chapter 5, we developed methodology which improved the trial-level validation of surrogate endpoints, when such validation is based on correlated binomial aggregate data within high or low proportions of events. BRMA-BC method modeled the within-study variability on the original binomial scale accounting for within-study associations using bivariate joint densities constructed with copula functions. This model can be extended in various ways. Under the current parameterisation, the model implies a linear relationship between the true probabilities of events on the first and the second outcome on a transformed scale using *logit* link function. Hence under this parameterisation, the correlation between true probabilities of events is expressed on the *logit* scale which is less intuitive. A more intuitive approach would allow for modeling the correlation on the original scale. This can be implemented by using copulas in a similar way as Chu et al. and Nikolopoulos have proposed [169, 170]. Furthermore, as discussed by Bujkiewicz et al. [89], BRMA (standard method for trial-level surrogate endpoint evaluation) can be extended to the multivariate case to account for multiple surrogate endpoints. Similarly, BRMA-BC can be extended to allow for modeling multiple surrogate endpoints (or the same surrogate endpoint but reported at multiple time points) with the use of vine-copulas.

In chapter 6, the proposed hierarchical method enhanced the trial-level validation of surrogate endpoints by combining RCTs and OBs in a single analysis. Further research can be undertaken by investigating other approaches. For instance, an alternative way to incorporate single-arm OBs in a single meta-analysis, is via a matching technique. Propensity scoring or matching adjusted indirect comparisons can be applied to match single-arm observational data, however, they require use of IPD [220–224]. Practically, very often this is not feasible, as IPD is rarely available. Matching techniques based only on aggregate data have been discussed by some authors [225–227], but their results need to be interpreted with caution as they may underestimate the uncertainty and consequently, are prone to bias [227]. Schmitz et al. [227] proposed a matching strategy which incorporates single-arm OBs in the analysis accounting for all the relevant uncertainty to connect disconnected networks. This strategy identifies the most important baseline characteristics (covariates) in a disease area, assigning weights to them. These characteristics are used to calculate a distance metric which is used to measure the of similarity between any two of single arm OBs included in the data set. Based on the values of the distance metric, the single-arm OBs can be matched to act as active treatment and control arms in a pseudo comparative study. Schmitz et al. highlighted the importance of exploring the space of possible matches and assessing the impact different matches have on the results. This approach can easily be applied to different settings and, therefore, can potentially be used to strengthen the inferences of the parameters describing surrogate relationships.

7.5 Conclusion

In conclusion, this thesis has proposed methodological tools to improve the trial-level validation of surrogate endpoints. The models were developed under the Bayesian framework and can be easily generalised to other research areas facing similar methodological challenges. The work presented in Chapter 4, can assist trial-level validation of surrogate endpoints in newer classes of treatment where the validation is problematic due to the sparsity of the data. This will accelerate the evaluation process of drugs which normally can last several years. The proposed methods in Chapter 5 are able to improve the trial-level validation of binary outcomes as surrogate endpoints in disease areas with high/low proportions of events. This is rather important when such validation is based on data from modern clinical trials assessing personalised treatments, as the increased effectiveness of those treatments often leads to high numbers of responses and reduces the numbers of events. The work presented in Chapter 6 can facilitate the approval of new targeted therapies, when RCTs offer limited evidence and the evaluation process is based on long term outcomes, as it allows for the inclusion of all available evidence.

Bibliography

[1] Manfred Kunz. Genomic signatures for individualized treatment of malignant tumors. *Current drug discovery technologies*, 5(1):9–14, 2008.

[2] Paulina Krzyszczyk, Alison Acevedo, Erika J Davidoff, Lauren M Timmins, Ileana Marrero-Berrios, Misaal Patel, Corina White, Christopher Lowe, Joseph J Sherba, Clara Hartmanshenn, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology*, 6(03 – 04):79–100, 2018.

[3] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[4] Stephen G O'Brien, François Guilhot, Richard A Larson, Insa Gathmann, Michele Baccarani, Francisco Cervantes, Jan J Cornelissen, Thomas Fischer, Andreas Hochhaus, Timothy Hughes, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *New England Journal of Medicine*, 348(11):994–1004, 2003.

[5] A Hochhaus, SG O'brien, François Guilhot, BJ Druker, S Branford, L Foroni, JM Goldman, MC Müller, JP Radich, M Rudoltz, et al. Six-year follow-up of patients receiving imatinib for the first-line treatment of chronic myeloid leukemia. *Leukemia*, 23(6):1054–1061, 2009.

[6] Hugues De Lavallade, Jane F Apperley, Jamshid S Khorashad, Dragana Milojkovic, Alistair G Reid, Marco Bua, Richard Szydlo, Eduardo Olavarria, Jaspal Kaeda, John M Goldman, et al. Imatinib for newly diagnosed patients with chronic myeloid leukemia: incidence of sustained responses in

an intention-to-treat analysis. *Journal of Clinical Oncology*, 26(20):3358–3363, 2008.

[7] EMA. European medicines agency, final report from the emea/chmp-think-tank group on innovative drug development. *available at http://bit.ly/2rCmLAH*, 2007.

[8] Stephen Joel Coons. The fda's critical path initiative: a brief introduction. *Clinical therapeutics*, 31(11):2572–2573, 2009.

[9] ED Saad and Marc Buyse. Statistical controversies in clinical research: end points other than overall survival are vital for regulatory approval of anticancer agents. *Annals of Oncology*, 27(3):373–378, 2015.

[10] S. Bujkiewicz, F. Achana, T. Papanikos, R.D. Riley, and K. Abrams. Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. *NICE DSU Technical Support*, Document 20, 2019.

[11] Oriana Ciani, Marc Buyse, Ruth Garside, Jaime Peters, Everardo D Saad, Ken Stein, and Rod S Taylor. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. *Journal of clinical epidemiology*, 68(7):833–842, 2015.

[12] Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Daniel J Sargent, Langdon L Miller, Gary L Elfring, Jean-Pierre Pignon, and Pascal Piedbois. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology*, 25(33):5218–5224, 2007.

[13] Michael J Daniels and Michael D Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in medicine*, 16(17):1965–1982, 1997.

[14] Hans C Van Houwelingen, Koos H Zwinderman, and Theo Stijnen. A bivariate approach to meta-analysis. *Statistics in medicine*, 12(24):2273–2284, 1993.

[15] Richard D Riley, KR Abrams, PC Lambert, AJ Sutton, and JR Thompson. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in medicine*, 26(1):78–97, 2007.

[16] Sylwia Bujkiewicz, John R Thompson, Enti Spata, and Keith R Abrams.

Uncertainty in the bayesian meta-analysis of normally distributed surrogate endpoints. *Statistical methods in medical research*, 26(5):2287–2318, 2015.

[17] Taye H Hamza, Hans C van Houwelingen, and Theo Stijnen. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of clinical epidemiology*, 61(1):41–51, 2008.

[18] M Baccarani, M Dreyling, and ESMO Guidelines Working Group. Chronic myeloid leukaemia: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of oncology*, 21(5):165–167, 2010.

[19] Hagop M Kantarjian, Neil P Shah, Jorge E Cortes, Michele Baccarani, Mohan B Agarwal, María Soledad Undurraga, Jianxiang Wang, Juan Julio Kassack Ipiña, Dong-Wook Kim, Michinori Ogura, et al. Dasatinib or imatinib in newly diagnosed chronic-phase chronic myeloid leukemia: 2-year follow-up from a randomized phase 3 trial (dasision). *Blood*, 119(5):1123–1129, 2012.

[20] Thomas R. Fleming. Surrogate endpoints in clinical trials. *Drug Information Journal*, 30(2):545–551, 1996.

[21] Ariel Alonso, Theophile Bigirumurame, Tomasz Burzykowski, Marc Buyse, Geert Molenberghs, Leacky Muchene, Nolen Joy Perualila, Ziv Shkedy, and Wim Van der Elst. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. CRC Press, 2016.

[22] Susan S Ellenberg and J Michael Hamilton. Surrogate endpoints in clinical trials: cancer. *Statistics in medicine*, 8(4):405–413, 1989.

[23] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95, 2001.

[24] Tomasz Burzykowski, Geert Molenberghs, and Marc Buyse. *The evaluation of surrogate endpoints*. Springer Science & Business Media, 2006.

[25] Thomas R Fleming and David L DeMets. Surrogate end points in clinical

trials: are we being misled? *Annals of internal medicine*, 125(7):605–613, 1996.

[26] Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine*, 321(6):406–412, 1989.

[27] Thomas R Fleming. Surrogate markers in aids and cancer trials. *Statistics in Medicine*, 13(13-14):1423–1435, 1994.

[28] V DeGruttola, T R. Fleming, DY Lin, and R Coombs. Perspective: Validating surrogate markers–are we being naive? *The Journal of infectious diseases*, 175:237–46, 03 1997.

[29] Ann E Ferentz. Integrating pharmacogenomics into drug development. *Pharmacogenomics*, 3(4):453–467, 2002.

[30] Arthur Schatzkin and Mitchell Gail. The promise and peril of surrogate end points in cancer research. *Nature Reviews Cancer*, 2(1):19, 2002.

[31] Phillips Alan and Haudiquet Vincent. International conference on harmonisation e9 guideline 'statistical principles for clinical trials': a case study. *Statistics in Medicine*, 22(1):1–11, 2003.

[32] Mitchell H Gail, Ruth Pfeiffer, Hans C Van Houwelingen, and Raymond J Carroll. On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1(3):231–246, 2000.

[33] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

[34] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[35] Thomas A Louis. Using empirical bayes methods in biopharmaceutical research. *Statistics in Medicine*, 10(6):811–829, 1991.

[36] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

[37] David J Spiegelhalter, Jonathan P Myles, David R Jones, and Keith R Abrams. An introduction to bayesian methods in health technology assessment. *Bmj*, 319(7208):508–512, 1999.

[38] Paul C Lambert, Alex J Sutton, Paul R Burton, Keith R Abrams, and David R Jones. 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, 2005.

[39] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[40] Rutger van Haasteren. Marginal likelihood calculation with mcmc methods. In *Gravitational Wave Detection and Data Analysis for Pulsar Timing Arrays*, pages 99–120. Springer, 2014.

[41] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[42] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[43] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6(6):721–741, 1984.

[44] Andrew Gelman, Kenneth Shirley, et al. Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, 6:163–174, 2011.

[45] Radford M Neal. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.

[46] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter.

Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[47] David Lunn, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC, 2012.

[48] David Spiegelhalter, Andrew Thomas, Nicky Best, and Dave Lunn. Winbugs user manual, 2003.

[49] Sibylle Sturtz, Uwe Ligges, and Andrew Gelman. R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, 12(3):1–16, 2005.

[50] Sibylle Sturtz, Uwe Ligges, and Andrew Gelman. R2openbugs: a package for running openbugs from r. *available at http://cran. rproject. org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS. pdf*, 2010.

[51] Mary Kathryn Cowles. *Applied Bayesian statistics: with R and OpenBUGS examples*, volume 98. Springer Science & Business Media, 2013.

[52] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[53] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[54] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[55] Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.

[56] Team Stan Development. Rstan: the r interface to stan. *available at https://mc-stan.org/users/interfaces/rstan*, 2016.

[57] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[58] Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.

[59] Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

[60] Maria I Gorinova, Dave Moore, and Matthew D Hoffman. Automatic reparameterisation of probabilistic programs. *arXiv preprint arXiv:1906.03028*, 2019.

[61] Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

[62] Gene V Glass. Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8, 1976.

[63] Richard Peto. Why do we need systematic overviews of randomized trials?(transcript of an oral presentation, modified by the editors). *Statistics in medicine*, 6(3):233–240, 1987.

[64] Salim Yusuf, Janet Wittes, and Lawrence Friedman. Overview of results of randomized clinical trials in heart disease: I. treatments following myocardial infarction. *Jama*, 260(14):2088–2093, 1988.

[65] R Barker Bausell, Yu-Fang Li, Meei-Ling Gau, and Karen L Soeken. The growth of meta-analytic literature from 1980 to 1993. *Evaluation & the Health Professions*, 18(3):238–251, 1995.

[66] Richard Peto, Rory Collins, and Richard Gray. Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of clinical epidemiology*, 48(1):23–40, 1995.

[67] Christopher H Schmid. Using bayesian inference to perform meta-analysis. *Evaluation & the health professions*, 24(2):165–189, 2001.

[68] Alex J Sutton, Keith R Abrams, David R Jones, David R Jones, Trevor A Sheldon, and Fujian Song. *Methods for meta-analysis in medical research*, volume 348. Wiley Chichester, 2000.

[69] Nicky J Welton, Alexander J Sutton, Nicola Cooper, Keith R Abrams, and AE Ades. *Evidence synthesis for decision making in healthcare*, volume 132. John Wiley & Sons, 2012.

[70] DR Cox. The analysis of binary data, methuen & co. *Ltd., London*, pages 48–52, 1970.

[71] Lincoln E Moses, David Shapiro, and Benjamin Littenberg. Combining independent studies of a diagnostic test into a summary roc curve: data-analytic approaches and some additional considerations. *Statistics in medicine*, 12(14):1293–1316, 1993.

[72] Teresa C Smith, David J Spiegelhalter, and Andrew Thomas. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in medicine*, 14(24):2685–2699, 1995.

[73] Michael Borenstein, H Cooper, L Hedges, and J Valentine. Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, 2:221–235, 2009.

[74] Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

[75] Anne Whitehead. *Meta-analysis of controlled clinical trials*, volume 7. John Wiley & Sons, 2002.

[76] Susanne Hempel, Jeremy NV Miles, Marika J Booth, Zhen Wang, Sally C Morton, and Paul G Shekelle. Risk of bias: a simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic reviews*, 2(1):107, 2013.

[77] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

[78] Laurence S Freedman, Barry I Graubard, and Arthur Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, 11(2):167–178, 1992.

[79] Marc Buyse and Geert Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, pages 1014–1029, 1998.

[80] Geert Molenberghs, Marc Buyse, and Tomasz Burzykowski. A meta-analytic validation framework for continuous outcomes. In *The evaluation of surrogate endpoints*, pages 95–120. Springer, 2005.

[81] DY Lin, TR Fleming, and V De Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine*, 16(13):1515–1527, 1997.

[82] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[83] Tomasz Burzykowski and Marc Buyse. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, 5(3):173–186, 2006.

[84] Marc Buyse, Geert Molenberghs, Tomasz Burzykowski, Didier Renard, and Helena Geys. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1):49–67, 2000.

[85] Tomasz Burzykowski, Geert Molenberghs, Marc Buyse, Helena Geys, and Didier Renard. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(4):405–422, 2001.

[86] Lindsay A Renfro, Qian Shi, Daniel J Sargent, and Bradley P Carlin. Bayesian adjusted r2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in medicine*, 31(8):743–761, 2012.

[87] Sylwia Bujkiewicz, John R Thompson, Richard D Riley, and Keith R Abrams. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics in medicine*, 2015.

[88] Hans C van Houwelingen, Lidia R Arends, and Theo Stijnen. Tutorial in biostatistics advanced methods in meta-analysis: multivariate approach and meta-regression. *Tutorials in Biostatistics, Tutorials in Biostatistics: Statistical Modelling of Complex Medical Data*, 2:289, 2005.

[89] Sylwia Bujkiewicz, John R Thompson, Alex J Sutton, Nicola J Cooper, Mark J Harrison, Deborah PM Symmons, and Keith R Abrams. Multivariate meta-analysis of mixed outcomes: a bayesian approach. *Statistics in Medicine*, 32(22):3926–3943, 2013.

[90] William DuMouchel. *Hierarchical Bayes linear models for meta-analysis.* National Institute of Statistical Sciences, 1974.

[91] William E Strawderman. Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):385–388, 1971.

[92] James O Berger. *Statistical decision theory and Bayesian analysis.* Springer Science & Business Media, 2013.

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

[94] Martin W McIntosh. The population risk as an explanatory variable in research synthesis of clinical trials. *Statistics in medicine*, 15(16):1713–1728, 1996.

[95] RD Riley, MJ Price, D Jackson, M Wardle, François Gueyffier, J Wang, Jan A Staessen, and IR White. Multivariate meta-analysis using individual participant data. *Research synthesis methods*, 6(2):157–174, 2015.

[96] Victor G De Gruttola, Pamela Clax, David L DeMets, Gregory J Downing, Susan S Ellenberg, Lawrence Friedman, Mitchell H Gail, Ross Prentice, Janet Wittes, and Scott L Zeger. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a national institutes of health workshop. *Controlled clinical trials*, 22(5):485–502, 2001.

[97] Jane Xu and Scott L Zeger. The evaluation of multiple surrogate endpoints. *Biometrics*, 57(1):81–87, 2001.

[98] Thomas R Fleming and John H Powers. Biomarkers and surrogate endpoints in clinical trials. *Statistics in medicine*, 31(25):2973–2984, 2012.

[99] Clemens Giessen, Ruediger Paul Laubender, Donna Pauler Ankerst, Sebastian Stintzing, Dominik Paul Modest, Ulrich Mansmann, and Volker Heinemann. Progression-free survival as a surrogate endpoint for median overall survival

in metastatic colorectal cancer: literature-based analysis from 50 randomized first-line trials. *Clinical Cancer Research*, 19(1):225–235, 2013.

[100] Costel Chirila, Dawn Odom, Giovanna Devercelli, Shahnaz Khan, Bintu N Sherif, James A Kaye, István Molnár, and Beth Sherrill. Meta-analysis of the association between progression-free survival and overall survival in metastatic colorectal cancer. *International journal of colorectal disease*, 27(5):623–634, 2012.

[101] Donald A. Berry. Subgroup analyses. *Biometrics*, 46(4):1227–1230, 1990.

[102] Beat Neuenschwander, Simon Wandel, Satrajit Roychoudhury, and Stuart Bailey. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical statistics*, 15(2):123–134, 2016.

[103] Tasos Papanikos, John R Thompson, Keith R Abrams, Nicolas Städler, Oriana Ciani, Rod Taylor, and Sylwia Bujkiewicz. Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data. *Statistics in medicine*, 39(8):1103–1124, 2020.

[104] Mariana Macedo, Fernanda M Melo, Héber Ribeiro, Márcio Carmona Marques, Luciane T Kagohara, Maria Begnami, Julio C Neto, Júlia S Ribeiro, Fernando Soares, Dirce Carraro, and Isabela Cunha. Kras mutation status is highly homogeneous between areas of the primary tumor and the corresponding metastasis of colorectal adenocarcinomas: One less problem in patient care. *American journal of cancer research*, 7:1978–1989, 09 2017.

[105] Kimberly Perez, Robert Walsh, Kate E. Brilliant, Leila Nobel, Evgeny Yakirevich, Virginia Breese, Cynthia Jackson, Devasis Chatterjee, Victor Pricolo, Leslie Roth, Nishit Shah, Thomas Cataldo, Howard Safran, Douglas Hixson, and Peter J. Quesenberry. Heterogeneity of colorectal cancer (crc) in reference to kras proto-oncogene utilizing wave technology. *Journal of Clinical Oncology*, 31(15):14637–14637, 2013.

[106] Helen A Dakin, Nicky J Welton, AE Ades, Sarah Collins, Michelle Orme, and Steven Kelly. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary

open-angle glaucoma and ocular hypertension. *Statistics in medicine*, 30(20):2511–2535, 2011.

[107] Fiona C Warren, Keith R Abrams, and Alex J Sutton. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics in medicine*, 33(14):2449–2466, 2014.

[108] Marta O Soares, Jo C Dumville, AE Ades, and Nicky J Welton. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 259–279, 2014.

[109] Rhiannon K Owen, Douglas G Tincello, and R Abrams Keith. Network meta-analysis: development of a three-level hierarchical modeling approach incorporating dose-related constraints. *Value in Health*, 18(1):116–126, 2015.

[110] Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[111] Thomas A. Louis. Estimating a population of parameter values using bayes and empirical bayes methods. *Journal of the American Statistical Association*, 79(386):393–398, 1984.

[112] Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis*, volume 88. Chapman & Hall/CRC Boca Raton, 2000.

[113] Jean-Marie Grouin, Maylis Coste, and John Lewis. Subgroup analyses in randomized clinical trials: Statistical and regulatory issues. *Journal of Biopharmaceutical Statistics*, 15(5):869–882, 2005. PMID: 16078390.

[114] Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934, 1995.

[115] Isabella Verdinelli and Larry Wasserman. Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995.

[116] Scott M Berry, Kristine R Broglio, Susan Groshen, and Donald A Berry. Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase ii oncology clinical trials. *Clinical Trials*, 10(5):720–734, 2013.

[117] Thall Peter F., Wathen J. Kyle, Bekele B. Nebiyou, Champlin Richard E., Baker Laurence H., and Benjamin Robert S. Hierarchical bayesian approaches to phase ii trials in diseases with multiple subtypes. *Statistics in Medicine*, 22(5):763–780, 2003.

[118] Rashmi Chugh, J Kyle Wathen, Robert G Maki, Robert S Benjamin, Shreyaskumar R Patel, PA Meyers, Dennis A Priebat, Denise K Reinke, Dafydd G Thomas, Mary L Keohan, et al. Phase ii multicenter trial of imatinib in 10 histologic subtypes of sarcoma using a bayesian hierarchical statistical model. *J Clin Oncol*, 27(19):3148–53, 2009.

[119] Bradley Efron. *Bootstrap methods: another look at the jackknife*. Springer, 1992.

[120] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, and Victor Aboyans et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2095–2128, 12 2012.

[121] GJ Poston, D Tait, S O'Connell, A Bennett, and S Berendse. Diagnosis and management of colorectal cancer: summary of nice guidance. *Bmj*, 343:d6751, 2011.

[122] Ahmedin Jemal, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, Taylor Murray, and Michael J Thun. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, 58(2):71–96, 2008.

[123] A de de Gramont, A Figer, M Seymour, M Homerin, A Hmissi, J Cassidy, C Boni, H Cortes-Funes, A Cervantes, G Freyer, et al. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *Journal of Clinical Oncology*, 18(16):2938–2947, 2000.

[124] Herbert Hurwitz, Louis Fehrenbacher, William Novotny, Thomas Cartwright, John Hainsworth, William Heim, Jordan Berlin, Ari Baron, Susan Griffing,

Eric Holmgren, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *New England journal of medicine*, 350(23):2335–2342, 2004.

[125] Eric Van Cutsem, Claus-Henning Köhne, Erika Hitre, Jerzy Zaluski, Chung-Rong Chang Chien, Anatoly Makhson, Geert D'haens, Tamás Pintér, Robert Lim, György Bodoky, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *New England Journal of Medicine*, 360(14):1408–1417, 2009.

[126] Jean-Yves Douillard, Salvatore Siena, James Cassidy, Josep Tabernero, Ronald Burkes, Mario Barugel, Yves Humblet, György Bodoky, David Cunningham, Jacek Jassem, et al. Randomized, phase iii trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (folfox4) versus folfox4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the prime study. *Journal of clinical oncology*, 28(31):4697–4705, 2010.

[127] Alexander Kumachev, Marie Yan, Scott Berry, Yoo-Joung Ko, Maria CR Martinez, Keya Shah, and Kelvin KW Chan. A systematic review and network meta-analysis of biologic agents in the first line setting for advanced colorectal cancer. *PloS one*, 10(10):e0140187, 2015.

[128] Tomasz Burzykowski, Geert Molenberghs, and Marc Buyse. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(1):103–124, 2004.

[129] Jaafar Bennouna, Javier Sastre, Dirk Arnold, Pia Österlund, Richard Greil, Eric Van Cutsem, Roger von Moos, Jose Maria Viéitez, Olivier Bouché, Christophe Borg, et al. Continuation of bevacizumab after first progression in metastatic colorectal cancer (ml18147): a randomised phase 3 trial. *The lancet oncology*, 14(1):29–37, 2013.

[130] ML Rothenberg, JV Cox, C Butts, M Navarro, Y-J Bang, R Goel, S Gollins, LL Siu, S Laguerre, and D Cunningham. Capecitabine plus oxaliplatin (xelox) versus 5-fluorouracil/folinic acid plus oxaliplatin (folfox-4) as second-line

therapy in metastatic colorectal cancer: a randomized phase iii noninferiority study. *Annals of Oncology*, 19(10):1720–1726, 2008.

[131] J Cassidy, S Clarke, E Díaz-Rubio, W Scheithauer, A Figer, R Wong, S Koski, K Rittweger, F Gilberg, and L Saltz. Xelox vs folfox-4 as first-line therapy for metastatic colorectal cancer: No16966 updated results. *British journal of cancer*, 105(1):58, 2011.

[132] Eleni G Elia, Nicolas Städler, Oriana Ciani, Rod S Taylor, and Sylwia Bujkiewicz. Combining tumour response and progression free survival as surrogate endpoints for overall survival in advanced colorectal cancer. *Cancer Epidemiology*, 64:101665, 2020.

[133] Daniel McNeish and Laura M Stapleton. Modeling clustered data with very few clusters. *Multivariate behavioral research*, 51(4):495–518, 2016.

[134] Nicholas R Latimer, Chris Henshall, Uwe Siebert, and Helen Bell. Treatment switching: statistical and decision-making challenges and approaches. *International journal of technology assessment in health care*, 32(3):160–166, 2016.

[135] Tait D Shanafelt, Charles Loprinzi, Randolph Marks, Paul Novotny, and Jeff Sloan. Are chemotherapy response rates related to treatment-induced survival prolongations in patients with advanced cancer. *Journal of clinical oncology*, 22(10):1966–1974, 2004.

[136] Sylwia Bujkiewicz, Dan Jackson, John R Thompson, Rebecca M Turner, Nicolas Städler, Keith R Abrams, and Ian R White. Bivariate network meta-analysis for surrogate endpoint evaluation. *Statistics in medicine*, 38(18):3322–3341, 2019.

[137] Richard D Riley, Dan Jackson, Georgia Salanti, Danielle L Burke, Malcolm Price, Jamie Kirkham, and Ian R White. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *bmj*, 358:3932, 2017.

[138] Haitao Chu and Stephen R Cole. Bivariate meta-analysis of sensitivity

and specificity with sparse data: a generalized linear mixed model approach. *Journal of clinical epidemiology*, 59(12):1331, 2006.

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

[140] Charles L Sawyers. Chronic myeloid leukemia. *New England Journal of Medicine*, 340(17):1330–1340, 1999.

[141] Xuelin Huang, Jorge Cortes, and Hagop Kantarjian. Estimations of the increasing prevalence and plateau prevalence of chronic myeloid leukemia in the era of tyrosine kinase inhibitor therapy. *Cancer*, 118(12):3123–3127, 2012.

[142] Renaud Capdeville, Elisabeth Buchdunger, Juerg Zimmermann, and Alex Matter. Glivec (sti571, imatinib), a rationally developed, targeted anticancer drug. *Nature reviews Drug discovery*, 1(7):493, 2002.

[143] Roger M Harbord, Jonathan J Deeks, Matthias Egger, Penny Whiting, and Jonathan AC Sterne. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*, 8(2):239–251, 2006.

[144] Harry Joe. *Multivariate models and multivariate dependence concepts*. CRC Press, 1997.

[145] Harry Joe. *Dependence modeling with copulas*. Chapman and Hall/CRC, 2014.

[146] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[147] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[148] Christian Genest and Johanna Nešlehová. A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515, 2007.

[149] Peter Xue-Kun Song. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

[150] Janet E Heffernan. A directory of coefficients of tail dependence. *Extremes*, 3(3):279–290, 2000.

[151] Zhi-Jie Kang, Yu-Fei Liu, Ling-Zhi Xu, Zi-Jie Long, Dan Huang, Ya Yang, Bing Liu, Jiu-Xing Feng, Yu-Jia Pan, Jin-Song Yan, et al. The philadelphia chromosome in leukemogenesis. *Chinese journal of cancer*, 35(1):48, 2016.

[152] Tomasz Sacha. Imatinib in chronic myeloid leukemia: an overview. *Mediterranean journal of hematology and infectious diseases*, 6(1), 2014.

[153] Hagop M Kantarjian, Andreas Hochhaus, Giuseppe Saglio, Carmino De Souza, Ian W Flinn, Leif Stenke, Yeow-Tee Goh, Gianantonio Rosti, Hirohisa Nakamae, Neil J Gallagher, et al. Nilotinib versus imatinib for the treatment of patients with newly diagnosed chronic phase, philadelphia chromosome-positive, chronic myeloid leukaemia: 24-month minimum follow-up of the phase 3 randomised enestnd trial. *The lancet oncology*, 12(9):841–851, 2011.

[154] Ciani Oriana, Hoyle Martin, Pavey Toby, Cooper Chris, Garside Ruth, Rudin Claudius, and Taylor Rod. Complete cytogenetic response and major molecular response as surrogate outcomes for overall survival in first-line treatment of chronic myelogenous leukemia: a case study for technology appraisal on the basis of surrogate outcomes evidence. *Value in Health*, 16(6):1081–1090, 2013.

[155] Hagop Kantarjian, Susan O'Brien, Jianqin Shan, Xuelin Huang, Guillermo Garcia-Manero, Stefan Faderl, Farhad Ravandi-Kashani, Srdan Verstovsek, Mary Beth Rios, and Jorge Cortes. Cytogenetic and molecular responses and outcome in chronic myelogenous leukemia: need for new response definitions? *Cancer*, 112(4):837–845, 2008.

[156] Michael W Deininger, Kenneth J Kopecky, Jerald P Radich, Suzanne Kamel-Reid, Wendy Stock, Elisabeth Paietta, Peter D Emanuel, Martin Tallman, Martha Wadleigh, Richard A Larson, et al. Imatinib 800 mg daily induces deeper molecular responses than imatinib 400 mg daily: results of swog s0325, an intergroup randomized phase ii trial in newly diagnosed chronic phase chronic myeloid leukaemia. *British journal of haematology*, 164(2):223–232, 2014.

[157] Claude Preudhomme, Joëlle Guilhot, Franck Emmanuel Nicolini, Agnès Guerci-Bresler, Françoise Rigal-Huguet, Frederic Maloisel, Valérie Coiteux,

Martine Gardembas, Christian Berthou, Anne Vekhoff, et al. Imatinib plus peginterferon alfa-2a in chronic myeloid leukemia. *New England Journal of Medicine*, 363(26):2511–2521, 2010.

[158] Jorge E Cortes, Anish Maru, Cármino Antonio De Souza, Francois Guilhot, Ladan Duvillie, Christine Powell, et al. Bosutinib versus imatinib in newly diagnosed chronic phase chronic myeloid leukemia-bela trial: 24-month follow-up. *Blood*, 118(21):455, 2011.

[159] Jerald P Radich, Kenneth J Kopecky, Frederick R Appelbaum, Suzanne Kamel-Reid, Wendy Stock, Greg Malnassy, Elisabeth Paietta, Martha Wadleigh, Richard A Larson, Peter Emanuel, et al. A randomized trial of dasatinib 100 mg versus imatinib 400 mg in newly diagnosed chronic-phase chronic myeloid leukemia. *Blood*, 120(19):3898–3905, 2012.

[160] Michele Baccarani, Gianantonio Rosti, Fausto Castagnetti, Ibrahim Haznedaroglu, Kimmo Porkka, Elisabetta Abruzzese, Giuliana Alimena, Hans Ehrencrona, Henrik Hjorth-Hansen, Veli Kairisto, et al. Comparison of imatinib 400 mg and 800 mg daily in the front-line treatment of high-risk, philadelphia-positive chronic myeloid leukemia: a european leukemianet study. *Blood*, 113(19):4497–4504, 2009.

[161] Rüdiger Hehlmann, Michael Lauseker, Susanne Jung-Munkwitz, Armin Leitner, Martin C Müller, Nadine Pletsch, Ulrike Proetel, Claudia Haferlach, Brigitte Schlegelberger, Leopold Balleisen, et al. Tolerability-adapted imatinib 800 mg/d versus 400 mg/d versus 400 mg/d plus interferon-a in newly diagnosed chronic myeloid leukemia. *J Clin Oncol*, 29(12):1634–1642, 2011.

[162] Jorge E Cortes, Michele Baccarani, François Guilhot, Brian J Druker, Susan Branford, Dong-Wook Kim, Fabrizio Pane, Ricardo Pasquini, Stuart L Goldberg, Matt Kalaycio, et al. Phase iii, randomized, open-label study of daily imatinib mesylate 400 mg versus 800 mg in patients with newly diagnosed, previously untreated chronic myeloid leukemia in chronic phase using molecular end points: tyrosine kinase inhibitor optimization and selectivity study. *Journal of Clinical Oncology*, 28(3):424, 2010.

[163] Jianxiang Wang, Zhi-Xiang Shen, Giuseppe Saglio, Jie Jin, He Huang, Yu Hu,

Xin Du, Jianyong Li, Fanyi Meng, Huanling Zhu, et al. Phase 3 study of nilotinib vs imatinib in chinese patients with newly diagnosed chronic myeloid leukemia in chronic phase: Enestchina. *Blood*, 125(18):2771–2778, 2015.

[164] Elias Jabbour, Hagop Kantarjian, Susan O'Brien, Jenny Shan, Alfonso Quintas-Cardama, Stefan Faderl, Guillermo Garcia-Manero, Farhad Ravandi, Mary Beth Rios, and Jorge Cortes. The achievement of an early complete cytogenetic response is a major determinant for outcome in patients with early chronic phase chronic myeloid leukemia treated with tyrosine kinase inhibitors. *Blood*, 118(17):4541–4546, 2011.

[165] Sumio Watanabe and Manfred Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

[166] D Spiegelhalter, A Thomas, N Best, W Gilks, and D Lunn. Bugs: Bayesian inference using gibbs sampling. mrc biostatistics unit, cambridge, england (1994, 2003).

[167] Aki Vehtari and Andrew Gelman. Waic and cross-validation in stan. submitted. *City*, 2014.

[168] StanDevelopment Team et al. Rstan, the r interface to stan. *R package version*, 2.19.2(1), 2019.

[169] Haitao Chu, Lei Nie, Yong Chen, Yi Huang, and Wei Sun. Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. *Statistical methods in medical research*, 21(6):621–633, 2012.

[170] Aristidis K Nikoloulopoulos. A mixed effect model for bivariate meta-analysis of diagnostic test accuracy studies using a copula representation of the random effects distribution. *Statistics in medicine*, 34(29):3842–3865, 2015.

[171] David M Eddy, Vic Hasselblad, and Ross Shachter. A bayesian method for synthesizing evidence: The confidence profile method. *International Journal of Technology Assessment in Health Care*, 6(1):31–55, 1990.

[172] David M Eddy, Vic Hasselblad, and Ross Shachter. An introduction to a bayesian method for meta-analysis: the confidence profile method, 1990.

[173] David M Eddy, Victor Hasselblad, Ross D Shachter, et al. *Meta-analysis by the confidence profile method*. Academic Press, 1992.

[174] Kjell Benson and Arthur J Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342(25):1878–1886, 2000.

[175] Mark Zimmerman, Iwona Chelminski, and Michael A Posternak. Exclusion criteria used in antidepressant efficacy trials: consistency across studies and representativeness of samples included. *The Journal of nervous and mental disease*, 192(2):87–94, 2004.

[176] Martin Fortin, Jonathan Dionne, Geneviève Pinho, Julie Gignac, José Almirall, and Lise Lapointe. Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *The Annals of Family Medicine*, 4(2):104–108, 2006.

[177] Ross L Prentice, Robert D Langer, Marcia L Stefanick, Barbara V Howard, Mary Pettinger, Garnet L Anderson, David Barad, J David Curb, Jane Kotchen, Lewis Kuller, et al. Combined analysis of women's health initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *American journal of epidemiology*, 163(7):589–599, 2006.

[178] Joel B Greenhouse, Eloise E Kaizar, Kelly Kelleher, Howard Seltman, and William Gardner. Generalizing from clinical trial data: a case study. the risk of suicidality among pediatric antidepressant users. *Statistics in medicine*, 27(11):1801–1813, 2008.

[179] Pablo E Verde and Christian Ohmann. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods*, 6(1):45–62, 2015.

[180] David M Eddy. The confidence profile method: a bayesian method for assessing health technologies. *Operations Research*, 37(2):210–228, 1989.

[181] Joseph W Segura, Glenn M Preminger, Dean G Assimos, Stephen P Dretler, Robert I Kahn, James E Lingeman, Joseph N Macaluso, and David L McCullough. Nephrolithiasis clinical guidelines panel summary report on the management of staghorn calculi. *The Journal of urology*, 151(6):1648–1651, 1994.

[182] Joseph W Segura, Glenn M Preminger, Dean G Assimos, Stephen P Dretler, Robert I Kahn, James E Lingeman, and Joseph N Macaluso. Ureteral stones clinical guidelines panel summary report on the management of ureteral calculi. *The Journal of urology*, 158(5):1915–1921, 1997.

[183] M Craig Hall, Sam S Chang, Guido Dalbagni, Raj Som Pruthi, John Derek Seigne, Eila Curlee Skinner, J Stuart Wolf, and Paul F Schellhammer. Guideline for the management of nonmuscle invasive bladder cancer (stages ta, t1, and tis): 2007 update. *The Journal of urology*, 178(6):2314–2330, 2007.

[184] Joseph A Smith, Richard F Labasky, Abraham TK Cockett, John A Fracchia, James E Montie, and Randall G Rowland. Bladder cancer clinical guidelines panel summary report on the management of nonmuscle invasive bladder cancer. *The Journal of urology*, 162(5):1697–1701, 1999.

[185] Judith Droitcour, George Silberman, and Eleanor Chelimsky. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care*, 9(3):440–449, 1993.

[186] Eleanor Chelimsky, George Silberman, and Judith Droitcour. Cross design synthesis. *The Lancet*, 341(8843):498, 1993.

[187] Colin B Begg. Cross design synthesis: A new strategy for medical effectiveness research. *Statistics in Medicine*, 11(12):1627–1628, 1992.

[188] Colin B Begg and Louise Pilote. A model for incorporating historical controls into a meta-analysis. *Biometrics*, pages 899–906, 1991.

[189] Pablo E Verde, Christian Ohmann, Stephan Morbach, and Andrea Icks. Bayesian evidence synthesis for exploring generalizability of treatment effects:

a case study of combining randomized and non-randomized results in diabetes. *Statistics in medicine*, 35(10):1654–1675, 2016.

[190] Pablo Emilio Verde. The hierarchical metaregression approach and learning from clinical evidence. *Biometrical Journal*, 61(3):535–557, 2019.

[191] Pablo Emilio Verde. A bias-corrected meta-analysis model for combining, studies of different types and quality. *Biometrical Journal*, 2020.

[192] Eduardo Díaz-Rubio, Auxiliadora Gómez-España, Bartomeu Massutí, Javier Sastre, Albert Abad, Manuel Valladares, Fernando Rivera, Maria J Safont, Purificación Martínez De Prado, Manuel Gallén, et al. First-line xelox plus bevacizumab followed by xelox plus bevacizumab or single-agent bevacizumab as maintenance therapy in patients with metastatic colorectal cancer: the phase iii macro ttd study. *The oncologist*, 17(1):15, 2012.

[193] Zhong-Zhen Guan, Jian-Ming Xu, Rong-Cheng Luo, Feng-Yi Feng, Li-Wei Wang, Lin Shen, Shi-Ying Yu, Yi Ba, Jun Liang, Dong Wang, et al. Efficacy and safety of bevacizumab plus chemotherapy in chinese patients with metastatic colorectal cancer: a randomized phase iii artist trial. *Chinese journal of cancer*, 30(10):682, 2011.

[194] J Randolph Hecht, Tanja Trarbach, John D Hainsworth, Pierre Major, Elke Jäger, Robert A Wolff, Katherine Lloyd-Salvant, György Bodoky, Kelly Pendergrass, William Berg, et al. Randomized, placebo-controlled, phase iii study of first-line oxaliplatin-based chemotherapy plus ptk787/zk 222584, an oral vascular endothelial growth factor receptor inhibitor, in patients with metastatic colorectal adenocarcinoma. *Journal of clinical oncology*, 29(15):1997–2003, 2011.

[195] J Randolph Hecht, Edith Mitchell, Tarek Chidiac, Carroll Scroggin, Christopher Hagenstad, David Spigel, John Marshall, Allen Cohn, David McCollum, Philip Stella, et al. A randomized phase iiib trial of chemotherapy, bevacizumab, and panitumumab compared with chemotherapy and bevacizumab alone for metastatic colorectal cancer. *J Clin Oncol*, 27(5):672–680, 2009.

[196] Paulo M Hoff, Andreas Hochhaus, Bernhard C Pestalozzi, Niall C Tebbutt, Jin

Li, Tae Won Kim, Krassimir D Koynov, Galina Kurteva, Tamás Pintér, Ying Cheng, et al. Cediranib plus folfox/capox versus placebo plus folfox/capox in patients with previously untreated metastatic colorectal cancer: a randomized, double-blind, phase iii study (horizon ii). *Journal of clinical oncology*, 30(29):3596–3603, 2012.

[197] Fairooz F Kabbinavar, Joseph Schulz, Michael McCleod, Taral Patel, John T Hamm, J Randolph Hecht, Robert Mass, Brent Perrou, Betty Nelson, and William F Novotny. Addition of bevacizumab to bolus fluorouracil and leucovorin in first-line metastatic colorectal cancer: results of a randomized phase ii trial. *Journal of Clinical Oncology*, 23(16):3697–3705, 2005.

[198] Hans-Joachim Schmoll, David Cunningham, Alberto Sobrero, Christos S Karapetis, Philippe Rougier, Sheryl L Koski, Ilona Kocakova, Igor Bondarenko, György Bodoky, Paul Mainwaring, et al. Cediranib with mfolfox6 versus bevacizumab with mfolfox6 as first-line treatment for patients with advanced colorectal cancer: a double-blind, randomized phase iii study (horizon iii). *Journal of clinical oncology*, 30(29):3588–3595, 2012.

[199] John Souglakos, N Ziras, S Kakolyris, I Boukovinas, N Kentepozidis, P Makrantonakis, S Xynogalos, Ch Christophyllakis, Ch Kouroussis, L Vamvakas, et al. Randomised phase-ii trial of capiri (capecitabine, irinotecan) plus bevacizumab vs folfiri (folinic acid, 5-fluorouracil, irinotecan) plus bevacizumab as first-line treatment of patients with unresectable/metastatic colorectal cancer (mcrc). *British journal of cancer*, 106(3):453–459, 2012.

[200] Niall C Tebbutt, Kate Wilson, Val J Gebski, Michelle M Cummins, Diana Zannino, Guy A Van Hazel, Bridget Robinson, Adam Broad, Vinod Ganju, Stephen P Ackland, et al. Capecitabine, bevacizumab, and mitomycin in first-line treatment of metastatic colorectal cancer: results of the australasian gastrointestinal trials group randomized phase iii max study. *Journal of clinical oncology*, 28(19):3191–3198, 2010.

[201] Eric Van Cutsem, Josep Tabernero, Radek Lakomy, Hans Prenen, Jana Prausová, Teresa Macarulla, Paul Ruff, Guy A Van Hazel, Vladimir Moiseyenko, David Ferry, et al. Addition of aflibercept to fluorouracil,

leucovorin, and irinotecan improves survival in a phase iii randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. *J Clin Oncol*, 30(28):3499–3506, 2012.

[202] Eric Van Cutsem, Emilio Bajetta, Juan Valle, Claus-Henning Köhne, J Randolph Hecht, Malcolm Moore, Colin Germond, William Berg, Bee-Lian Chen, Tarja Jalava, et al. Randomized, placebo-controlled, phase iii study of oxaliplatin, fluorouracil, and leucovorin with or without ptk787/zk 222584 in patients with previously treated metastatic colorectal adenocarcinoma. *Journal of clinical oncology*, 29(15):2004–2010, 2011.

[203] Johanna C Bendell, Tanios S Bekaii-Saab, Allen L Cohn, Herbert I Hurwitz, Mark Kozloff, Haluk Tezcan, Nancy Roach, Yong Mun, Susan Fish, E Dawn Flick, et al. Treatment patterns and clinical outcomes in patients with metastatic colorectal cancer initially treated with folfox–bevacizumab or folfiri–bevacizumab: results from aries, a bevacizumab observational cohort study. *The oncologist*, 17(12):1486, 2012.

[204] HI Hurwitz, TS Bekaii-Saab, JC Bendell, AL Cohn, M Kozloff, N Roach, Y Mun, S Fish, ED Flick, A Grothey, et al. Safety and effectiveness of bevacizumab treatment for metastatic colorectal cancer: final results from the avastin® registry–investigation of effectiveness and safety (aries) observational cohort study. *Clinical Oncology*, 26(6):323–332, 2014.

[205] Eric Van Cutsem, F Rivera, S Berry, A Kretzschmar, M Michael, M DiBartolomeo, M-A Mazier, J-L Canon, V Georgoulias, Marc Peeters, et al. Safety and efficacy of first-line bevacizumab with folfox, xelox, folfiri and fluoropyrimidines in metastatic colorectal cancer: the beat study. *Annals of Oncology*, 20(11):1842–1847, 2009.

[206] Jaafar Bennouna, Jean-Marc Phelip, Thierry André, Bernard Asselain, Sébastien Koné, and Michel Ducreux. Observational cohort study of patients with metastatic colorectal cancer initiating chemotherapy in combination with bevacizumab (concert). *Clinical Colorectal Cancer*, 16(2):129–140, 2017.

[207] Tomas Buchler, Tomas Pavlik, Bohuslav Melichar, Zbynek Bortlicek, Zuzana Usiakova, Ladislav Dusek, Igor Kiss, Milan Kohoutek, Vera Benesova, Rostislav

Vyzula, et al. Bevacizumab with 5-fluorouracil, leucovorin, and oxaliplatin versus bevacizumab with capecitabine and oxaliplatin for metastatic colorectal carcinoma: results of a large registry-based cohort analysis. *BMC cancer*, 14(1):323, 2014.

[208] Janja Ocvirk, Martina Rebersek, and Marko Boc. Bevacizumab in first-line therapy of metastatic colorectal cancer: a retrospective comparison of folfiri and xeliri. *Anticancer research*, 31(5):1777–1782, 2011.

[209] Toshikazu Moriwaki, Hideaki Bando, Atsuo Takashima, Kentaro Yamazaki, Taito Esaki, Keishi Yamashita, Mutsumi Fukunaga, Yasuhiro Miyake, Kenji Katsumata, Satoshi Kato, et al. Bevacizumab in combination with irinotecan, 5-fluorouracil, and leucovorin (folfiri) in patients with metastatic colorectal cancer who were previously treated with oxaliplatin-containing regimens: a multicenter observational cohort study (tctg 2nd-bv study). *Medical Oncology*, 29(4):2842–2848, 2012.

[210] Masahito Kotaka, Fusao Ikeda, Masaki Tsujie, Shinichi Yoshioka, Yoshihiko Nakamoto, Takaaki Ishii, Takahisa Kyogoku, Takeshi Kato, Akihito Tsuji, and Michiya Kobayashi. Observational cohort study focused on treatment continuity of patients administered xelox plus bevacizumab for previously untreated metastatic colorectal cancer. *OncoTargets and therapy*, 9:4113, 2016.

[211] Andreas Hochhaus, Richard A Larson, François Guilhot, Jerald P Radich, Susan Branford, Timothy P Hughes, Michele Baccarani, Michael W Deininger, Francisco Cervantes, Satoko Fujihara, et al. Long-term outcomes of imatinib treatment for chronic myeloid leukemia. *New England Journal of Medicine*, 376(10):917–927, 2017.

[212] Jorge E Cortes, Andreas Hochhaus, Dong-Wook Kim, Neil P Shah, Jiri Mayer, Philip Rowlings, Hirohisa Nakamae, M Brigid Bradley-Garelik, Hesham Mohamed, Hagop M Kantarjian, et al. Four-year (yr) follow-up of patients (pts) with newly diagnosed chronic myeloid leukemia in chronic phase (cml-cp) receiving dasatinib or imatinib: efficacy based on early response, 2013.

[213] A Hochhaus, Giuseppe Saglio, TP Hughes, RA Larson, DW Kim, S Issaragrisil,

PD Le Coutre, G Etienne, PE Dorlhiac-Llacer, RE Clark, et al. Long-term benefits and risks of frontline nilotinib vs imatinib for chronic myeloid leukemia in chronic phase: 5-year update of the randomized enestnd trial. *Leukemia*, 30(5):1044–1054, 2016.

[214] Masahiro Kizaki, Naoto Takahashi, Noriyoshi Iriyama, Shinichiro Okamoto, Takaaki Ono, Noriko Usui, Koiti Inokuchi, Chiaki Nakaseko, Mineo Kurokawa, Masahiko Sumi, et al. Efficacy and safety of tyrosine kinase inhibitors for newly diagnosed chronic-phase chronic myeloid leukemia over a 5-year period: results from the japanese registry obtained by the new target system. *International journal of hematology*, 109(4):426–439, 2019.

[215] Verena Sophia Hoffmann, M Baccarani, Joerg Hasford, F Castagnetti, F Di Raimondo, LF Casado, A Turkina, D Zackova, G Ossenkoppele, A Zaritskey, et al. Treatment and outcome of 2904 cml patients from the eutos population-based registry. *Leukemia*, 31(3):593–601, 2017.

[216] Ian R White, Rebecca M Turner, Amalia Karahalios, and Georgia Salanti. A comparison of arm-based and contrast-based models for network meta-analysis. *Statistics in medicine*, 38(27):5197–5213, 2019.

[217] Stephen Senn. Hans van houwelingen and the art of summing up. *Biometrical Journal*, 52(1):85–94, 2010.

[218] Sofia Dias and AE Ades. Absolute or relative effects? arm-based synthesis of trial data. *Research synthesis methods*, 7(1):23, 2016.

[219] Felix A Achana, Nicola J Cooper, Sofia Dias, Guobing Lu, Stephen JC Rice, Denise Kendrick, and Alex J Sutton. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Statistics in medicine*, 32(5):752–771, 2013.

[220] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[221] Ralph B D'Agostino Jr. Propensity score methods for bias reduction in the

comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17(19):2265–2281, 1998.

[222] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[223] James E Signorovitch, Vanja Sikirica, M Haim Erder, Jipan Xie, Mei Lu, Paul S Hodgkins, Keith A Betts, and Eric Q Wu. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value in Health*, 15(6):940–947, 2012.

[224] David M Phillippo, Anthony E Ades, Sofia Dias, Stephen Palmer, Keith R Abrams, and Nicky J Welton. Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making*, 38(2):200–211, 2018.

[225] Paul R Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

[226] Michael R Jaff, Teresa Nelson, Nicole Ferko, Melissa Martinson, Louise H Anderson, and Sarah Hollmann. Endovascular interventions for femoropopliteal peripheral artery disease: a network meta-analysis of current technologies. *Journal of Vascular and Interventional Radiology*, 28(12):1617–1627, 2017.

[227] Susanne Schmitz, Áine Maguire, James Morris, Kai Ruggeri, Elisa Haller, Isla Kuhn, Joy Leahy, Natalia Homer, Ayesha Khan, Jack Bowden, et al. 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, 2018.

Appendix A

Appendix

A.1 Centred and non-centred parameterisations

This sections includes the codes of the two "normal" models used to illustrate the difference between centred and non-centred parameterisations in Figure 2.3.

Normal model with centred parameterisation

```
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] s;}

parameters {
  real tau;
  real m;
  vector[N] mu;}

model {
  m ~normal(0,5);
  tau~normal(0,2.5);
  mu ~ normal(m,exp(tau)/2);//centred parameterisation
  y ~ normal(mu, s);}
```

Normal model with non-centred parameterisation

```
data {
 int<lower=0> N;
 vector[N] y;
 vector[N] s;}

parameters {
 real tau;
 real m;
 vector[N] z;}

transformed parameters {
  vector[N] mu;
  for (i in 1:N){
  mu[i] = (exp(tau)/2)*z[i]+m;}}//non-centred parameterisation

model {
  z~std_normal();
  m ~normal(0,5);
  tau~normal(0,2.5);
  y ~ normal(mu, s);}
```

Appendix B

Appendix

B.1 Results for the estimates of the slopes in each treatment class separately

1st scenario

Table B.1: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the first scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$	Probability of strong association$_j$
subgroup analysis					
1^{st} treatment class	0.95	0.07	0.08		0.80
2^{nd} treatment class	0.95	0.07	0.10		0.79
3^{rd} treatment class	0.96	0.07	0.10		0.81
4^{th} treatment class	0.97	0.08	0.10		0.81
5^{th} treatment class	0.94	0.08	0.10		0.81
F-EX model					
1^{st} treatment class	0.94	0.06	0.07	0.77	0.85
2^{nd} treatment class	0.97	0.05	0.06	0.71	0.84
3^{rd} treatment class	0.98	0.05	0.06	0.70	0.85
4^{th} treatment class	0.97	0.05	0.07	0.69	0.84

5th treatment class	0.90	0.07	0.09	0.72	0.83
P-EX					
1st treatment class	0.94	0.06	0.07	0.78	0.85
2st treatment class	0.97	0.05	0.06	0.72	0.85
3st treatment class	0.98	0.05	0.06	0.70	0.84
4st treatment class	0.97	0.05	0.07	0.70	0.84
5st treatment class	0.91	0.07	0.09	0.72	0.84

2nd scenario

Table B.2: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the second scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$	Probability of strong association$_j$
subgroup analysis					
1^{st} treatment class	0.98	0.10	0.13		0.68
2^{nd} treatment class	0.98	0.11	0.14		0.66
3^{rd} treatment class	0.98	0.11	0.15		0.70
4^{th} treatment class	0.98	0.11	0.15		0.75
5^{th} treatment class	0.98	0.11	0.16		0.78
F-EX model					
1^{st} treatment class	0.97	0.07	0.09	0.64	0.90
2^{nd} treatment class	0.98	0.06	0.08	0.60	0.90
3^{rd} treatment class	0.98	0.06	0.08	0.59	0.90
4^{th} treatment class	0.98	0.07	0.08	0.59	0.90
5^{th} treatment class	0.94	0.09	0.10	0.59	0.90
P-EX					
1^{st} treatment class	0.98	0.07	0.09	0.65	0.90
2^{st} treatment class	0.98	0.06	0.08	0.61	0.90
3^{st} treatment class	0.99	0.06	0.08	0.60	0.90
4^{st} treatment class	0.98	0.07	0.08	0.59	0.90
5^{st} treatment class	0.94	0.09	0.11	0.60	0.90

3rd scenario

Table B.3: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the third scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$	Probability of strong association$_j$
subgroup analysis					
1^{st} treatment class	1.00	0.19	0.27		0.03
2^{nd} treatment class	0.99	0.10	0.14		0.71
3^{rd} treatment class	0.99	0.14	0.18		0.52
4^{th} treatment class	0.98	0.10	0.13		0.81
5^{th} treatment class	0.97	0.13	0.19		0.72
F-EX model					
1^{st} treatment class	0.99	0.08	0.10	0.26	0.84
2^{nd} treatment class	0.99	0.06	0.08	0.63	0.90
3^{rd} treatment class	0.99	0.07	0.08	0.48	0.90
4^{th} treatment class	0.98	0.08	0.08	0.69	0.88
5^{th} treatment class	0.97	0.08	0.10	0.55	0.91
P-EX					
1^{st} treatment class	0.99	0.08	0.10	0.27	0.83
2^{st} treatment class	0.99	0.06	0.08	0.64	0.90
3^{st} treatment class	0.99	0.06	0.08	0.49	0.90
4^{st} treatment class	0.98	0.07	0.09	0.69	0.88
5^{st} treatment class	0.97	0.08	0.10	0.56	0.92

4th scenario

Table B.4: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the forth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$	Probability of strong association$_j$
subgroup analysis					
1^{st} treatment class	0.96	0.08	0.10		0.85
2^{nd} treatment class	0.93	0.09	0.11		0.90
3^{rd} treatment class	0.94	0.09	0.11		0.91
4^{th} treatment class	0.95	0.09	0.11		0.90
5^{th} treatment class	0.93	0.10	0.12		0.90
F-EX model					
1^{st} treatment class	0.96	0.07	0.09	0.93	0.88
2^{nd} treatment class	0.94	0.08	0.10	0.89	0.91
3^{rd} treatment class	0.93	0.08	0.10	0.89	0.93
4^{th} treatment class	0.94	0.08	0.10	0.90	0.92
5^{th} treatment class	0.93	0.09	0.11	0.90	0.91
P-EX					
1^{st} treatment class	0.96	0.07	0.09	0.92	0.89
2^{st} treatment class	0.96	0.07	0.09	0.84	0.92
3^{st} treatment class	0.95	0.07	0.09	0.83	0.93
4^{st} treatment class	0.94	0.07	0.09	0.83	0.92
5^{st} treatment class	0.93	0.09	0.10	0.84	0.91

5th scenario

Table B.5: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the fifth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$	Probability of strong association$_j$
subgroup analysis					
1^{st} treatment class	0.98	0.13	0.16		0.85
2^{nd} treatment class	0.96	0.13	0.17		0.90
3^{rd} treatment class	0.96	0.13	0.17		0.91
4^{th} treatment class	0.98	0.14	0.16		0.90
5^{th} treatment class	0.97	0.14	0.17		0.90
F-EX model					
1^{st} treatment class	0.96	0.14	0.19	1.06	0.90
2^{nd} treatment class	0.96	0.11	0.14	0.80	0.94
3^{rd} treatment class	0.96	0.11	0.14	0.80	0.92
4^{th} treatment class	0.98	0.11	0.14	0.81	0.92
5^{th} treatment class	0.96	0.12	0.15	0.80	0.91
P-EX					
1^{st} treatment class	0.98	0.10	0.14	0.90	0.88
2^{st} treatment class	0.98	0.09	0.11	0.76	0.94
3^{st} treatment class	0.97	0.09	0.11	0.74	0.93
4^{st} treatment class	0.97	0.10	0.12	0.75	0.92
5^{st} treatment class	0.95	0.11	0.13	0.74	0.92

6th scenario

Table B.6: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the sixth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$	Probability of strong association$_j$
subgroup analysis					
1^{st} treatment class	1.00	0.22	0.31		0.07
2^{nd} treatment class	0.97	0.12	0.15		0.90
3^{rd} treatment class	0.98	0.15	0.20		0.86
4^{th} treatment class	0.97	0.11	0.14		0.90
5^{th} treatment class	0.98	0.14	0.20		0.90
F-EX model					
1^{st} treatment class	0.91	0.46	0.51	0.58	0.71
2^{nd} treatment class	0.98	0.08	0.11	0.76	0.93
3^{rd} treatment class	0.98	0.10	0.13	0.64	0.92
4^{th} treatment class	0.95	0.10	0.12	0.82	0.93
5^{th} treatment class	0.96	0.12	0.15	0.70	0.92
P-EX					
1^{st} treatment class	0.96	0.34	0.40	0.62	0.68
2^{st} treatment class	0.98	0.08	0.10	0.75	0.93
3^{st} treatment class	0.99	0.09	0.13	0.62	0.92
4^{st} treatment class	0.95	0.09	0.12	0.81	0.92
5^{st} treatment class	0.96	0.12	0.14	0.70	0.92

7th scenario

Table B.7: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the seventh scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
subgroup analysis				
1^{st} treatment class	0.96	0.07	0.09	
2^{nd} treatment class	0.96	0.14	0.18	
3^{rd} treatment class	0.96	0.08	0.11	
4^{th} treatment class	0.95	0.16	0.20	
5^{th} treatment class	0.94	0.09	0.12	
F-EX model				
1^{st} treatment class	0.95	0.07	0.08	0.91
2^{nd} treatment class	0.98	0.08	0.11	0.71
3^{rd} treatment class	0.97	0.06	0.08	0.80
4^{th} treatment class	0.97	0.10	0.13	0.68
5^{th} treatment class	0.89	0.10	0.13	0.86
P-EX				
1^{st} treatment class	0.95	0.07	0.08	0.91
2^{st} treatment class	0.98	0.08	0.11	0.71
3^{st} treatment class	0.97	0.06	0.08	0.81
4^{st} treatment class	0.97	0.10	0.13	0.69
5^{st} treatment class	0.88	0.10	0.12	0.86

8th scenario

Table B.8: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the eighth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
subgroup analysis				
1^{st} treatment class	0.98	0.11	0.14	
2^{nd} treatment class	0.96	0.21	0.27	
3^{rd} treatment class	0.97	0.13	0.17	
4^{th} treatment class	0.96	0.24	0.31	
5^{th} treatment class	0.97	0.13	0.18	
F-EX model				
1^{st} treatment class	0.97	0.09	0.11	0.78
2^{nd} treatment class	0.98	0.11	0.14	0.59
3^{rd} treatment class	0.98	0.08	0.11	0.68
4^{th} treatment class	0.98	0.13	0.16	0.56
5^{th} treatment class	0.90	0.13	0.11	0.71
P-EX				
1^{st} treatment class	0.97	0.09	0.11	0.78
2^{st} treatment class	0.98	0.11	0.14	0.60
3^{st} treatment class	0.98	0.08	0.11	0.68
4^{st} treatment class	0.98	0.13	0.16	0.57
5^{st} treatment class	0.90	0.13	0.11	0.72

9th scenario

Table B.9: Performance measures of $\hat{\lambda}_{1j}$ averaged over 1000 replications in the ninth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
subgroup analysis				
1^{st} treatment class	1.00	0.20	0.14	
2^{nd} treatment class	0.96	0.21	0.27	
3^{rd} treatment class	0.99	0.16	0.17	
4^{th} treatment class	0.96	0.21	0.31	
5^{th} treatment class	0.98	0.15	0.18	
F-EX model				
1^{st} treatment class	0.98	0.13	0.15	0.34
2^{nd} treatment class	0.98	0.12	0.15	0.61
3^{rd} treatment class	0.99	0.09	0.11	0.56
4^{th} treatment class	0.98	0.12	0.16	0.63
5^{th} treatment class	0.93	0.13	0.15	0.56
P-EX				
1^{st} treatment class	0.99	0.13	0.15	0.35
2^{st} treatment class	0.98	0.12	0.15	0.62
3^{st} treatment class	0.99	0.09	0.11	0.57
4^{st} treatment class	0.97	0.13	0.16	0.64
5^{st} treatment class	0.94	0.13	0.15	0.67

B.2 Results for the predictions of the true treatment effects in each treatment class separately

1st scenario

Table B.10: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the first scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.95	0.09	0.10	
2^{nd} treatment class	0.95	0.09	0.10	
3^{rd} treatment class	0.95	0.09	0.11	
4^{th} treatment class	0.95	0.09	0.11	
5^{th} treatment class	0.95	0.09	0.11	
F-EX model				
1^{st} treatment class	0.95	0.08	0.10	0.96
2^{nd} treatment class	0.95	0.08	0.10	0.93
3^{rd} treatment class	0.95	0.08	0.10	0.92
4^{th} treatment class	0.95	0.08	0.10	0.92
5^{th} treatment class	0.95	0.08	0.10	0.93
P-EX model				
1^{st} treatment class	0.95	0.08	0.10	0.95
2^{nd} treatment class	0.95	0.08	0.10	0.93
3^{rd} treatment class	0.95	0.08	0.10	0.93
4^{th} treatment class	0.95	0.08	0.10	0.92
5^{th} treatment class	0.95	0.08	0.10	0.93

2nd scenario

Table B.11: Performance measures of $\hat{\hat{\delta}}_{2ij}$ averaged over 1000 replications and the number of studies in the second scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.98	0.10	0.12	
2^{nd} treatment class	0.98	0.11	0.12	
3^{rd} treatment class	0.98	0.11	0.13	
4^{th} treatment class	0.98	0.11	0.13	
5^{th} treatment class	0.98	0.11	0.13	
F-EX model				
1^{st} treatment class	0.98	0.08	0.10	0.81
2^{nd} treatment class	0.98	0.08	0.10	0.79
3^{rd} treatment class	0.98	0.08	0.10	0.79
4^{th} treatment class	0.98	0.08	0.10	0.79
5^{th} treatment class	0.98	0.09	0.11	0.80
P-EX model				
1^{st} treatment class	0.98	0.08	0.10	0.82
2^{nd} treatment class	0.98	0.08	0.10	0.80
3^{rd} treatment class	0.98	0.08	0.10	0.79
4^{th} treatment class	0.98	0.08	0.10	0.79
5^{th} treatment class	0.98	0.09	0.11	0.80

3rd scenario

Table B.12: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the third scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	1.00	0.18	0.38	
2^{nd} treatment class	0.99	0.10	0.12	
3^{rd} treatment class	0.99	0.11	0.14	
4^{th} treatment class	0.98	0.09	0.12	
5^{th} treatment class	1.00	0.10	0.13	
F-EX model				
1^{st} treatment class	1.00	0.08	0.11	0.32
2^{nd} treatment class	0.99	0.09	0.10	0.81
3^{rd} treatment class	0.99	0.08	0.11	0.62
4^{th} treatment class	0.97	0.08	0.10	0.88
5^{th} treatment class	0.99	0.09	0.11	0.74
P-EX model				
1^{st} treatment class	1.00	0.08	0.11	0.34
2^{nd} treatment class	0.99	0.09	0.11	0.81
3^{rd} treatment class	0.99	0.09	0.11	0.62
4^{th} treatment class	0.97	0.09	0.11	0.88
5^{th} treatment class	0.99	0.09	0.11	0.74

4th scenario

Table B.13: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the forth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.95	0.09	0.11	
2^{nd} treatment class	0.95	0.15	0.19	
3^{rd} treatment class	0.95	0.15	0.19	
4^{th} treatment class	0.95	0.15	0.19	
5^{th} treatment class	0.95	0.16	0.20	
F-EX model				
1^{st} treatment class	0.95	0.08	0.11	0.98
2^{nd} treatment class	0.95	0.14	0.18	0.97
3^{rd} treatment class	0.95	0.14	0.18	0.97
4^{th} treatment class	0.95	0.15	0.19	0.97
5^{th} treatment class	0.95	0.15	0.19	0.97
P-EX model				
1^{st} treatment class	0.95	0.08	0.10	0.96
2^{nd} treatment class	0.96	0.14	0.18	0.96
3^{rd} treatment class	0.96	0.14	0.18	0.96
4^{th} treatment class	0.96	0.15	0.18	0.95
5^{th} treatment class	0.95	0.15	0.19	0.95

5th scenario

Table B.14: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the fifth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.99	0.10	0.13	
2^{nd} treatment class	0.99	0.16	0.20	
3^{rd} treatment class	0.99	0.17	0.21	
4^{th} treatment class	0.99	0.17	0.21	
5^{th} treatment class	0.99	0.17	0.22	
F-EX model				
1^{st} treatment class	0.98	0.11	0.15	1.08
2^{nd} treatment class	0.99	0.15	0.19	0.88
3^{rd} treatment class	0.98	0.15	0.19	0.88
4^{th} treatment class	0.99	0.16	0.20	0.88
5^{th} treatment class	0.98	0.16	0.20	0.88
P-EX model				
1^{st} treatment class	0.98	0.10	0.10	0.93
2^{nd} treatment class	0.99	0.14	0.18	0.86
3^{rd} treatment class	0.98	0.14	0.18	0.85
4^{th} treatment class	0.99	0.15	0.18	0.85
5^{th} treatment class	0.98	0.16	0.19	0.85

6th scenario

Table B.15: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the sixth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	1.00	0.18	0.30	
2^{nd} treatment class	0.99	0.17	0.20	
3^{rd} treatment class	0.99	0.18	0.23	
4^{th} treatment class	0.98	0.17	0.20	
5^{th} treatment class	0.99	0.18	0.23	
F-EX model				
1^{st} treatment class	0.99	0.25	0.31	0.62
2^{nd} treatment class	0.99	0.15	0.19	0.88
3^{rd} treatment class	0.99	0.16	0.20	0.73
4^{th} treatment class	0.98	0.15	0.19	0.93
5^{th} treatment class	0.99	0.16	0.20	0.82
P-EX model				
1^{st} treatment class	1.00	0.15	0.20	0.56
2^{nd} treatment class	0.98	0.15	0.18	0.87
3^{rd} treatment class	0.99	0.15	0.19	0.71
4^{th} treatment class	0.98	0.15	0.19	0.91
5^{th} treatment class	0.99	0.16	0.19	0.80

7th scenario

Table B.16: Performance measures of $\hat{\hat{\delta}}_{2ij}$ averaged over 1000 replications and the number of studies in the seventh scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.95	0.08	0.11	
2^{nd} treatment class	0.95	0.26	0.33	
3^{rd} treatment class	0.95	0.09	0.12	
4^{th} treatment class	0.95	0.27	0.34	
5^{th} treatment class	0.96	0.11	0.13	
F-EX model				
1^{st} treatment class	0.95	0.08	0.11	0.99
2^{nd} treatment class	0.95	0.25	0.32	0.94
3^{rd} treatment class	0.95	0.09	0.11	0.95
4^{th} treatment class	0.96	0.25	0.32	0.94
5^{th} treatment class	0.96	0.10	0.13	0.96
P-EX model				
1^{st} treatment class	0.95	0.08	0.11	0.98
2^{nd} treatment class	0.95	0.25	0.32	0.95
3^{rd} treatment class	0.95	0.09	0.11	0.95
4^{th} treatment class	0.96	0.25	0.32	0.94
5^{th} treatment class	0.96	0.10	0.13	0.96

8th scenario

Table B.17: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the eighth scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.98	0.10	0.12	
2^{nd} treatment class	0.97	0.29	0.37	
3^{rd} treatment class	0.98	0.10	0.13	
4^{th} treatment class	0.96	0.30	0.38	
5^{th} treatment class	0.98	0.12	0.15	
F-EX model				
1^{st} treatment class	0.98	0.09	0.12	0.89
2^{nd} treatment class	0.96	0.25	0.32	0.84
3^{rd} treatment class	0.98	0.09	0.12	0.84
4^{th} treatment class	0.96	0.26	0.32	0.83
5^{th} treatment class	0.98	0.10	0.13	0.86
P-EX model				
1^{st} treatment class	0.98	0.09	0.12	0.89
2^{nd} treatment class	0.96	0.25	0.32	0.84
3^{rd} treatment class	0.98	0.09	0.12	0.84
4^{th} treatment class	0.96	0.26	0.32	0.83
5^{th} treatment class	0.98	0.10	0.13	0.86

9th scenario

Table B.18: Performance measures of $\hat{\delta}_{2ij}$ averaged over 1000 replications and the number of studies in the second scenario

Methods	Coverage probability$_j$	Absolute Bias	RMSE$_j$	Width Ratio$_j$
Subgroup analysis				
1^{st} treatment class	0.98	0.19	0.34	
2^{nd} treatment class	0.97	0.29	0.36	
3^{rd} treatment class	0.98	0.12	0.15	
4^{th} treatment class	0.96	0.28	0.35	
5^{th} treatment class	0.98	0.12	0.16	
F-EX model				
1^{st} treatment class	1.00	0.11	0.13	0.37
2^{nd} treatment class	0.96	0.25	0.32	0.85
3^{rd} treatment class	0.99	0.10	0.12	0.67
4^{th} treatment class	0.96	0.26	0.32	0.89
5^{th} treatment class	0.99	0.11	0.14	0.80
P-EX model				
1^{st} treatment class	1.00	0.11	0.13	0.40
2^{nd} treatment class	0.96	0.25	0.32	0.85
3^{rd} treatment class	0.99	0.10	0.12	0.67
4^{th} treatment class	0.96	0.26	0.32	0.89
5^{th} treatment class	0.99	0.11	0.14	0.80

B.3 Bootstrap method

The following function provided by Prof. Sylwia Bujkiewicz was used to calculate the within-study correlations in section 4.4.

```
function(data,Nb){
  s<-nrow(data) #number of observation in the data
  y1<-y2<-y3<-array(0,Nb)
  for (i in 1:Nb){
    sam<-sample(s, replace=T)
    boot.RND<-data$RND[sam]
    boot.TTPFS<-data$TTPFS[sam]
    boot.CSPFS<-data$CSPFS[sam]
    boot.TTDIED<-data$TTDIED[sam]
    boot.CSDIED<-data$CSDIED[sam]
    boot.response<-data$response[sam]
    lmodres<-glm(boot.response~boot.RND, family="binomial")
    y1[i]<-coef(lmodres)[2]
    smodPFS<-coxph(Surv(boot.TTPFS,boot.CSPFS)~boot.RND)
    y2[i]<-coef(smodPFS)[1]
    smodOS<-coxph(Surv(boot.TTDIED,boot.CSDIED)~boot.RND)
    y3[i]<-coef(smodOS)[1]}
  rho<-cor(data.frame(y1,y2,y3),use = "pairwise",method= "pearson")
  colnames(rho) <- rownames(rho) <- c("ORR","PFS","OS")
   #the correlations between logHR_PFS and logHR_OS logOR_response)
return(list(rho=rho))}
```

B.4 Implementation of F-EX model in BUGS

```
model{
#within study precision matrix
for (i in 1:ns) {
Prec_w[i,1:2,1:2] <- inverse(Sigma[i,1:2,1:2])
#covariance matrix for the i-th study
Sigma[i,1,1]<-pow(se[i,1],2)
Sigma[i,2,2]<-pow(se[i,2],2)
Sigma[i,1,2]<-sqrt(Sigma[i,1,1])*sqrt(Sigma[i,2,2])*rho_w[i]
Sigma[i,2,1]<-sqrt(Sigma[i,1,1])*sqrt(Sigma[i,2,2])*rho_w[i]
}
# Random effects model
for (i in 1:ns) {
y[i,1:2]~dmnorm(mu[i,1:2], Prec_w[i,1:2,1:2])
#  product normal formulation for the between study part:
mu[i,1]~dnorm(0,1.0E-3)
mu[i,2]~dnorm(eta[i,class[i]],prec_fin[class[i]])
for (j in 1:nclass) {
eta[i,j]<-lambda0[j]+lambda1[j]*mu[i,1]
}
}
for (j in 1:nclass) {
lambda0[j]~dnorm(beta1,pr1)
lambda1[j]~dnorm(beta2,pr2)
gam_fin[j]~dnorm(0,2)I(0,)
gam_fin.sq[j]<-gam_fin[j]*gam_fin[j]
prec_fin[j]<-1/gam_fin.sq[j]
}
gamma1~dnorm(0,0.01)I(0,)
gamma.sq1<-pow(gamma1,2)
pr1<-1/gamma.sq1

gamma2~dnorm(0,0.01)I(0,)

gamma.sq2<-pow(gamma2,2)

pr2<-1/gamma.sq2

beta1~dnorm(0,1.0E-3)

beta2~dnorm(0,1.0E-3)

}
```

B.5 Implementation of P-EX model in BUGS

```
model{
#within study precision matrix
for (i in 1:ns) {
Prec_w[i,1:2,1:2] <- inverse(Sigma[i,1:2,1:2])
Sigma[i,1,1]<-pow(se[i,1],2)
Sigma[i,2,2]<-pow(se[i,2],2)
Sigma[i,1,2]<-sqrt(Sigma[i,1,1])*sqrt(Sigma[i,2,2])*rho_w[i]
Sigma[i,2,1]<-sqrt(Sigma[i,1,1])*sqrt(Sigma[i,2,2])*rho_w[i]
}
# Random effects model
for (i in 1:ns) {
y[i,1:2]~dmnorm(mu[i,1:2], Prec_w[i,1:2,1:2])
mu[i,1]~dnorm(0,1.0E-3)
mu[i,2]~dnorm(eta[i,class[i]],prec_fin[class[i]])
for (j in 1:nclass) {
eta[i,j]<-lambda0[j]+lambda1[j]*mu[i,1]
}}
for (j in 1:nclass) {
lambda0[j]~dnorm(beta1,pr1)
sd[j]~dnorm(0,2)I(0,)
gam_fin.sq[j]<-pow(sd[j],2)
prec_fin[j]<-1/gam_fin.sq[j]
c[j]~dbern(p[j])
#exchangeability branch
l1.branch[j,1]~dnorm(beta2,pr2)
#Non-exchangeability branch
l1.branch[j,2]~dnorm(0,0.001)
#construct partial exchangeability
#1 for the exchangeable #2  for the non-exchangeable
if_branch[j]<-1+step(-(c[j] - 0.5))
lambda1[j]<-l1.branch[j,if_branch[j]]
}
gamma1~dnorm(0,0.01)I(0,)
gamma.sq1<-pow(gamma1,2)
pr1<-1/gamma.sq1
gamma2~dnorm(0,0.01)I(0,)
gamma.sq2<-pow(gamma2,2)
pr2<-1/gamma.sq2
beta1~dnorm(0,1.0E-3)
beta2~dnorm(0,1.0E-3)}
```

B.6 Bayes factor calculation with Savage Dickey ratio

The following R code calculates Bayes factors for the conditional variance ψ^2 using Savage Dickey density ratio.

```
#calculation of height of the prior distribution of psi#
    fit1 = bugs(data = 'nodata.txt',inits = 'nodata.txt',
                para='gam', model.file = 'prior.txt',
                n.chains = 1, n.burnin = n.burnin,
                n.iter =n.iter, n.thin = 1,
                DIC=F, debug = F,
                save.history = F, OpenBUGS.pgm=obugspath,
                working.directory = wd)

    res.coda1    = as.mcmc.list(fit1)
    a            = res.coda1[,1]
    b            = a[[1]]
    b2           = density(b)
    b2fun        = splinefun(b2$x,b2$y)
    #this is the height of the prior distribution at 0
    prior.height = b2fun(0)

    #calculation of height of the posterior distribution given data#
    fit2 = bugs(data = data, inits = ints,
                para=para, model.file = 'model.txt',
                n.chains = 1, n.burnin = n.burnin,
                n.iter =n.iter, n.thin = 1,
                DIC=F, debug = F,
                save.history = F, OpenBUGS.pgm=obugspath,
                working.directory = wd)
    res.coda      = as.mcmc.list(fit2)
    c             = res.coda[,1]
    d             = c[[1]]
    d2            = density(d)
    d2fun         = splinefun(d2$x,d2$y)
    #this is the height of the posterior distribution at 0
    post.height   = d2fun(0)

    #Bayes factor is the ratio of the heights
    BF            = post.height/prior.height
```

The next two BUGS models were used for the calculation of Bayes factors :

```
# Prior distribution placed on psi
model{
    psi~dnorm(0,2)I(0,)
    psi.sq<-pow(gamma,2)
    }

# Hierarchical model using the same prior distribution placed on psi
model{
    #within study precision matrix
    for (i in 1:ns) {
     prec_w[i,1:2,1:2] <- inverse(delta[i,1:2,1:2])
    #covariance matrix for the j-th study
     delta[i,1,1]<-pow(se[i,1],2)
     delta[i,2,2]<-pow(se[i,2],2)
     delta[i,1,2]<-sqrt(delta[i,1,1])*sqrt(delta[i,2,2])*rho_w[i]
     delta[i,2,1]<-sqrt(delta[i,1,1])*sqrt(delta[i,2,2])*rho_w[i]}
    # Random effects model
    for (i in 1:ns) {
     y[i,1:2]~dmnorm(mu[i,1:2], prec_w[i,1:2,1:2])
    #  product normal formulation for the between study part:
     mu[i,1]~dnorm(0,1.0E-3)
```

```
 mu[i,2]~dnorm(eta[i],prec_fin)
 eta[i]<-lambda0+lambda1*mu[i,1]
 }
psi~dnorm(0,2)I(0,)
psi.sq<-psi*psi
prec_fin<-1/psi.sq
lambda0~dnorm(0,0.001)
lambda1~dnorm(0,0.001)
 }
```

B.7 Convergence plots of F-EX and P-EX models

B.7.1 Convergence plots of F-EX model

The following figures display trace and density plots of the parameters describing the surrogate relationships on PFS-OS pair of outcomes in the aCRC data set in chapter 4. These two types of figures are used as tools to illustrate the convergence of the key parameters of the model.

Figure B.1: Trace - density plots of 3 chains consisting of 50000 iterations each after 20000 iterations burn-in period

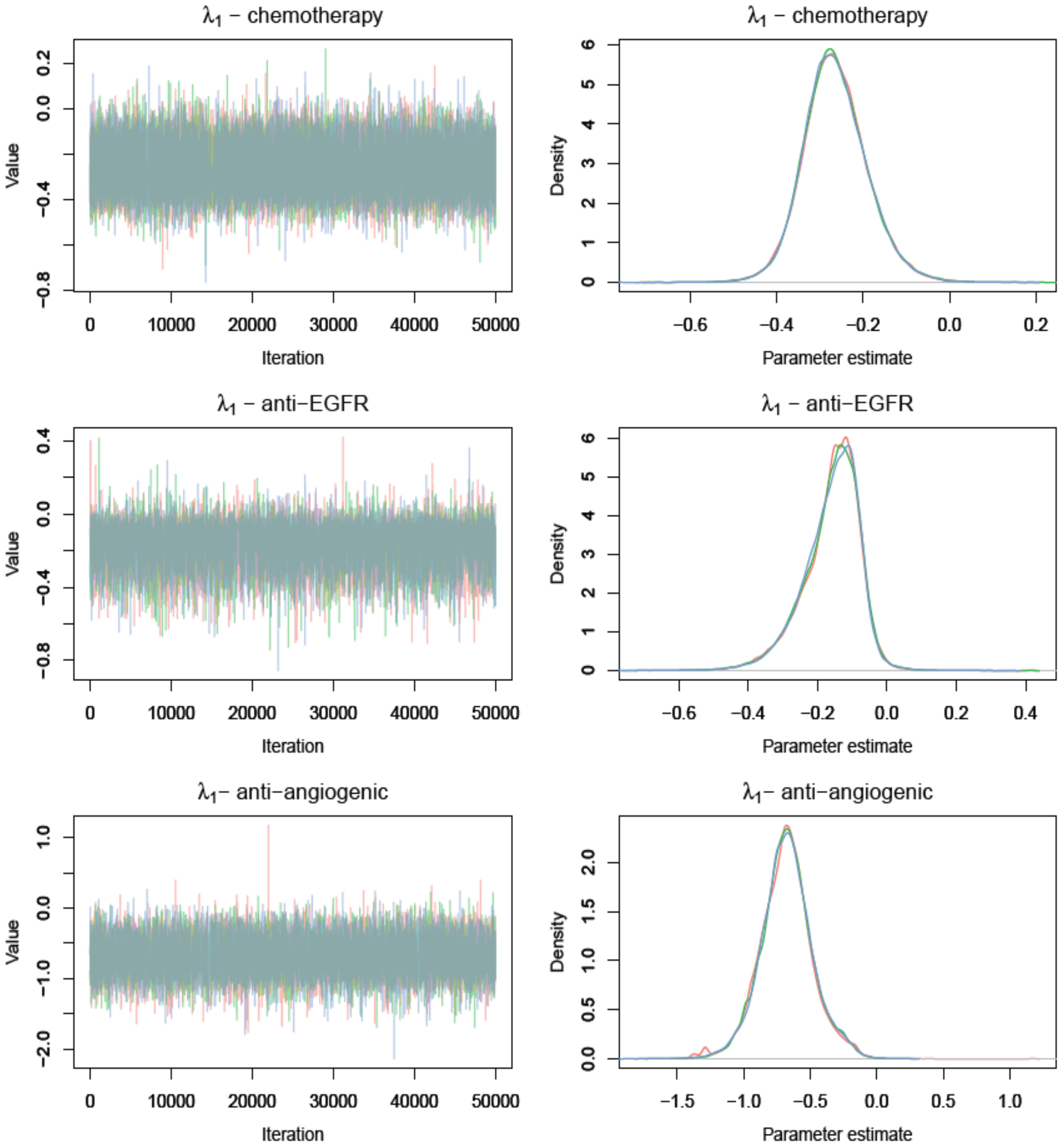


Figure B.2: Trace - density plots of 3 chains consisting of 50000 iterations each after 20000 iterations burn-in period

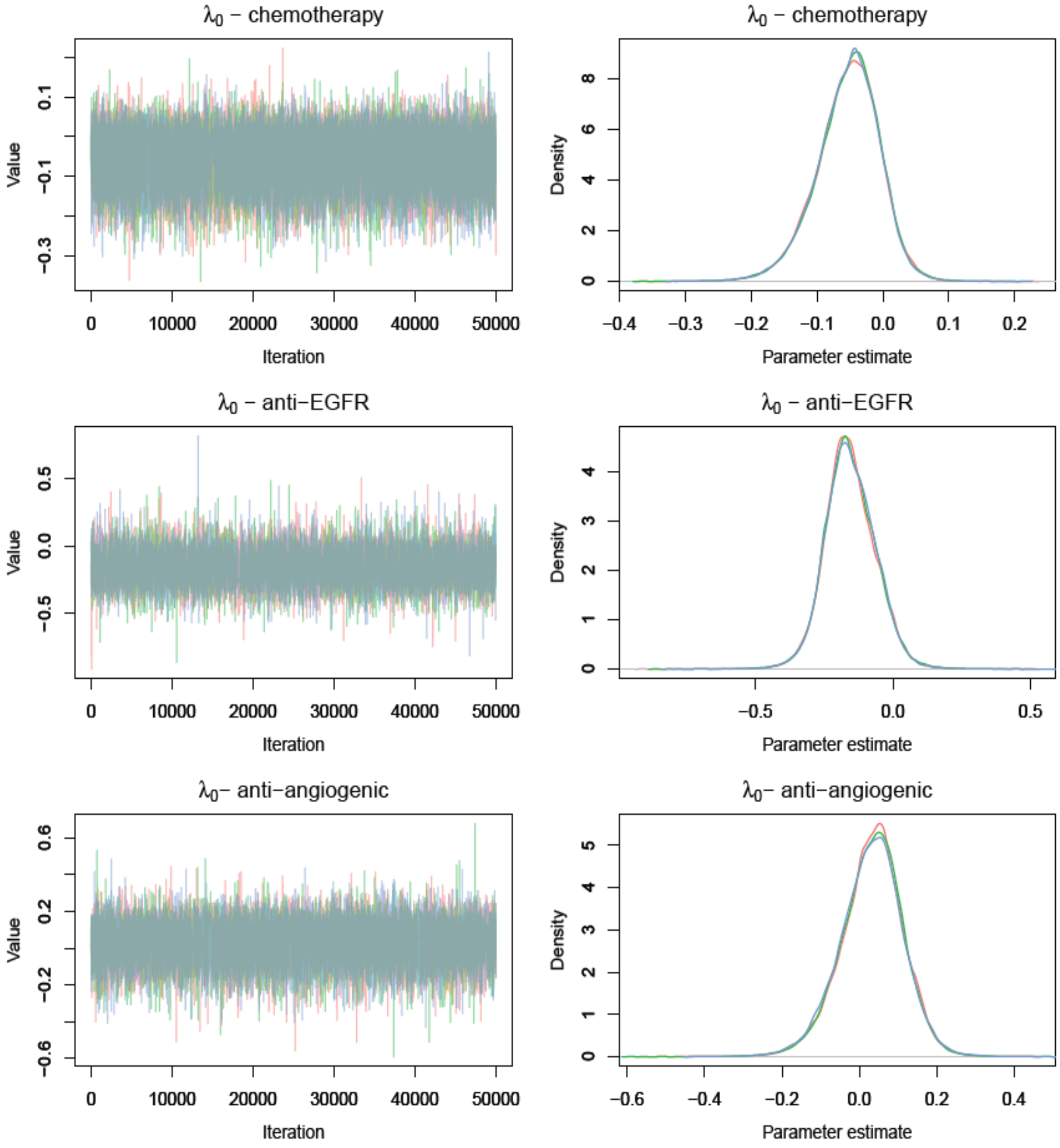
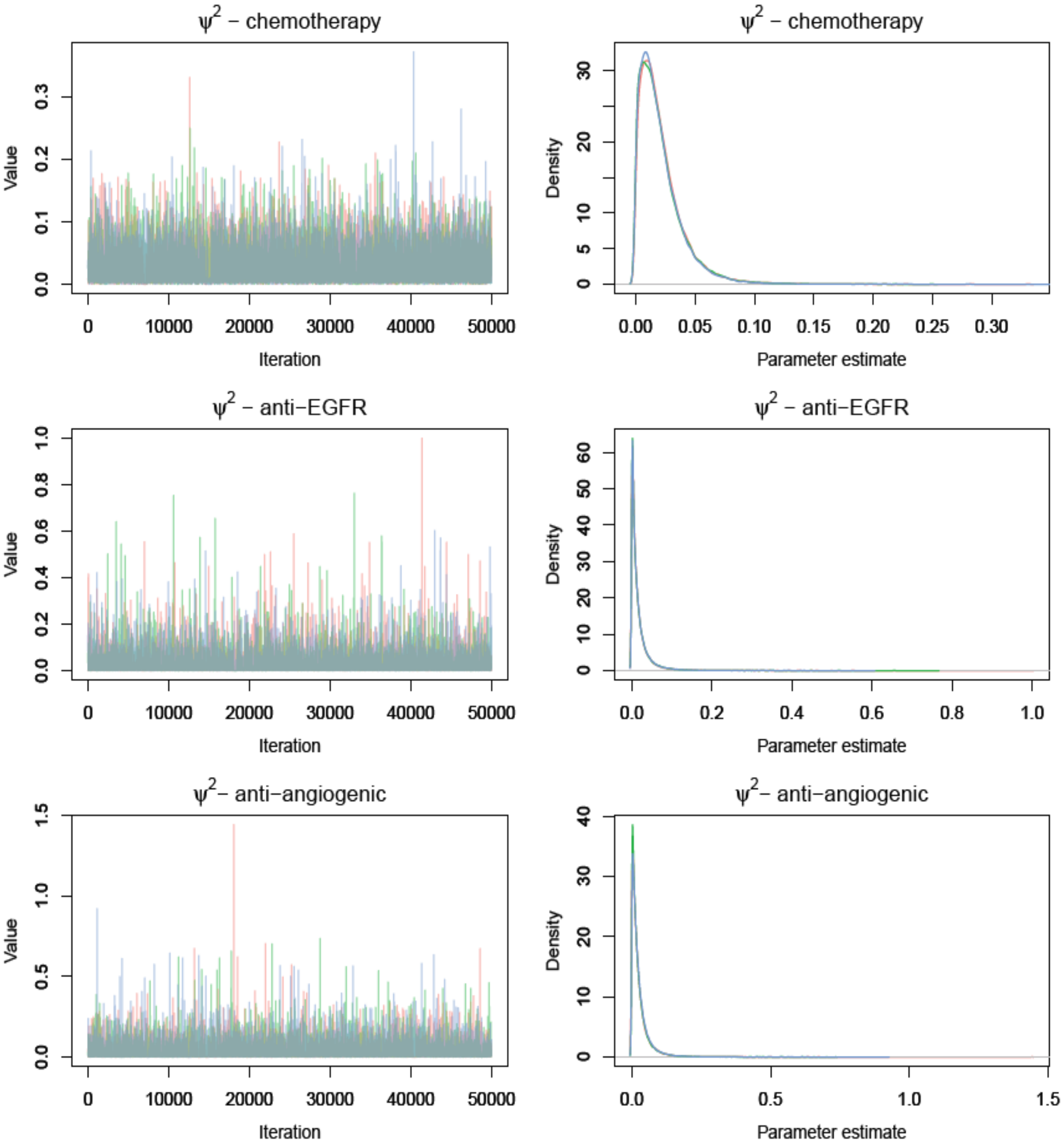


Figure B.3: Trace - density plots of 3 chains consisting of 50000 iterations each after 20000 iterations burn-in period



B.7.2 Convergence plots of P-EX model

The following figures display trace and density plots of the parameters describing the surrogate relationships on PFS-OS pair of outcomes in the aCRC data set in chapter 4.

Figure B.4: Trace - density plots of 3 chains consisting of 50000 iterations each after 20000 iterations burn-in period

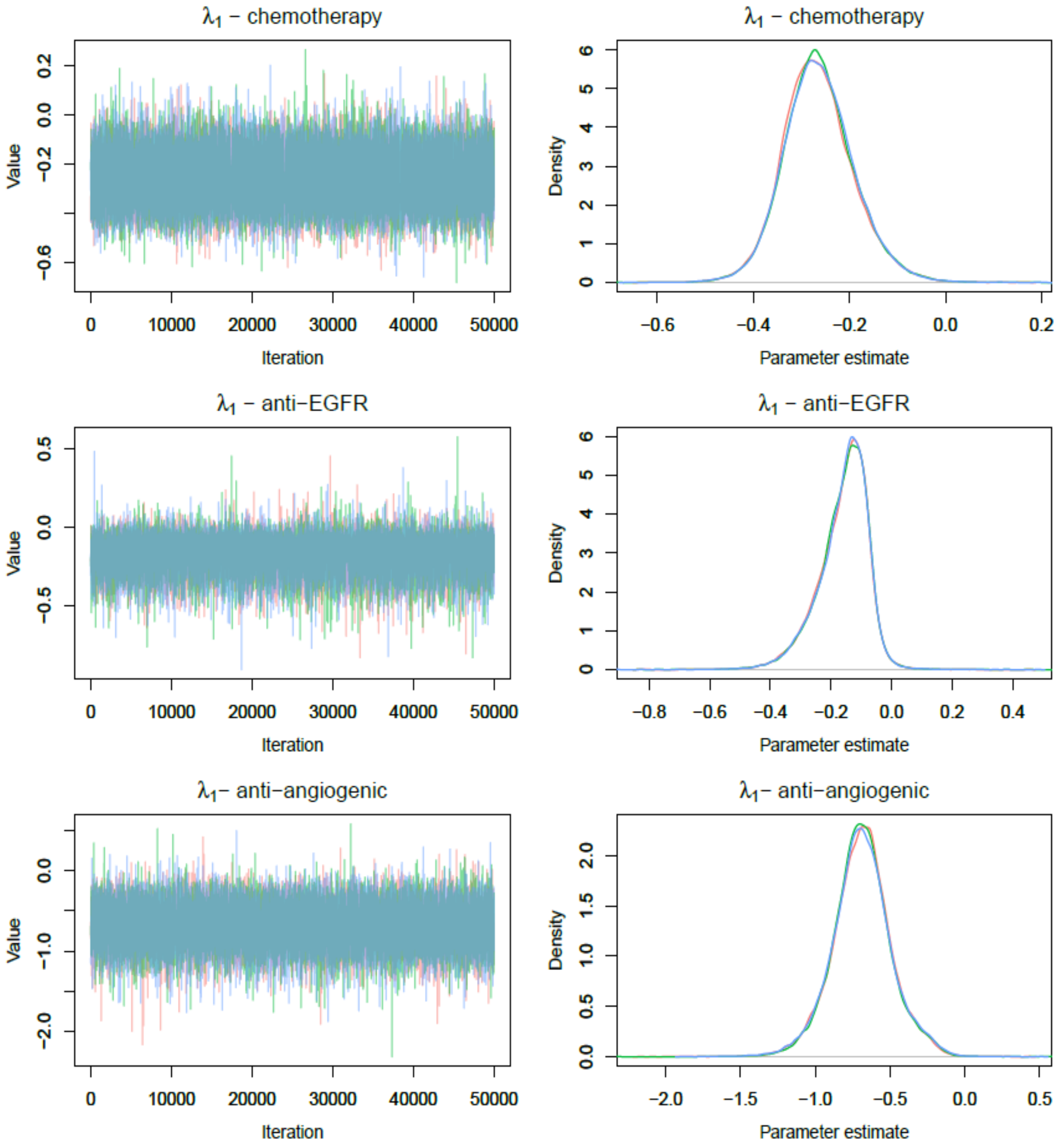


Figure B.5: Trace - density plots of 3 chains consisting of 50000 iterations each after 20000 iterations burn-in period

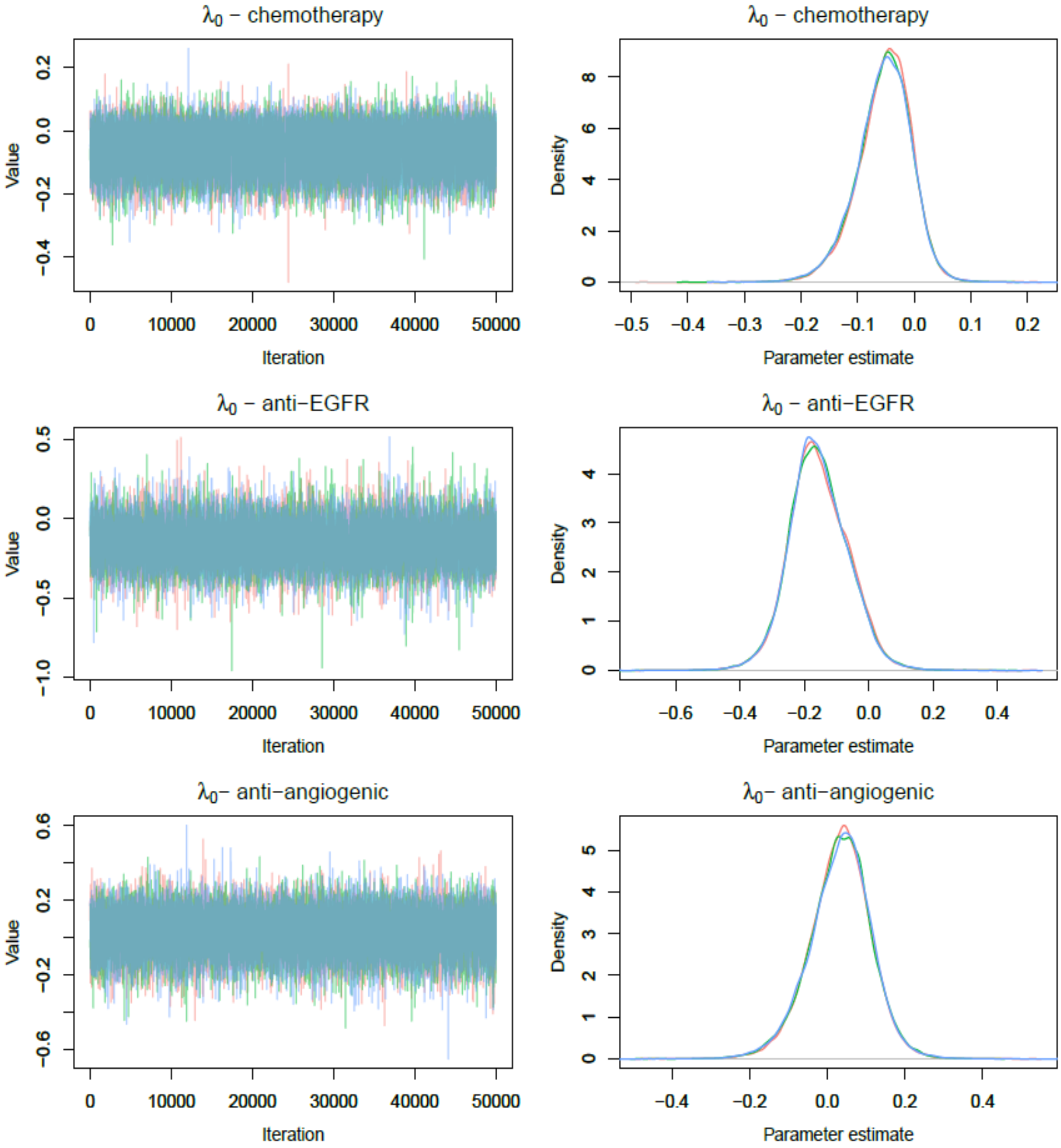
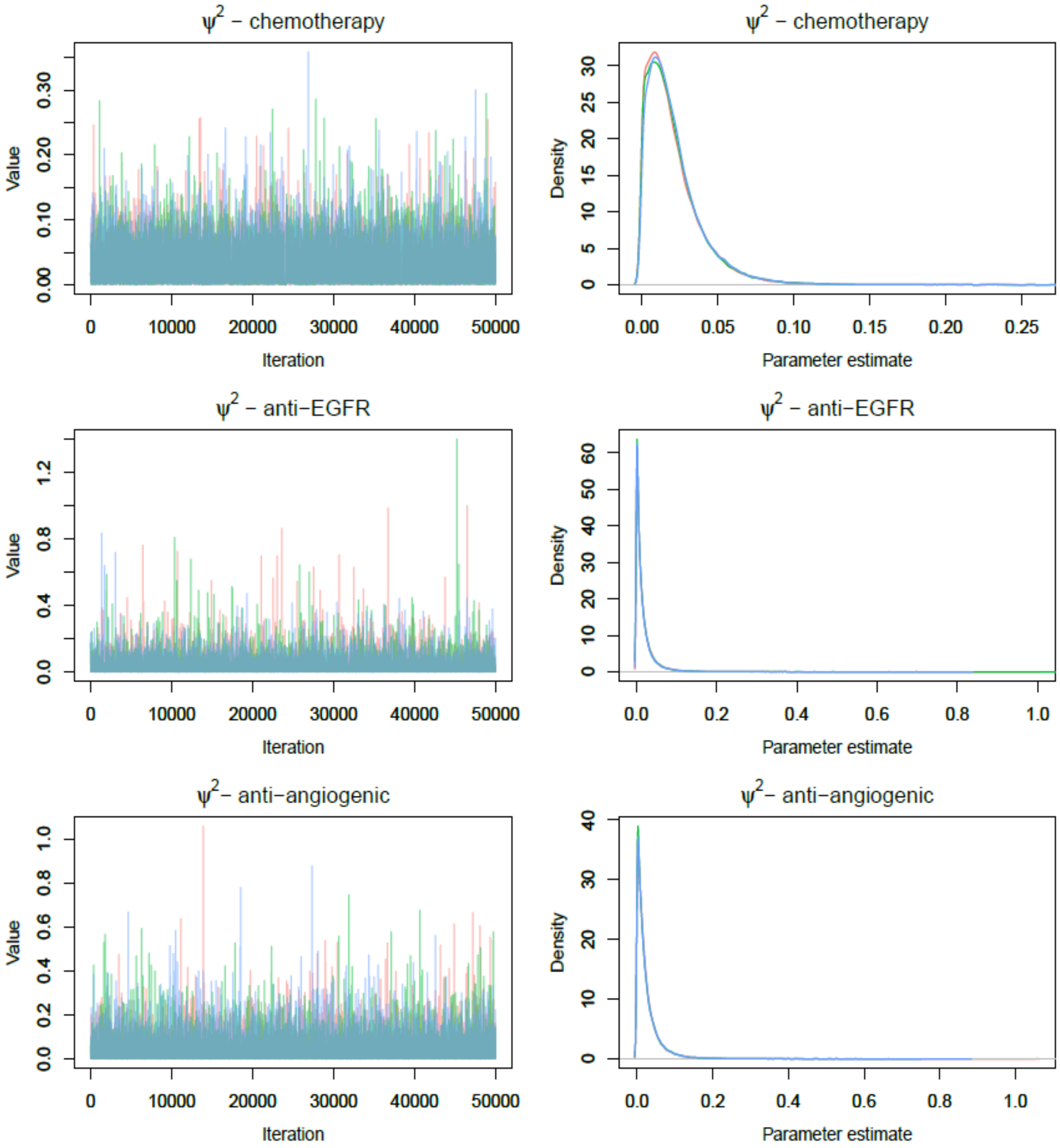


Figure B.6: Trace - density plots of 3 chains consisting of 50000 iterations each after 20000 iterations burn-in period



Appendix C

Appendix

C.1 Implementation of BRMA model in Stan

```
data{
int<lower = 0> Ns;
int<lower = 0> nA[Ns,2];
int<lower = 0> nB[Ns,2];
int<lower = 0> rA[Ns,2];
int<lower = 0> rB[Ns,2];
real rho_w[Ns];}

transformed data{
//Calculate log odds ratios
vector[2] Y[Ns];
vector[2] S[Ns];
for (i in 1:Ns){
Y[i,1]=log(((rB[i,1]+0.5)*(nA[i,1]-rA[i,1]+0.5))/
           ((nB[i,1]-rB[i,1]+0.5)*(rA[i,1]+0.5)));
Y[i,2]=log(((rB[i,2]+0.5)*(nA[i,2]-rA[i,2]+0.5))/
           ((nB[i,2]-rB[i,2]+0.5)*(rA[i,2]+0.5)));
S[i,1]=sqrt((1/(rB[i,1]+0.5))+(1/(nB[i,1]-rB[i,1]+0.5))+
(1/(rA[i,1]+0.5))+(1/(nA[i,1]-rA[i,1]+0.5)));
```

```
S[i,2]=sqrt((1/(rB[i,2]+0.5))+(1/(nB[i,2]-rB[i,2]+0.5))+
(1/(rA[i,2]+0.5))+(1/(nA[i,2]-rA[i,2]+0.5)));}}

parameters{
real rr;
vector[2]  b;
vector[2]  z[Ns];
vector<lower=0, upper=5>[2] tau;}

transformed parameters{
matrix[2,2] Tau;
matrix[2,2] L;
matrix[2,2] Sigma1[Ns];
vector[2] delta[Ns];
real<lower= -1, upper=1> rho1;
rho1           = tanh(rr);
for (i in 1:Ns){
Sigma1[i,1, 1] = S[i,1]^2;
Sigma1[i,1, 2] = S[i,1]*S[i,2]*rho_w[Ns];
Sigma1[i,2, 1] = S[i,1]*S[i,2]*rho_w[Ns];
Sigma1[i,2, 2] = S[i,2]^2;}
Tau[1,1]       = tau[1]^2;
Tau[2,2]       = tau[2]^2;
Tau[1,2]       = tau[1]*tau[2]*rho1;
Tau[2,1]       = tau[1]*tau[2]*rho1;
L              = cholesky_decompose(Tau);
//non-centred parameterisation for delta~multi_normal(b,Tau)
for (i in 1:Ns){
delta[i] = b + (L*z[i]);}

model{
//priors
```

```
rr       ~ std_normal();
b        ~ normal(0, 10);
for (i in 1:Ns){
z[i]     ~ std_normal();
//likelihood
Y[i]     ~ multi_normal(delta[i],Sigma1[i]);}}
```

C.2 Implementation of BRMA-IB model in Stan

```
data{
int<lower = 0> Ns;
int<lower = 0> nA[Ns,2];
int<lower = 0> nB[Ns,2];
int<lower = 0> rA[Ns,2];
int<lower = 0> rB[Ns,2];}

parameters {
real rr;
vector[2] b;
vector<lower = 0, upper = 5>[2] tau;
vector[2] z[Ns];
vector[2] mu[Ns];}

transformed parameters{
matrix[2,2] Tau;
matrix[2,2] L;
vector[2] delta[Ns];
real<lower= -1, upper=1> rho1;
rho1      = tanh(rr);
Tau[1, 1] = tau[1]^2;
Tau[1, 2] = tau[1]*tau[2]*rho1;
Tau[2, 1] = tau[1]*tau[2]*rho1;
Tau[2, 2] = tau[2]^2;
L         = cholesky_decompose(Tau);
//non-centred parameterisation for delta~multi_normal(b,Tau)
for (i in 1:Ns){
delta[i]  = b+ L*z[i];}}

model {
```

```
//priors
b         ~ normal(0, 10);
rr        ~ std_normal();
for (i in 1:Ns){
z[i]      ~ std_normal();
mu[i]     ~ normal(0, 10);
//likelihoods
rA[i,1]   ~ binomial_logit(nA[i,1], mu[i,1]);
rA[i,2]   ~ binomial_logit(nA[i,2], mu[i,2]);
rB[i,1]   ~ binomial_logit(nB[i,1], mu[i,1]+delta[i,1]);
rB[i,2]   ~ binomial_logit(nB[i,2], mu[i,2]+delta[i,2]); }}
```

C.3 Implementation of BRMA-BC model in Stan

```
functions {
// frank Copula CDF
real fcop(real theta, real u, real v) {
real a = -1 / theta * log1p((expm1((-theta * u)) *
         (expm1(-theta * v)) / (expm1(-theta))));
return  a;
}
//Gumbel Copula CDF
real fcop2( real theta, real u,real v) {
    real a;
    real t1 = u;
    real t2 = v;
    real neg_log_u;
    real neg_log_v;
    if (t1>.999999) {t1=.999999;}//boundary condition
    if (t2>.999999) {t2=.999999;}//boundary condition
    neg_log_u = -log(t1);
    neg_log_v = -log(t2);
    a = exp(-(neg_log_u^theta+neg_log_v^theta)^(1/theta));
    return a;
  }
//Gaussian Copula CDF
real fcop3(real theta, real u1, real u2){
    real t1 = u1;real z1;
    real t2 = u2;real z2;
    if (t1 > .9999999) t1 = .9999999;//boundary condition
    if (t2 > .9999999) t2 = .9999999;//boundary condition
    z1 = inv_Phi(t1);
    z2 = inv_Phi(t2);
    if (z1 != 0 || z2 != 0) {
      real denom = fabs(theta) < 1.0 ? sqrt((1 + theta) *
                   (1 - theta)) : not_a_number();
      real a1 = (z2 / z1 - theta) / denom;
      real a2 = (z1 / z2 - theta) / denom;
      real product = z1 * z2;
      real delta = product < 0 || (product == 0 && (z1 + z2) < 0);
      real a = log(0.5 * (Phi(z1) + Phi(z2) - delta) -
               owens_t(z1, a1) - owens_t(z2, a2));
      return exp(a);
    }
    if (theta == 1){
      vector[2] z;
      z[1]=z1;z[2]=z2;
      return min(Phi(z));
    }
    return 0.25 + asin(theta) / (2 * pi());
  }


//Bivariate pmf to model binomial aggregate data jointly
real Bivfcop_lpmf(int[] r,int n1, int n2, real theta, vector mu){
real p1  = inv_logit(mu[1]);
real p2  = inv_logit(mu[2]);
real f11 = binomial_cdf(r[1]-1, n1, p1);
real f12 = binomial_cdf(r[2]-1, n2, p2);
real f1  = f11 + exp(binomial_logit_lpmf(r[1] |n1, mu[1]));
real f2  = f12 + exp(binomial_logit_lpmf(r[2] |n2, mu[2]));
real prob= fcop(theta,f1,f2)-fcop(theta,f1,f12)-
           fcop(theta,f11,f2)+fcop(theta,f11,f12);
return log(prob);}}

data{
```

```
int<lower = 0> Ns;
int<lower = 0> nA[Ns,2];
int<lower = 0> nB[Ns,2];
int<lower = 0> rA[Ns,2];
int<lower = 0> rB[Ns,2];
real theta1[Ns];real theta2[Ns];}


parameters{
real rr;
vector[2] b;
vector<lower = 0, upper = 5>[2] tau;
vector[2] z[Ns];
vector[2] mu[Ns];}

transformed parameters{
matrix[2,2] L;
matrix[2,2] Tau;
vector[2] delta[Ns];
real<lower= -1, upper=1> rho1;
rho1      = tanh(rr);
Tau[1,1]  = tau[1]^2;
Tau[2,2]  = tau[2]^2;
Tau[1,2]  = tau[1]*tau[2]*rho1;
Tau[2,1]  = tau[1]*tau[2]*rho1;
L         = cholesky_decompose(Tau);
//non-centred parameterisation for delta~multi_normal(b,Tau)
for (i in 1:Ns){
delta[i]  = b + (L*z[i]);}}

model{
//priors
```

```
rr       ~ std_normal();
b        ~ normal(0, 10);
for (i in 1:Ns){
z[i]     ~ std_normal();
mu[i]    ~ normal(0, 10);
//likelihoods
rA[i]    ~ Bivfcop(nA[i,1],nA[i,2], theta1[i], mu[i]);
rB[i]    ~ Bivfcop(nB[i,1],nB[i,2], theta2[i], mu[i]+delta[i]);}}
```

C.4 True values of the dependence parameters of the generation process in Section 5.3.1

To simulate IPD with low, moderate and strong association, we used a joint density made with Bernoulli marginal distributions constructed with Frank copula. The following table presents the values of the dependence parameters across the scenarios and the approximate values of the corresponding Spearman's correlation.

Table C.1: Values of dependence parameters and their corresponding Spearman's correlation

Strength of association	Parameter	Average proportion of events = 0.5	Average proportion of events =0.95
Low within- study	$\theta_A = \theta_B$	1.2	6.4
association	ρ_S	0.15	0.16
Moderate within- study	$\theta_A = \theta_B$	4.2	25
association	ρ_S	0.45	0.48
High within- study	$\theta_A = \theta_B$	14	100
association	ρ_S	0.75	0.76

C.5 Bootstrap method used to estimate the within-study correlations of BRMA

This section presents the bootstrap method used to estimate the within-study correlations of BRMA in each study. Specifically, the treatment effects (on the log odds ratio scale) on each outcome and were calculated for each bootstrap sample by using the standard formulas and the Pearson's correlation coefficient between the treatment effects were obtained.

```
bootstrap1 = function(df,Nb){
#Nb=number of bootstrap samples, df= dataframe containing IPD
names(df) = paste(c('Y1A','Y2A','Y1B','Y2B'))
s         = length(df$Y1A)#number of observations in the data
y1=y2=array(0,Nb) #Nb=2000 was used
for (d in 1:Nb){
sam       = sample(s, replace=T)
boot.1    = df$Y1A[sam]
boot.2    = df$Y1B[sam]
boot.3    = df$Y2A[sam]
boot.4    = df$Y2B[sam]
r1A       = sum(boot.1)
r1B       = sum(boot.2)
r2A       = sum(boot.3)
r2B       = sum(boot.4)
#Log odds ratio on the first outcome
LOR1      = log(((r1B+0.5)*(s-r1A+0.5))/((s-r1B+0.5)*(r1A+0.5)))
#Log odds ratio on the second outcome
LOR2      = log(((r2B+0.5)*(s-r2A+0.5))/((s-r2B+0.5)*(r2A+0.5)))
y1[d]     = LOR1
y2[d]     = LOR2}
#the correlations between log odds ratios across bootstrap samples
rho = cor(y1,y2,method= "pearson")
return(list(rho=rho)) }
```

C.6 Bootstrap method used to estimate the within-study association parameters of BRMA-BC

In this section we present the implementation of a second bootstrap method. This method was used to estimate the dependence parameters θ_{Ai}, θ_{Bi} of the joint density made with copula study between the first and the second outcome in each arm. Specifically, summary data were calculated for each bootstrap sample and then dependence parameters of the Frank and the Gaussian copulas were estimated by using a optimiser such as *nlm* or *optimize* in R. We also present the Frank and the Gaussian bivariate pdfs used to estimate the dependence parameters.

```
bootstrap2 = function (df,Nb){
#Nb=number of bootstrap samples, df= dataframe containing IPD
names(df)  = paste(c('Y1A','Y2A','Y1B','Y2B'))
s          = length(df$Y1A)#number of observations in the data
y1A=y1B=y2A=y2B<-array(0,Nb)
#Generate bootstrap samples and calculate the summary data for each one
for (k in 1:Nb){
sam     = sample(s, replace=T)
y1A[k]  = sum(df$Y1A[sam])
y1B[k]  = sum(df$Y1B[sam])
y2A[k]  = sum(df$Y2A[sam])
y2B[k]  = sum(df$Y2B[sam])}
#Binomial likelihoods for each arm and outcome
llik1   = function(p)-sum(dbinom(y1A,prob=p,size=s,log=TRUE))
llik2   = function(p)-sum(dbinom(y1B,prob=p,size=s,log=TRUE))
llik3   = function(p)-sum(dbinom(y2A,prob=p,size=s,log=TRUE))
llik4   = function(p)-sum(dbinom(y2B,prob=p,size=s,log=TRUE))
p1A.hat = nlm(llik1, p=0.5)
p1B.hat = nlm(llik2, p=0.5)
p2A.hat = nlm(llik3, p=0.5)
p2B.hat = nlm(llik4, p=0.5)
uA      = pbinom(y1A,s,p1A.hat$estimate)
vA      = pbinom(y2A,s,p2A.hat$estimate)
uB      = pbinom(y1B,s,p1B.hat$estimate)
vB      = pbinom(y2B,s,p2B.hat$estimate)
fA      = function(theta1) {-sum(log(dfrk(uA,vA,theta1)))}
fB      = function(theta2) {-sum(log(dfrk(uB,vB,theta2)))}
gA      = function(theta1) {-sum(log(dbvncop(uA,vA,theta1)))}
gB      = function(theta2) {-sum(log(dbvncop(uB,vB,theta2)))}
#dependence parameters in each arms
thetafA  = optimize(fA, c(-30,31))$min
thetagA  = optimize(gA, c(-.99,.99))$min
thetafB  = optimize(fB, c(-30,31))$min
thetagB  = optimize(gB, c(-.99,.99))$min
return(list(thetafA=thetafA, thetafB=thetafB,
            thetagA=thetagA, thetagB=thetagB))}



#####################
#PDF of Frank copula#
#####################
dfrk = function(u,v,cpar) {
    t1=1.-exp(-cpar);
    tem1=exp(-cpar*u); tem2=exp(-cpar*v);
    pdf=cpar*tem1*tem2*t1;
    tem=t1-(1.-tem1)*(1.-tem2);
    pdf=pdf/(tem*tem);
return(pdf)}
```

```
############################
#PDF of the Gaussian copula#
############################
 dbvncop=function(u,v,cpar){
 #boundary conditions to avoid errors
 #when the proportions are close to 1 or 0
        u[u>=.999999]=1-.000001; v[v>=.999999]=1-0.000001
        u[u<0.000001]=0.000001; v[v<0.000001]=0.000001
        x1=qnorm(u); x2=qnorm(v)
        qf=x1^2+x2^2-2*cpar*x1*x2
        qf=qf/(1-cpar^2)
        con=sqrt(1-cpar^2)*(2*pi)
        pdf=exp(-.5*qf)/con
        pdf=pdf/(dnorm(x1)*dnorm(x2))
    return(pdf)}
```

C.7 Bootstrap method used to estimate the within-study association parameters in section 6.4.1

In this section we present the bootstrap method which was used to estimate the within-study associations in section 6.4.1. The method estimated the within-study correlations between the log odds on the first and the second outcome in arm A and B

```
bootstrap3  = function(df,Nb){
  names(df)      = paste(c('Y1A','Y2A','Y1B','Y2B'))
  s<-length(df$Y1A)#number of observation in the data
  y1A<-y2A<-y1B<-y2B<-array(0,Nb)
  for (i in 1:Nb){
    sam<-sample(s, replace=T)
    boot.1   <-df$Y1A[sam]
    boot.2   <-df$Y1B[sam]
    boot.3   <-df$Y2A[sam]
    boot.4   <-df$Y2B[sam]
    r1A       <- sum(boot.1)
    r1B       <- sum(boot.2)
    r2A       <- sum(boot.3)
    r2B       <- sum(boot.4)
    #Log odds on the first and the second outcome
    LOA1      <- log((r1A+0.5)/(s-r1A+0.5))
    LOA2      <- log((r2A+0.5)/(s-r2A+0.5))
    LOB1      <- log((r1B+0.5)/(s-r1B+0.5))
    LOB2      <- log((r2B+0.5)/(s-r2B+0.5))
    y1A[i]    <-LOA1
    y2A[i]    <-LOA2
    y1B[i]    <-LOB1
    y2B[i]    <-LOB2
  }
  rhoA<-cor(y1A,y2A,method= "pearson")
  rhoB<-cor(y1B,y2B,method= "pearson")
  return(list(rhoA=rhoA,rhoB=rhoB)) #the correlations )
}
```

C.8 Convergence plots of BRMA, BRMA-IB and the three versions of BRMA-BC models

The following figures display trace and density plots of the parameters between studies parameters including the parameter of the intercept (λ_0) on CCyR-EFS pair of outcomes in the CML data-set in Chapter 5, illustrating the performance of the models in terms of convergence

C.8.1 Convergence plots of BRMA model

Figure C.1: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period

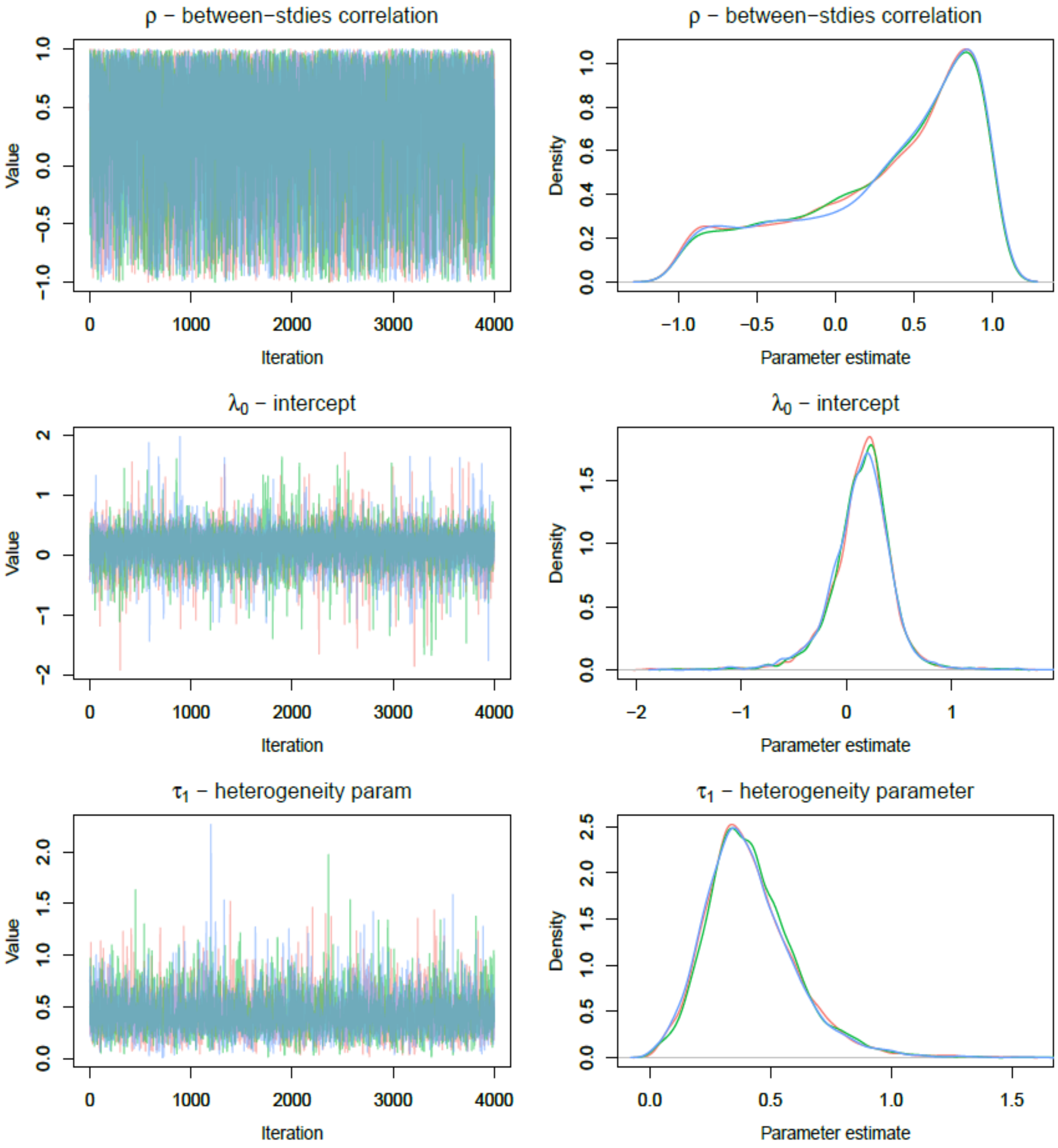
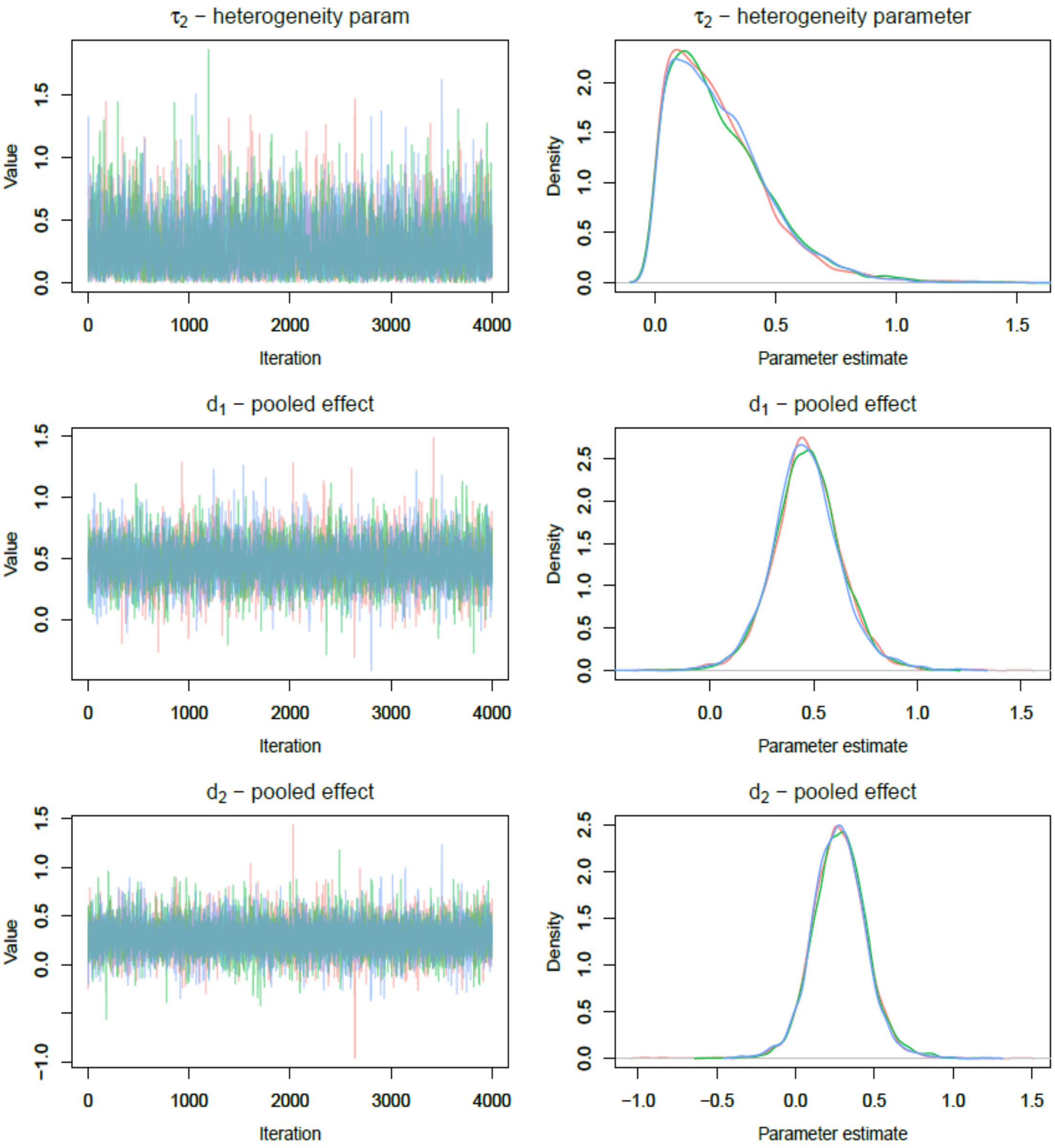


Figure C.2: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period



C.8.2 Convergence plots of BRMA-IB model

Figure C.3: Trace - density plots of 3 chains consisting of 2000 iterations each after 1000 iterations burn-in period

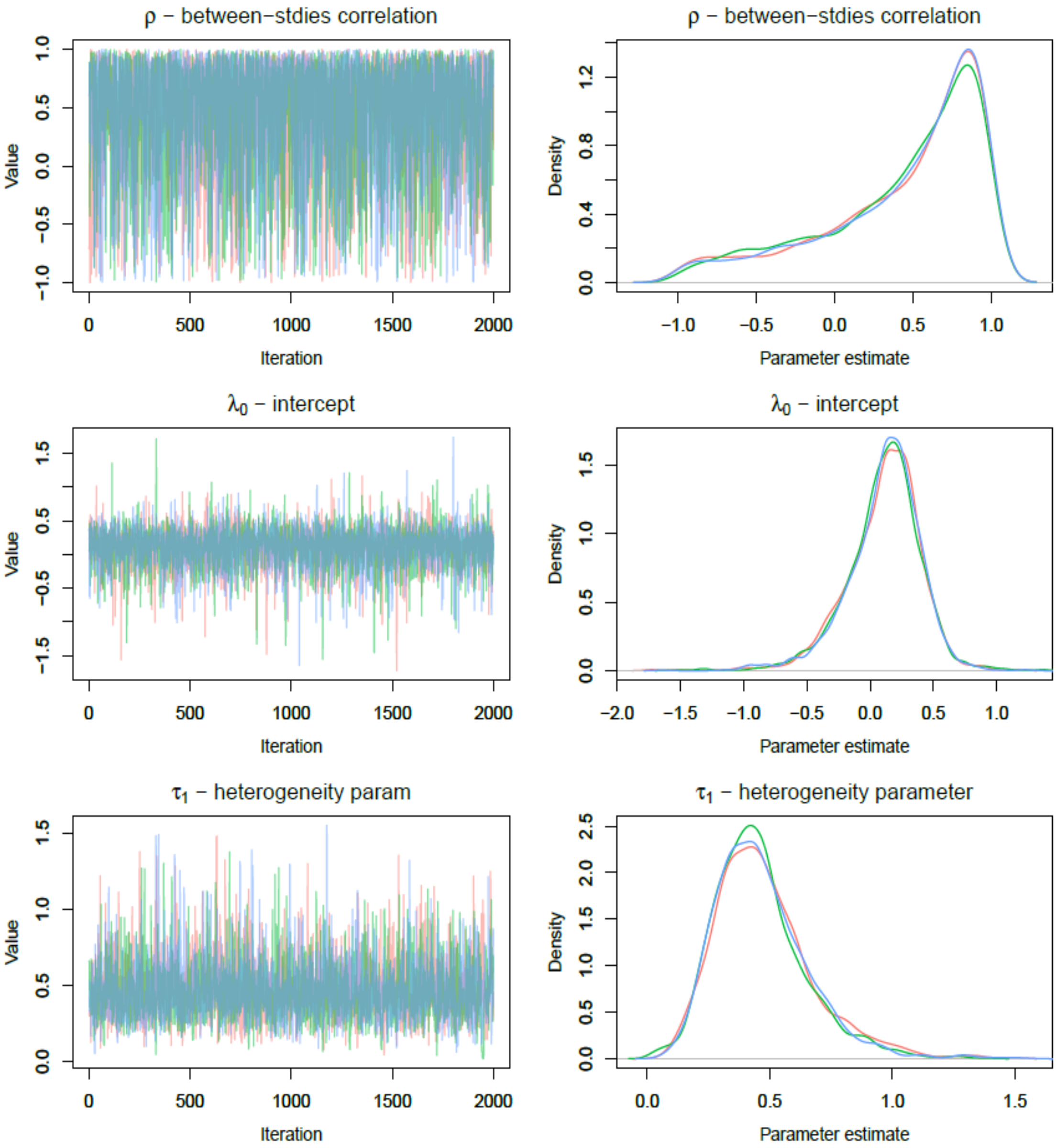
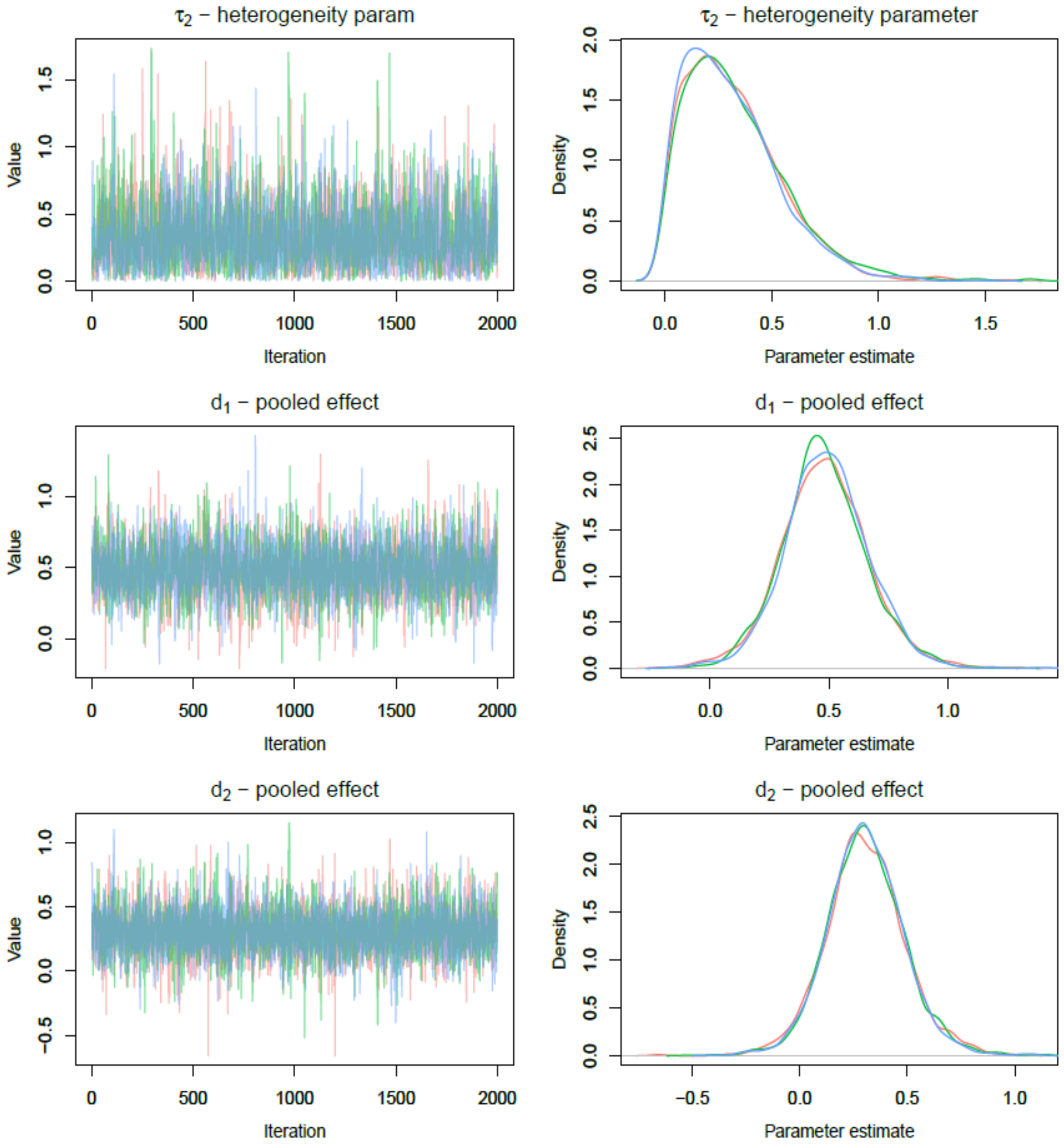


Figure C.4: Trace - density plots of 3 chains consisting of 2000 iterations each after 1000 iterations burn-in period



C.8.3 Convergence plots of BRMA-BC with frank copula model

Figure C.5: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period

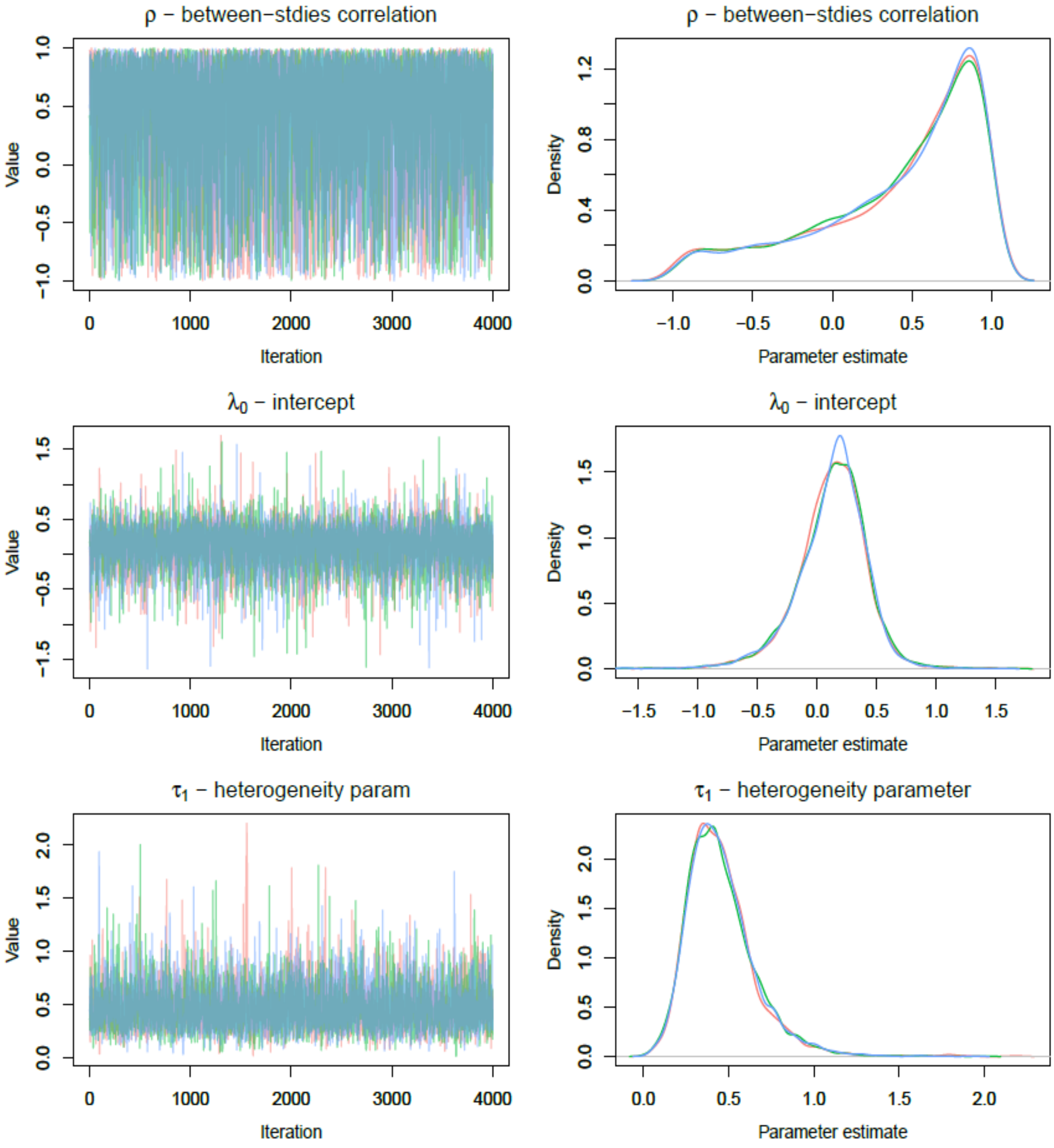
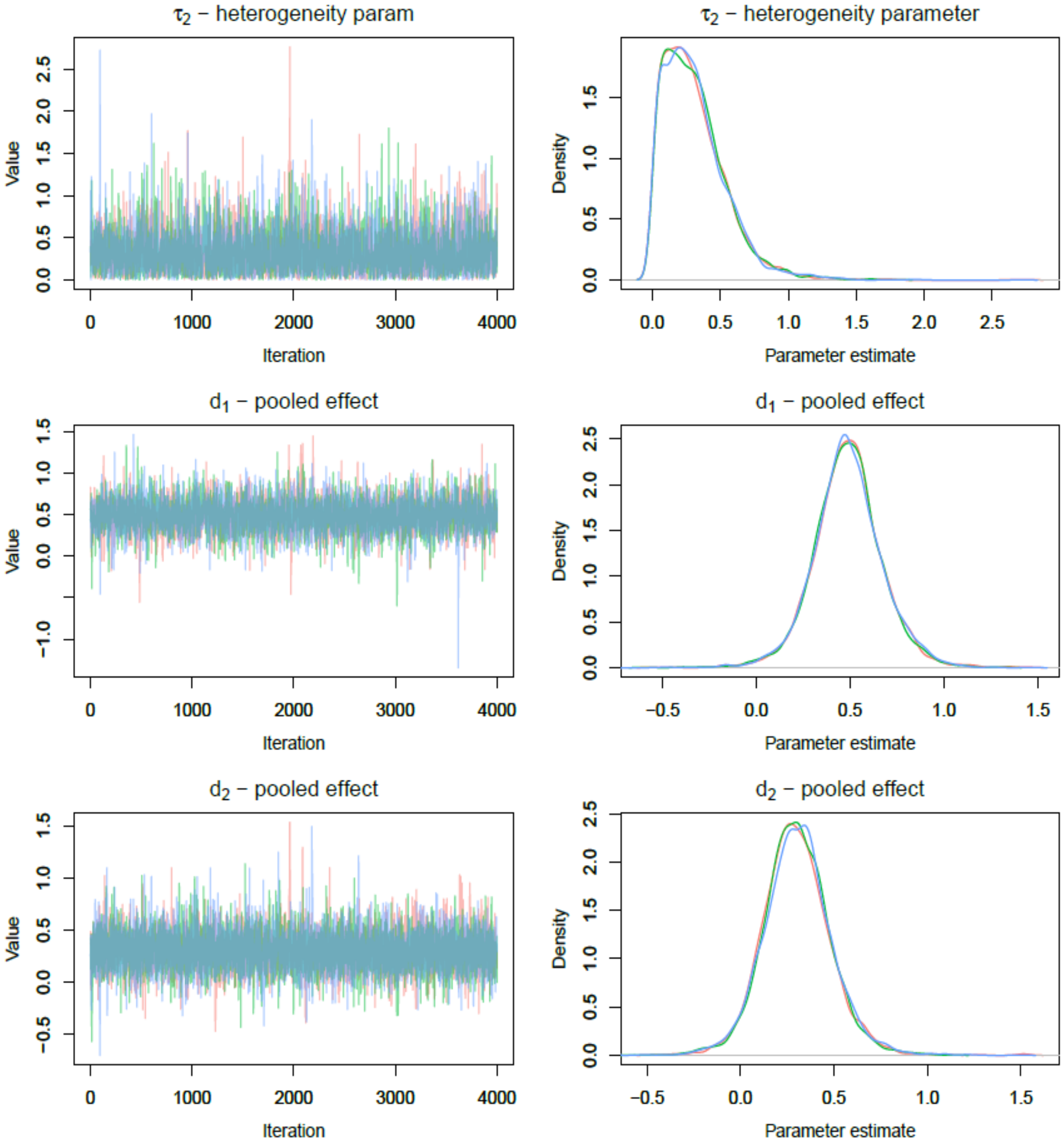


Figure C.6: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period



C.8.4 Convergence plots of BRMA-BC with Gaussian copula model

Figure C.7: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period

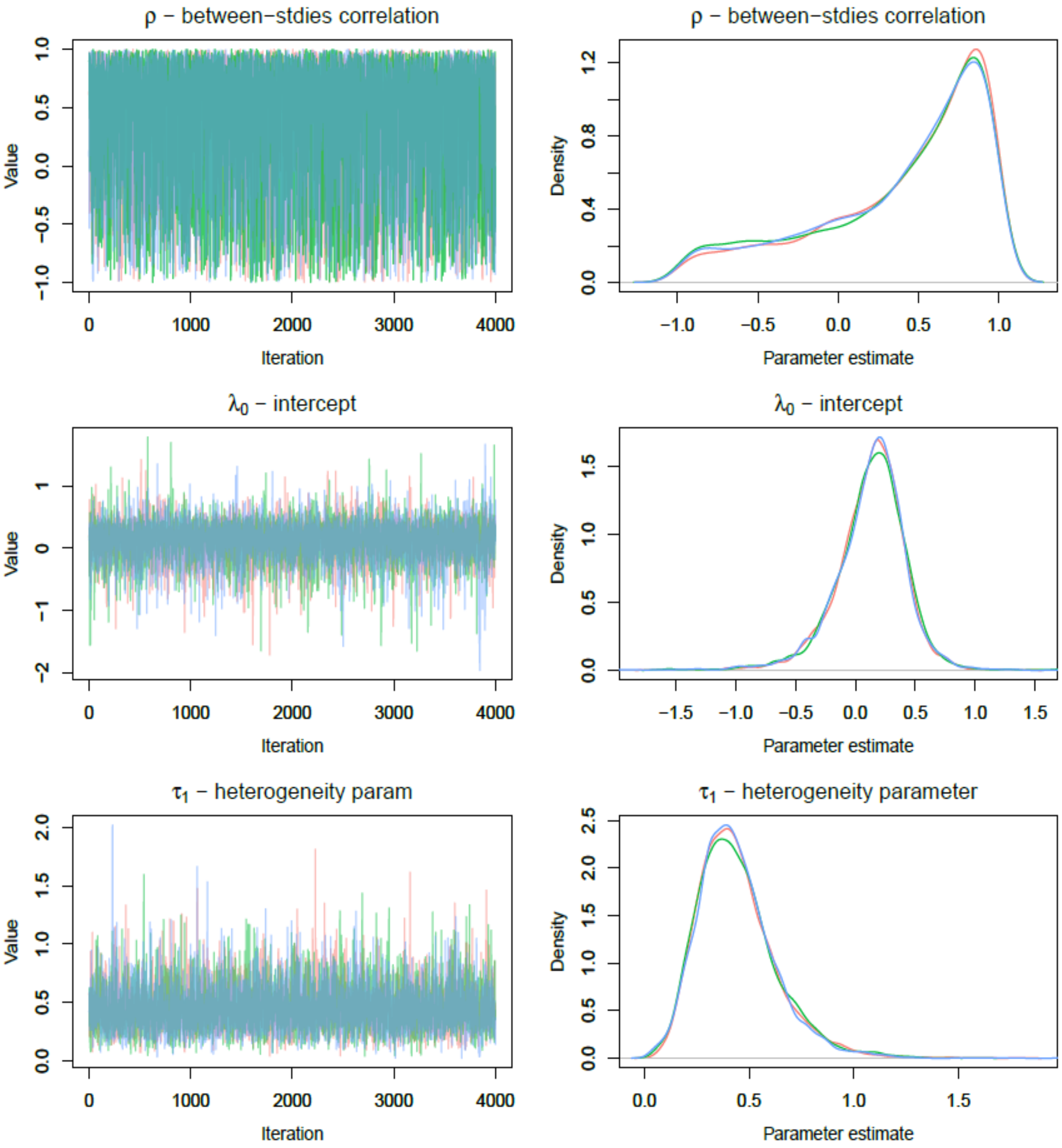
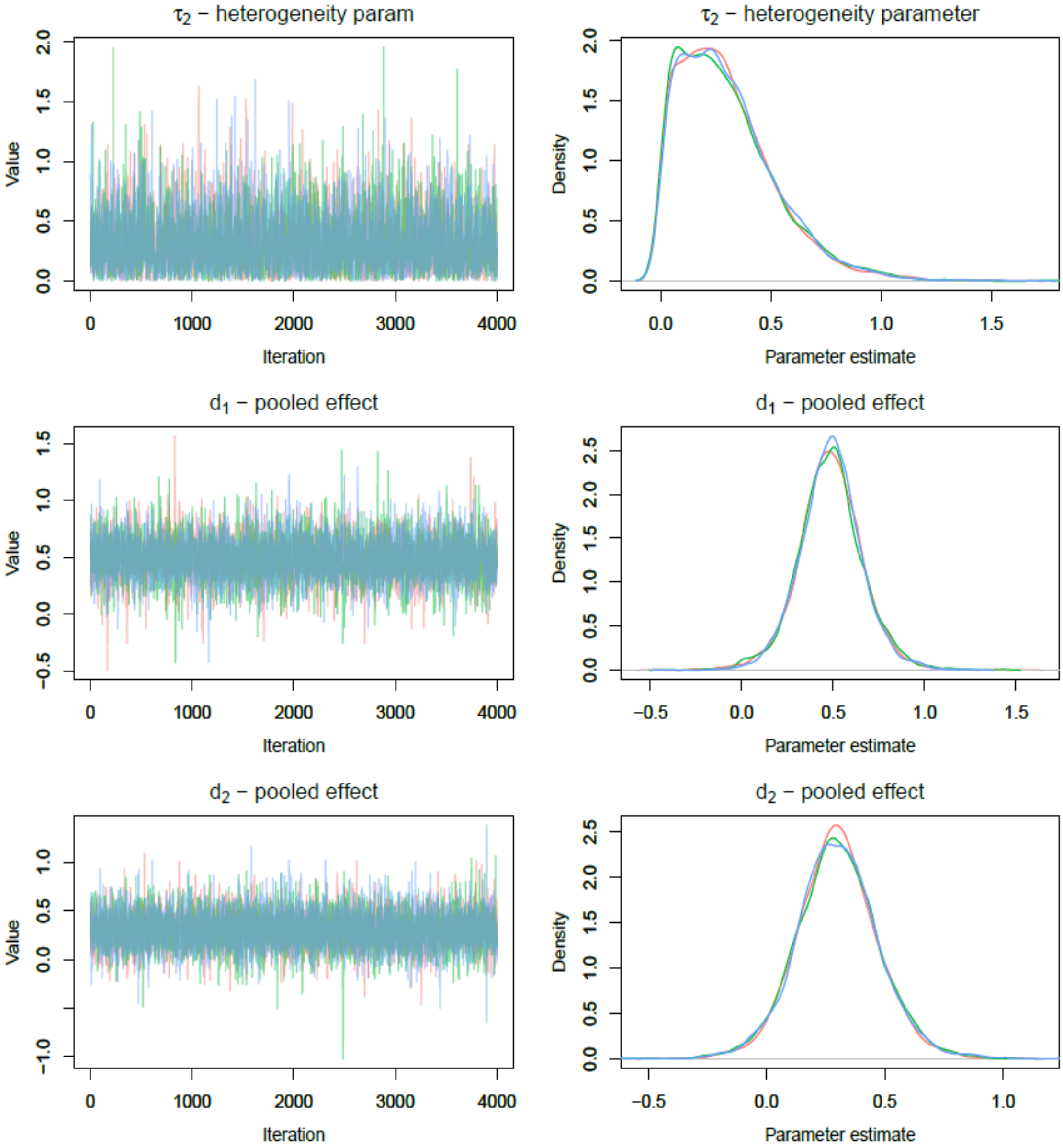


Figure C.8: Trace - density plots of 3 chains consisting of 5000 iterations each after 1000 iterations burn-in period



C.8.5 Convergence plots of BRMA-BC with Gumbel copula model

Figure C.9: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period

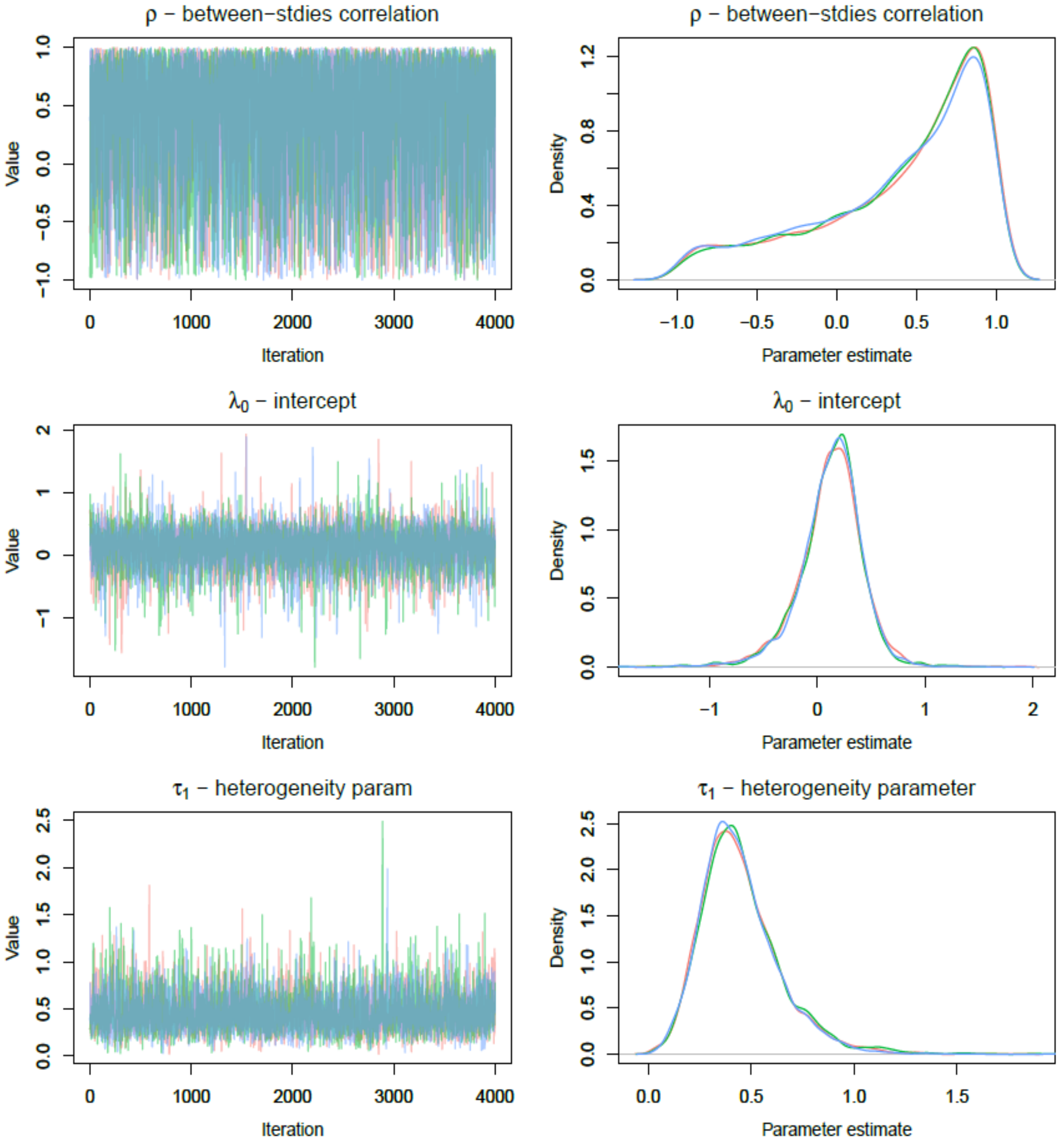
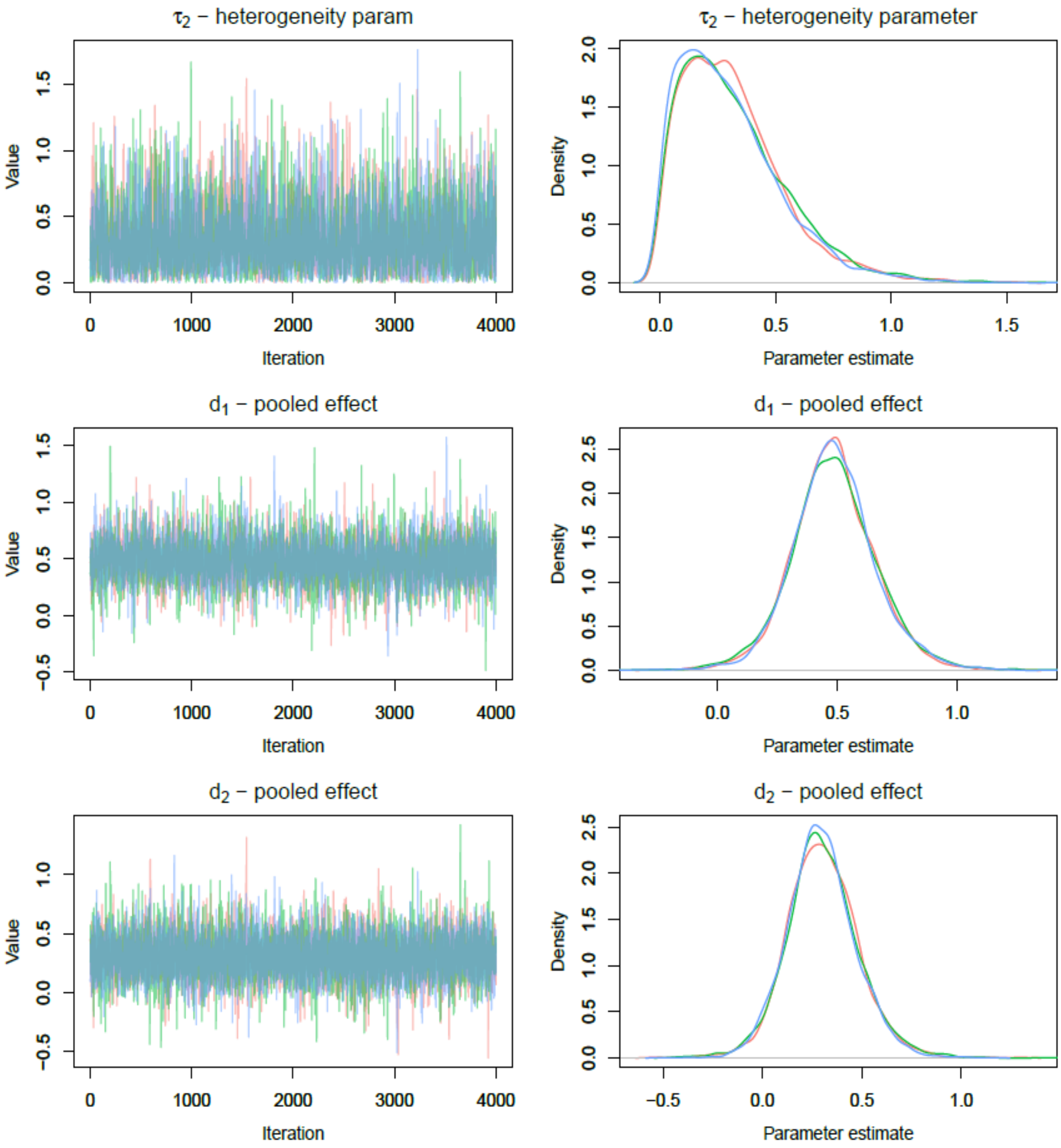


Figure C.10: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period



Appendix D

Appendix

D.1 Implementation of M1 model in Stan

```
data{
  int<lower = 0> Ns;//number of RCTs
  int<lower = 0> Ns2;//number of OBs for arm A
  int<lower = 0> Ns3;//number of OBs for arm B
  int<lower = 0> nA[Ns,2];//RCT data
  int<lower = 0> nB[Ns,2];//RCT data
  int<lower = 0> rA[Ns,2];//RCT data
  int<lower = 0> rB[Ns,2];//RCT data
  int<lower = 0> nAC[Ns2,2];// OB data for arm A
  int<lower = 0> rAC[Ns2,2];// OB data for arm A
  int<lower = 0> nBC[Ns3,2];// OB data for arm B
  int<lower = 0> rBC[Ns3,2];// OB data for arm B
  real rho_wA[Ns+Ns2];/within-study correlations
  real rho_wB[Ns+Ns3];/within-study correlations
}

transformed data{
//Calculate log odds and the corresponding variances
//for the RCT and OB data
vector[2] Ya[Ns];
vector[2] Yac[Ns2];
vector[2] Ybc[Ns3];
vector[2] Yb[Ns];
vector[2] Sa[Ns];
vector[2] Sac[Ns2];
vector[2] Sbc[Ns3];
vector[2] Sb[Ns];
for (i in 1:Ns){
Ya[i,1]=log((rA[i,1]+0.5)/((nA[i,1]-rA[i,1]+0.5)));
Ya[i,2]=log((rA[i,2]+0.5)/((nA[i,2]-rA[i,2]+0.5)));
Yb[i,1]=log((rB[i,1]+0.5)/((nB[i,1]-rB[i,1]+0.5)));
Yb[i,2]=log((rB[i,2]+0.5)/((nB[i,2]-rB[i,2]+0.5)));
Sa[i,1]=sqrt((1/(rA[i,1]+0.5))+(1/(nA[i,1]-rA[i,1]+0.5)));
Sa[i,2]=sqrt((1/(rA[i,2]+0.5))+(1/(nA[i,2]-rA[i,2]+0.5)));
Sb[i,1]=sqrt((1/(rB[i,1]+0.5))+(1/(nB[i,1]-rB[i,1]+0.5)));
Sb[i,2]=sqrt((1/(rB[i,2]+0.5))+(1/(nB[i,2]-rB[i,2]+0.5)));
}
for (i in 1:Ns2){
Yac[i,1]=log((rAC[i,1]+0.5)/((nAC[i,1]-rAC[i,1]+0.5)));
Yac[i,2]=log((rAC[i,2]+0.5)/((nAC[i,2]-rAC[i,2]+0.5)));
Sac[i,1]=sqrt((1/(rAC[i,1]+0.5))+(1/(nAC[i,1]-rAC[i,1]+0.5)));
Sac[i,2]=sqrt((1/(rAC[i,2]+0.5))+(1/(nAC[i,2]-rAC[i,2]+0.5)));
}
for (i in 1:Ns3){
Ybc[i,1]=log((rBC[i,1]+0.5)/((nBC[i,1]-rBC[i,1]+0.5)));
Ybc[i,2]=log((rBC[i,2]+0.5)/((nBC[i,2]-rBC[i,2]+0.5)));
Sbc[i,1]=sqrt((1/(rBC[i,1]+0.5))+(1/(nBC[i,1]-rBC[i,1]+0.5)));
Sbc[i,2]=sqrt((1/(rBC[i,2]+0.5))+(1/(nBC[i,2]-rBC[i,2]+0.5)));
}
}
```

```
parameters{
  real rr;
  real rr2;
  vector[2]  b;
  vector[2]  z[Ns+Ns2+Ns3];
  vector[2] k;
  vector<lower = 0, upper = 5>[2] s;
  vector[2] q[Ns+Ns2+Ns3];
  vector<lower = 0, upper = 5>[2] tau;
}

transformed parameters{
  matrix[2,2] Tau;
  matrix[2,2] L;
  matrix[2,2] Sig;
  matrix[2,2] H;
  matrix[2,2] SigmaA[Ns+Ns2];
  matrix[2,2] SigmaB[Ns+Ns3];
  vector[2] delta[Ns+Ns2+Ns3];
  vector[2] mu[Ns+Ns2+Ns3];
  real<lower= -1, upper=1> rho1;
  real<lower= -1, upper=1> rho2;
  rho1      = tanh(rr);
  rho2      = tanh(rr2);
  for (i in 1:(Ns)){
  SigmaA[i,1,1] = Sa[i,1]^2;
  SigmaA[i,1,2] = Sa[i,1]*Sa[i,2]*rho_wA[i];
  SigmaA[i,2,1] = Sa[i,1]*Sa[i,2]*rho_wA[i];
  SigmaA[i,2,2] = Sa[i,2]^2;
  SigmaB[i,1,1] = Sb[i,1]^2;
  SigmaB[i,1,2] = Sb[i,1]*Sb[i,2]*rho_wB[i];
  SigmaB[i,2,1] = Sb[i,1]*Sb[i,2]*rho_wB[i];
```

```
SigmaB[i,2,2] = Sb[i,2]^2;
}
for (i in 1:(Ns2)){
SigmaA[i+Ns,1,1] = Sac[i,1]^2;
SigmaA[i+Ns,1,2] = Sac[i,1]*Sac[i,2]*rho_wA[i+Ns];
SigmaA[i+Ns,2,1] = Sac[i,1]*Sac[i,2]*rho_wA[i+Ns];
SigmaA[i+Ns,2,2] = Sac[i,2]^2;
}
for (i in 1:(Ns3)){
SigmaB[i+Ns,1,1] = Sbc[i,1]^2;
SigmaB[i+Ns,1,2] = Sbc[i,1]*Sbc[i,2]*rho_wB[i+Ns];
SigmaB[i+Ns,2,1] = Sbc[i,1]*Sbc[i,2]*rho_wB[i+Ns];
SigmaB[i+Ns,2,2] = Sbc[i,2]^2;
}
Sig[1, 1] = s[1]^2;
Sig[1, 2] = s[1]*s[2]*rho2;
Sig[2, 1] = s[1]*s[2]*rho2;
Sig[2, 2] = s[2]^2;
H = cholesky_decompose(Sig);
Tau[1,1]       = tau[1]^2;
Tau[2,2]       = tau[2]^2;
Tau[1,2]       = tau[1]*tau[2]*rho1;
Tau[2,1]       = tau[1]*tau[2]*rho1;
L              = cholesky_decompose(Tau);
//non-centred parameterisation for mus and deltas
for (i in 1:(Ns)){
mu[i]    = k + H*q[i];
delta[i] = b + (L*z[i]);}
for (i in 1:Ns2){
mu[i+Ns]    = k + H*q[i+Ns];
delta[i+Ns] = b + L*z[i+Ns];}
for (i in 1:Ns3){
  mu[i+Ns+Ns2]    = k + H*q[i+Ns+Ns2];
  delta[i+Ns+Ns2] = b + L*z[i+Ns+Ns2];}
}

model{
//priors
  rr          ~ std_normal();
  rr2         ~ std_normal();
  k           ~ normal(0, 10);
  b           ~ normal(0, 10);
  for (i in 1:Ns){
  q[i]        ~ std_normal();
  z[i]        ~ std_normal();
//likelihood of the RCTs
  Ya[i]       ~ multi_normal(mu[i],SigmaA[i]);
  Yb[i]       ~ multi_normal(mu[i]+delta[i],SigmaB[i]);}
  for (i in 1:Ns2){
  q[i+Ns]     ~ std_normal();
  z[i+Ns]     ~ std_normal();
//likelihood of the OBs for arm A
  Yac[i]      ~ multi_normal(mu[i+Ns],SigmaA[i+Ns]);}
  for (i in 1:Ns3){
  z[i+Ns+Ns2] ~ std_normal();
  q[i+Ns+Ns2] ~ std_normal();
//likelihood of the OBs for arm B
  Ybc[i]       ~ multi_normal(mu[i+Ns+Ns2]+delta[i+Ns+Ns2],SigmaB[i+Ns]);}
}
```

D.2 Implementation of M2 model in Stan

```
data{
  int<lower = 0> Ns;//number of RCTs
  int<lower = 0> Ns2;//number of OBs for arm A
  int<lower = 0> Ns3;//number of OBs for arm B
  int<lower = 0> nA[Ns,2];//RCT data
  int<lower = 0> nB[Ns,2];//RCT data
  int<lower = 0> rA[Ns,2];//RCT data
  int<lower = 0> rB[Ns,2];//RCT data
  int<lower = 0> nAC[Ns2,2];// OB data for arm A
  int<lower = 0> rAC[Ns2,2];// OB data for arm A
  int<lower = 0> nBC[Ns3,2];// OB data for arm B
  int<lower = 0> rBC[Ns3,2];// OB data for arm B
  real rho_wA[Ns+Ns2];/within-study correlations
  real rho_wB[Ns+Ns3];/within-study correlations
}


transformed data{
//Calculate log odds and the corresponding variances
//for the RCT and OB data
  vector[2] Ya[Ns];
  vector[2] Yac[Ns2];
  vector[2] Ybc[Ns3];
  vector[2] Yb[Ns];
  vector[2] Sa[Ns];
  vector[2] Sac[Ns2];
  vector[2] Sbc[Ns3];
  vector[2] Sb[Ns];
  for (i in 1:Ns){
  Ya[i,1]=log((rA[i,1]+0.5)/((nA[i,1]-rA[i,1]+0.5)));
  Ya[i,2]=log((rA[i,2]+0.5)/((nA[i,2]-rA[i,2]+0.5)));
  Yb[i,1]=log((rB[i,1]+0.5)/((nB[i,1]-rB[i,1]+0.5)));
  Yb[i,2]=log((rB[i,2]+0.5)/((nB[i,2]-rB[i,2]+0.5)));
  Sa[i,1]=sqrt((1/(rA[i,1]+0.5))+(1/(nA[i,1]-rA[i,1]+0.5)));
  Sa[i,2]=sqrt((1/(rA[i,2]+0.5))+(1/(nA[i,2]-rA[i,2]+0.5)));
  Sb[i,1]=sqrt((1/(rB[i,1]+0.5))+(1/(nB[i,1]-rB[i,1]+0.5)));
  Sb[i,2]=sqrt((1/(rB[i,2]+0.5))+(1/(nB[i,2]-rB[i,2]+0.5)));
  }
  for (i in 1:Ns2){
  Yac[i,1]=log((rAC[i,1]+0.5)/((nAC[i,1]-rAC[i,1]+0.5)));
  Yac[i,2]=log((rAC[i,2]+0.5)/((nAC[i,2]-rAC[i,2]+0.5)));
  Sac[i,1]=sqrt((1/(rAC[i,1]+0.5))+(1/(nAC[i,1]-rAC[i,1]+0.5)));
  Sac[i,2]=sqrt((1/(rAC[i,2]+0.5))+(1/(nAC[i,2]-rAC[i,2]+0.5)));
  }
  for (i in 1:Ns3){
  Ybc[i,1]=log((rBC[i,1]+0.5)/((nBC[i,1]-rBC[i,1]+0.5)));
  Ybc[i,2]=log((rBC[i,2]+0.5)/((nBC[i,2]-rBC[i,2]+0.5)));
  Sbc[i,1]=sqrt((1/(rBC[i,1]+0.5))+(1/(nBC[i,1]-rBC[i,1]+0.5)));
  Sbc[i,2]=sqrt((1/(rBC[i,2]+0.5))+(1/(nBC[i,2]-rBC[i,2]+0.5)));
  }
  }

parameters{
  real rr;
  real rr2;
  vector[2]  b;
  vector[2]  z[Ns+Ns2+Ns3];
  vector[2] k;
  vector[2] eta;//Bias term
  vector[2] xi;//Bias term
  vector<lower = 0, upper = 5>[2] s;
  vector[2] q[Ns+Ns2+Ns3];
  vector<lower = 0, upper = 5>[2] tau;
```

```
}

transformed parameters{
  matrix[2,2] Tau;
  matrix[2,2] L;
  matrix[2,2] Sig;
  matrix[2,2] H;
  matrix[2,2] SigmaA[Ns+Ns2];
  matrix[2,2] SigmaB[Ns+Ns3];
  vector[2] delta[Ns+Ns2+Ns3];
  vector[2] mu[Ns+Ns2+Ns3];
  real<lower= -1, upper=1> rho1;
  real<lower= -1, upper=1> rho2;
  rho1      = tanh(rr);
  rho2      = tanh(rr2);
  for (i in 1:(Ns)){
  SigmaA[i,1,1] = Sa[i,1]^2;
  SigmaA[i,1,2] = Sa[i,1]*Sa[i,2]*rho_wA[i];
  SigmaA[i,2,1] = Sa[i,1]*Sa[i,2]*rho_wA[i];
  SigmaA[i,2,2] = Sa[i,2]^2;
  SigmaB[i,1,1] = Sb[i,1]^2;
  SigmaB[i,1,2] = Sb[i,1]*Sb[i,2]*rho_wB[i];
  SigmaB[i,2,1] = Sb[i,1]*Sb[i,2]*rho_wB[i];
  SigmaB[i,2,2] = Sb[i,2]^2;
  }
  for (i in 1:(Ns2)){
  SigmaA[i+Ns,1,1] = Sac[i,1]^2;
  SigmaA[i+Ns,1,2] = Sac[i,1]*Sac[i,2]*rho_wA[i+Ns];
  SigmaA[i+Ns,2,1] = Sac[i,1]*Sac[i,2]*rho_wA[i+Ns];
  SigmaA[i+Ns,2,2] = Sac[i,2]^2;
  }
  for (i in 1:(Ns3)){
```

```
  SigmaB[i+Ns,1,1] = Sbc[i,1]^2;
  SigmaB[i+Ns,1,2] = Sbc[i,1]*Sbc[i,2]*rho_wB[i+Ns];
  SigmaB[i+Ns,2,1] = Sbc[i,1]*Sbc[i,2]*rho_wB[i+Ns];
  SigmaB[i+Ns,2,2] = Sbc[i,2]^2;
  }
  Sig[1, 1] = s[1]^2;
  Sig[1, 2] = s[1]*s[2]*rho2;
  Sig[2, 1] = s[1]*s[2]*rho2;
  Sig[2, 2] = s[2]^2;
  H = cholesky_decompose(Sig);
  Tau[1,1]       = tau[1]^2;
  Tau[2,2]       = tau[2]^2;
  Tau[1,2]       = tau[1]*tau[2]*rho1;
  Tau[2,1]       = tau[1]*tau[2]*rho1;
  L              = cholesky_decompose(Tau);
  //non-centred parameterisation for mus and deltas
  for (i in 1:(Ns)){
  mu[i]    = k + H*q[i];
  delta[i] = b + (L*z[i]);}
  for (i in 1:Ns2){
  mu[i+Ns]    = k + H*q[i+Ns];
  delta[i+Ns] = b + L*z[i+Ns];}
  for (i in 1:Ns3){
  mu[i+Ns+Ns2]    = k + H*q[i+Ns+Ns2];
  delta[i+Ns+Ns2] = b + L*z[i+Ns+Ns2];}
}

model{
//priors
  rr         ~ std_normal();
  rr2        ~ std_normal();
  k          ~ normal(0, 10);
  b          ~ normal(0, 10);
  eta        ~ normal(0, 10);
  xi         ~ normal(0, 10);
  for (i in 1:Ns){
  q[i]       ~ std_normal();
  z[i]       ~ std_normal();
//likelihood of the RCTs
  Ya[i]      ~ multi_normal(mu[i],SigmaA[i]);
  Yb[i]      ~ multi_normal(mu[i]+delta[i],SigmaB[i]);}
  for (i in 1:Ns2){
  q[i+Ns]    ~ std_normal();
  z[i+Ns]    ~ std_normal();
//likelihood of the OBs for arm A with bias term
  Yac[i]     ~ multi_normal(mu[i+Ns]+eta,SigmaA[i+Ns]);}
  for (i in 1:Ns3){
  z[i+Ns+Ns2] ~ std_normal();
  q[i+Ns+Ns2] ~ std_normal();
//likelihood of the OBs for arm B with bias term
  Ybc[i]      ~ multi_normal(mu[i+Ns+Ns2]+delta[i+Ns+Ns2]+xi,SigmaB[i+Ns]);}
}
```

D.3 Implementation of M3 model in Stan

```
//Bivariate Binomial density with frank_copula
functions {
//Gaussian copula CDF
 real fcop3(real theta, real u1, real u2){
    real t1 = u1;real z1;
    real t2 = u2;real z2;
    if (t1 > 0.9999999) t1 = 0.9999999;
    if (t2 > 0.9999999) t2 = 0.9999999;
    z1 = inv_Phi(t1);
    z2 = inv_Phi(t2);
    if (z1 != 0 || z2 != 0) {
      real denom = fabs(theta) < 1.0 ? sqrt((1 + theta) *
                   (1 - theta)) : not_a_number();
      real a1 = (z2 / z1 - theta) / denom;
      real a2 = (z1 / z2 - theta) / denom;
      real product = z1 * z2;
      real delta = product < 0 || (product == 0 && (z1 + z2) < 0);
      real a = log(0.5 * (Phi(z1) + Phi(z2) - delta) -
               owens_t(z1, a1) - owens_t(z2, a2));
      return exp(a);
    }
    if (theta == 1){
      vector[2] z;
      z[1]=z1;z[2]=z2;
      return min(Phi(z));
    }
    return 0.25 + asin(theta) / (2 * pi());
  }

real Bivfcop_lpmf(int[] r,int n1, int n2, real theta, vector mu){
   real f1;
   real f2;
   real f11;
   real f12;
   real prob;
   f11 = binomial_cdf(r[1]-1, n1, inv_logit(mu[1]));
   f12 = binomial_cdf(r[2]-1, n2, inv_logit(mu[2]));
   f1 = f11 + exp(binomial_logit_lpmf(r[1] |n1, mu[1]));
   f2 = f12 + exp(binomial_logit_lpmf(r[2] |n2, mu[2]));
   prob = fcop3(theta,f1,f2)-fcop3(theta,f1,f12)-
          fcop3(theta,f11,f2)+fcop3(theta,f11,f12);
   return log(prob);}


   }

data{
  int<lower = 0> Ns;//number of RCTs
  int<lower = 0> Ns2;//number of OBs for arm A
  int<lower = 0> Ns3;//number of OBs for arm B
  int<lower = 0> nA[Ns,2];//RCT data
  int<lower = 0> nB[Ns,2];//RCT data
  int<lower = 0> rA[Ns,2];//RCT data
  int<lower = 0> rB[Ns,2];//RCT data
  int<lower = 0> nAC[Ns2,2];// OB data for arm A
  int<lower = 0> rAC[Ns2,2];// OB data for arm A
  int<lower = 0> nBC[Ns3,2];// OB data for arm B
  int<lower = 0> rBC[Ns3,2];// OB data for arm B
  real theta1[Ns+Ns2];
  real theta2[Ns+Ns3];
}
```

```
parameters{
  real rr;
  real rr2;
  vector[2] b;
  vector[2] z[Ns+Ns2+Ns3];
  vector[2] k;
  vector[2] q[Ns+Ns2+Ns3];
  vector[2] eta;//bias terms
  vector[2] xi;//bias terms
  vector<lower=0, upper=5>[2] tau;
  vector<lower = 0, upper=5>[2] s;
}

transformed parameters{
  matrix[2,2] Tau;
  matrix[2,2] L;
  matrix[2,2] Sig;
  matrix[2,2] H;
  vector[2] delta[Ns+Ns2+Ns3];
  vector[2] mu[Ns+Ns2+Ns3];
  real<lower  = -1, upper=1> rho1;
  real<lower  = -1, upper=1> rho2;
  rho1        = tanh(rr);
  rho2        = tanh(rr2);
  Sig[1, 1]   = s[1]^2;
  Sig[1, 2]   = s[1]*s[2]*rho2;
  Sig[2, 1]   = s[1]*s[2]*rho2;
  Sig[2, 2]   = s[2]^2;
  H = cholesky_decompose(Sig);
  Tau[1, 1]   = tau[1]^2;
  Tau[1, 2]   = tau[1]*tau[2]*rho1;
  Tau[2, 1]   = tau[1]*tau[2]*rho1;
  Tau[2, 2]   = tau[2]^2;
  L = cholesky_decompose(Tau);
  for (i in 1:Ns){
  mu[i]       = k + H*q[i];
  delta[i]    = b + L*z[i];}
  for (i in 1:Ns2){
  mu[i+Ns]    = k + H*q[i+Ns];
  delta[i+Ns] = b + L*z[i+Ns];}
  for (i in 1:Ns3){
  mu[i+Ns+Ns2]   = k + H*q[i+Ns+Ns2];
  delta[i+Ns+Ns2]= b + L*z[i+Ns+Ns2];}
}

model{
//priors
  rr         ~ std_normal();
  rr2        ~ std_normal();
  b          ~ normal(0, 10);
  k          ~ normal(0, 10);
  eta        ~ normal(0, 10);
  xi         ~ normal(0, 10);
  for (i in 1:Ns){
  z[i]       ~ std_normal();
  q[i]       ~ std_normal();
//likelihood of the RCTs
  rA[i]      ~ Bivfcop(nA[i,1],nA[i,2], theta1, mu[i]);
  rB[i]      ~ Bivfcop(nB[i,1],nB[i,2], theta2, mu[i]+delta[i]);}
  for (i in 1:Ns2){
  q[i+Ns]    ~ std_normal();
  z[i+Ns]    ~ std_normal();
//likelihood of the OBs for arm A with bias term
  rAC[i]     ~ Bivfcop(nAC[i,1],nAC[i,2],theta1, mu[i+Ns]+eta);}
  for (i in 1:Ns3){
  q[i+Ns+Ns2]~ std_normal();
  z[i+Ns+Ns2]~ std_normal();
//likelihood of the OBs for arm B with bias term
  rBC[i]     ~ Bivfcop(nBC[i,1],nBC[i,2],theta2,
                       mu[i+Ns+Ns2]+xi+delta[i+Ns+Ns2]);}
}
```

D.4 Convergence plots of M2 and M3 models

The following figures display trace and density plots of the parameters between studies parameters including the parameter of the intercept (λ_0) in the aCRC data-set discussed in section 6.5.1.

D.4.1 Convergence plots of M2 model

Figure D.1: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period

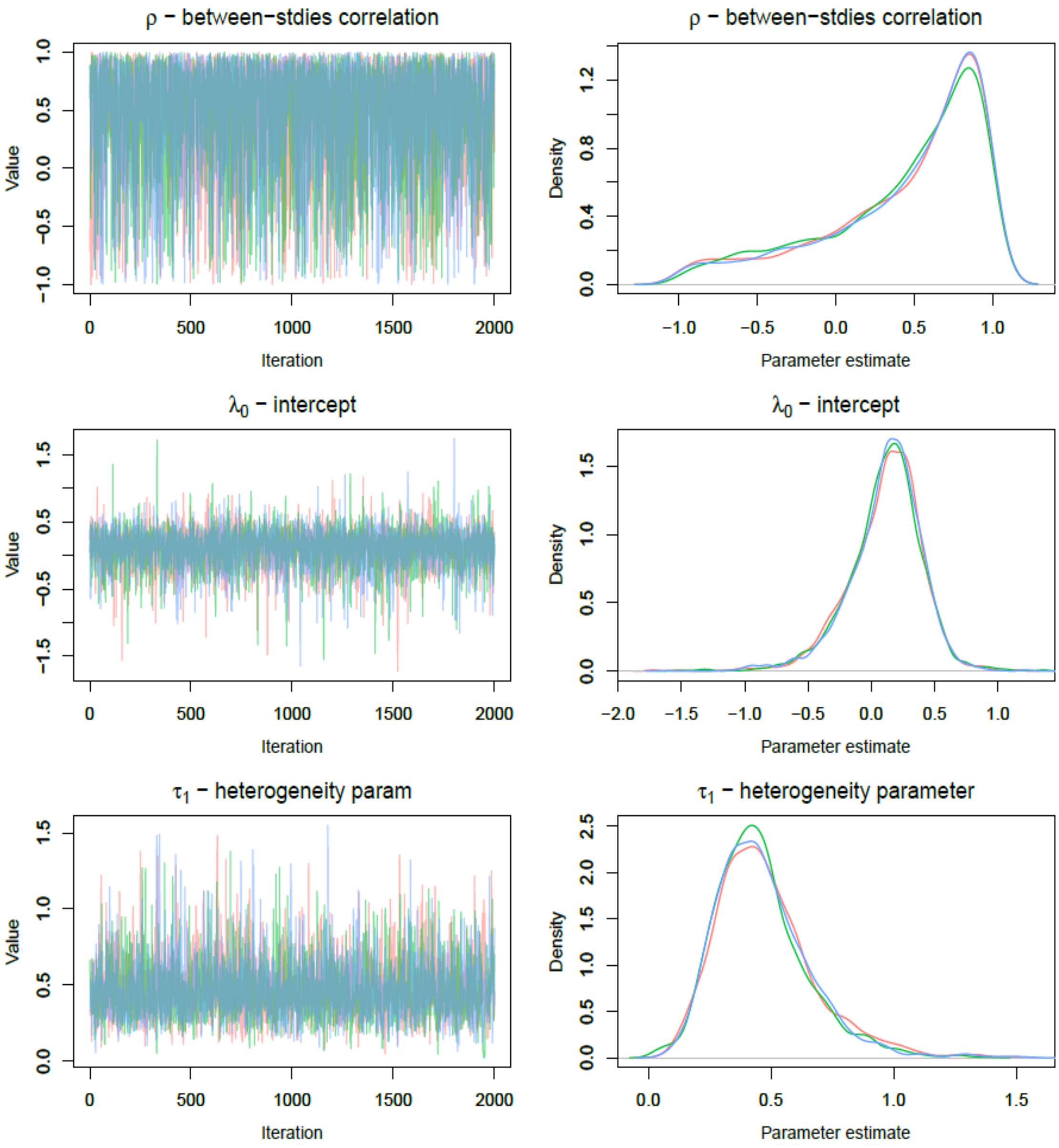
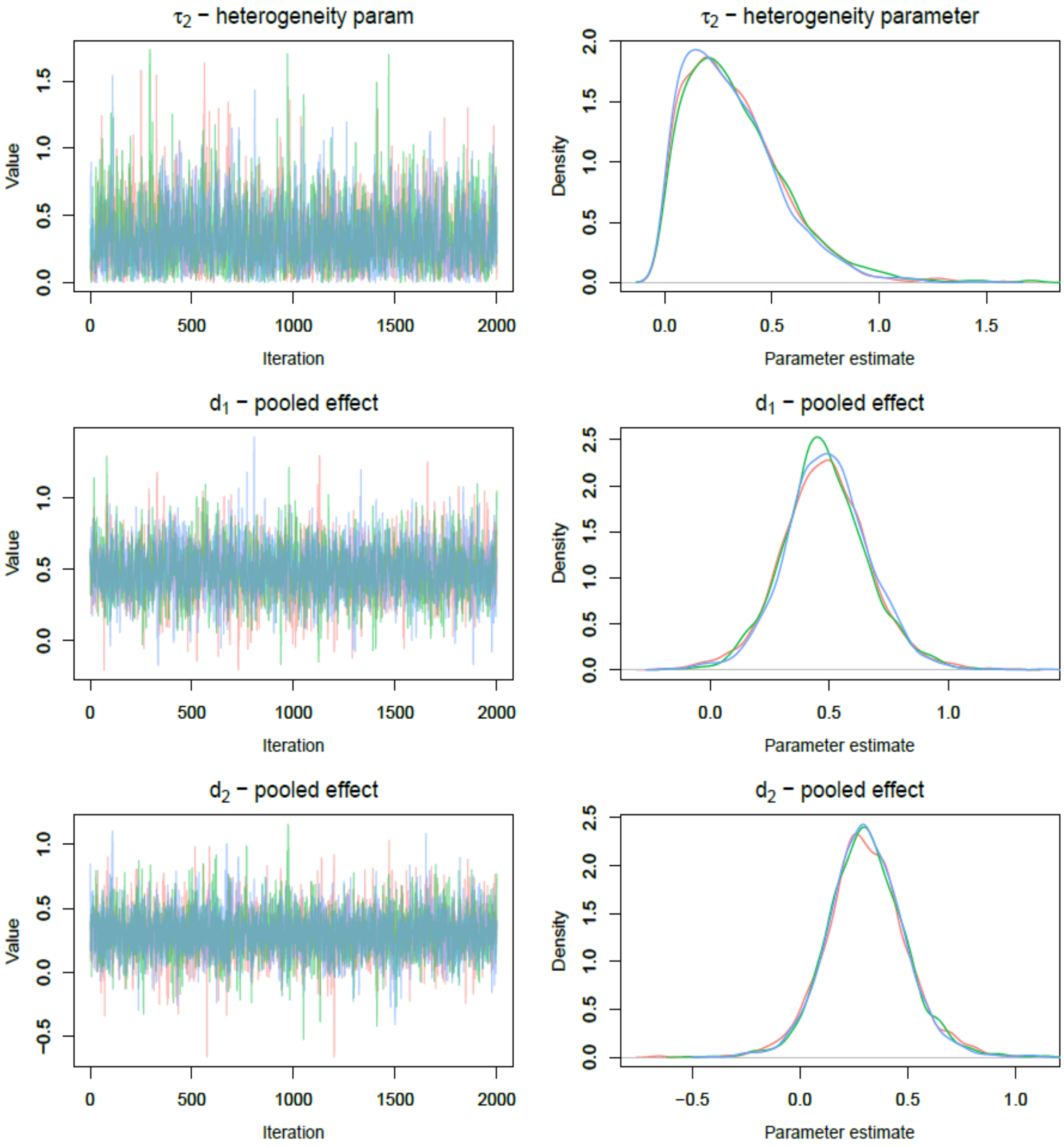


Figure D.2: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period



D.4.2 Convergence plots of M3 model

Figure D.3: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period

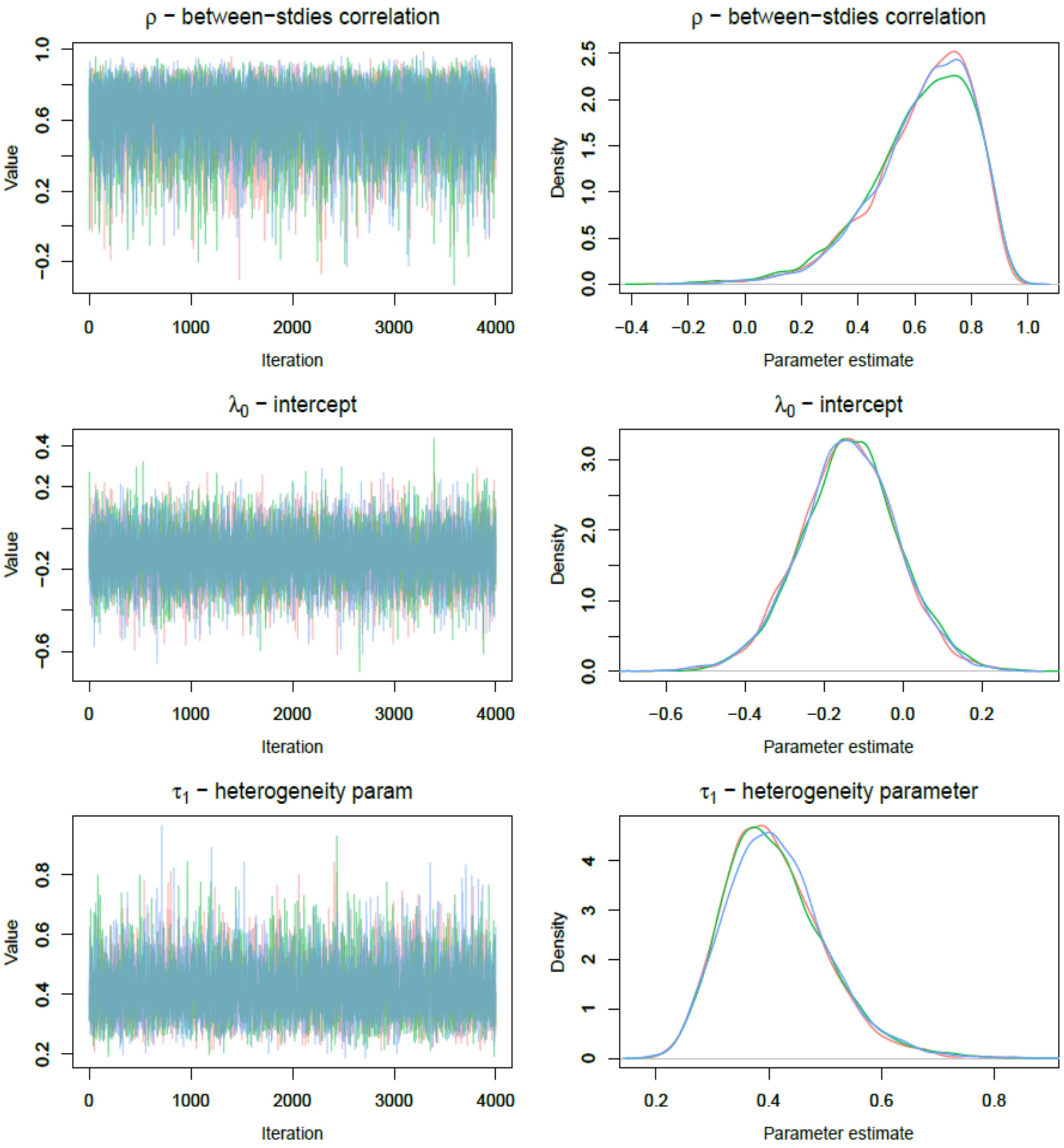
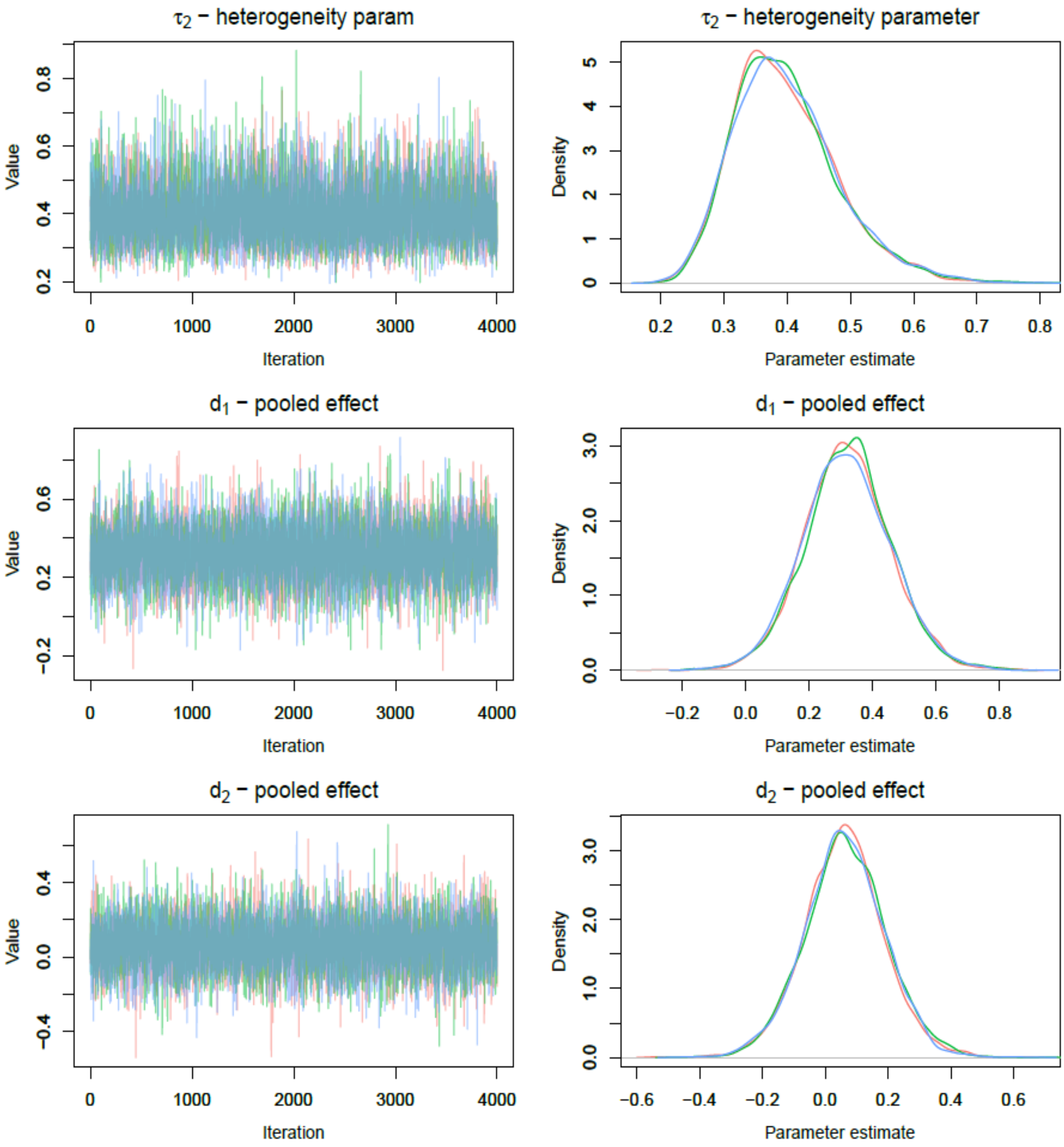


Figure D.4: Trace - density plots of 3 chains consisting of 4000 iterations each after 1000 iterations burn-in period



Appendix E

Appendix

E.1 Statistics in Medicine paper

Received: 14 December 2018 | Revised: 10 September 2019 | Accepted: 13 December 2019

DOI: 10.1002/sim.8465

RESEARCH ARTICLE

Statistics in Medicine WILEY

Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data

Tasos Papanikos[1] | John R. Thompson[2] | Keith R. Abrams[1] | Nicolas Städler[3] | Oriana Ciani[4,5] | Rod Taylor[4,6] | Sylwia Bujkiewicz[1]

[1]Biostatistics Group, Department of Health Sciences, University of Leicester, Leicester, UK

[2]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, UK

[3]Roche Innovation Center, F. Hoffmann-La Roche Ltd, Basel, Switzerland

[4]College of Medicine and Health, University of Exeter Medical School, Exeter, UK

[5]Centre for Research on Health and Social Care Management, SDA Bocconi University, Milan, Italy

[6]MRC/CSO Social and Public Health Sciences Unit & Robertson Centre for Biostatistics, University of Glasgow, Glasgow, UK

Correspondence
Tasos Papanikos, Biostatistics Research Group, Department of Health Sciences, University of Leicester, George Davies Centre University Road, Leicester LE1 7RH, UK.
Email: ap659@leicester.ac.uk

Funding information
Medical Research Council, MR/L009854/1

Surrogate endpoints play an important role in drug development when they can be used to measure treatment effect early compared to the final clinical outcome and to predict clinical benefit or harm. Such endpoints are assessed for their predictive value of clinical benefit by investigating the surrogate relationship between treatment effects on the surrogate and final outcomes using meta-analytic methods. When surrogate relationships vary across treatment classes, such validation may fail due to limited data within each treatment class. In this paper, two alternative Bayesian meta-analytic methods are introduced which allow for borrowing of information from other treatment classes when exploring the surrogacy in a particular class. The first approach extends a standard model for the evaluation of surrogate endpoints to a hierarchical meta-analysis model assuming full exchangeability of surrogate relationships across all the treatment classes, thus facilitating borrowing of information across the classes. The second method is able to relax this assumption by allowing for partial exchangeability of surrogate relationships across treatment classes to avoid excessive borrowing of information from distinctly different classes. We carried out a simulation study to assess the proposed methods in nine data scenarios and compared them with subgroup analysis using the standard model within each treatment class. We also applied the methods to an illustrative example in colorectal cancer which led to obtaining the parameters describing the surrogate relationships with higher precision.

KEYWORDS

hierarchical models, meta-analysis, partial exchangeability, surrogate endpoints, treatment classes

1 | INTRODUCTION

New advances in science have led to discovering of promising therapies which often are targeted to specific patient populations, for example, defined by a genetic biomarker. This leads to clinical trials of smaller size, while the increased effectiveness of these therapies reduces the number of events or deaths and consequently lead to measurement of treatment effect on overall survival (OS) with large uncertainty. Therefore, surrogate endpoints allowing the measurement of treatment effect with higher precision have been investigated to accelerate the availability of these treatments to the patients. These alternative endpoints often can be considered a cost-effective replacement of final clinical outcome, as they are particularly useful when they can be measured earlier, easier, more frequently compared to the final clinical endpoint or if they require smaller sample size and shorter follow up times. [1]

Potential surrogate endpoints have been investigated as candidate endpoints in clinical trials in a number of disease areas. However, before these candidate endpoints are used, either as primary endpoints in trial design or in regulatory decision-making, they need to be validated.[2] In practice, the most common approach to validate a candidate outcome is to examine whether it satisfies three levels of association, proposed by the International Conference on Harmonisation Guidelines on Statistical Principles for Clinical Trials.[3] First, the biological plausibility of the association of the surrogate and final outcomes is investigated which involves biological rather than statistical considerations. Furthermore, the individual-level association is evaluated to establish whether the candidate surrogate endpoint can be used to predict the course of the disease in an individual patient. Last but not least, the study-level association is investigated to ensure that the treatment effects on the final outcome can be predicted from the effect on the surrogate endpoint. Study-level association requires data from a number of randomized controlled trials (RCTs) and can be investigated carrying out a bivariate meta-analysis.[4-7] In this paper we focus on the third level of association only.

A bivariate meta-analytical method that was developed by Daniels and Hughes[4] can be used to validate a candidate surrogate endpoint, by evaluating the association pattern between the treatment effects on the surrogate and the final outcomes, and to predict treatment effects on the final clinical outcome from the effects on surrogate endpoint. This method, implemented in a Bayesian framework, can be used to evaluate a surrogate endpoint in a disease area overall, or in each treatment class separately through a subgroup analysis.

Traditionally, surrogate relationships between treatment effects on a surrogate endpoint and treatment effects on a final outcome have been investigated in a disease area using data from all trials regardless of treatment classes or trials of the same class of treatments. For instance, in advance colorectal cancer (aCRC) progression-free survival (PFS), tumor response (TR) or time to progression have been investigated as potential surrogate endpoints for OS.[8-11] In previous work, Buyse et al[8] found a strong association between treatment effects on PFS and OS in this disease area, by including in their meta-analysis studies on one treatment class only (modern chemotherapy). More recently, Ciani et al[10] investigated the surrogate relationship in aCRC across all modern treatments, including a range of targeted therapies, which led to suboptimal surrogate relationship in this disease area. They concluded that in aCRC the association patterns could vary across treatment classes and a surrogate relationship observed in a specific treatment class may not directly apply across other treatment classes or lines of treatment. This may be particularly important for targeted treatments used only in a subset of population. For example anti-EGFR treatments are recommended for patients without a KRAS/panRAS mutation as these mutations are associated with resistance to the anti-EGFR therapies[12,13] and the association pattern might be different for this particular treatment class in this subset of population with this unique characteristic. Furthermore, Giessen et al[9] who investigated the surrogate relationships in aCRC including all available treatments and subgroups of therapies, inferred that for validation of surrogacy in targeted treatments such as anti-EGFR therapies or anti-VEGF treatments further research is required once more data become available. Consequently, the assumption that a surrogate relationship remains the same across different treatment classes or lines of treatment does not seem reasonable in aCRC, which may be the case in other disease areas. Therefore, potential differences in surrogate relationships across classes should be investigated. This can be achieved by performing subgroup analysis using a standard model (eg, Daniels and Hughes model[4]) or extending the standard model by adding another level to the hierarchical structure of the model for a surrogate relationship accounting for differences between treatment classes. In this paper, we propose two new methods which allow different degrees of borrowing of information for surrogate relationships across treatment classes aiming to obtain estimates of surrogate relationships with higher precision.[14-16] The first approach assumes full exchangeability of the parameters describing the surrogate relationships exploiting the similarity of surrogate relationships and borrowing information across treatment classes. The second method is able to relax this assumption, by allowing for partial exchangeability[17] of surrogate relationships across treatment classes to avoid excessive borrowing of information from distinctly different treatment classes. In this model, the parameters describing surrogate relationships can be either exchangeable or nonexchangeable giving more flexibility when the assumption of exchangeability is not reasonable.

The modeling techniques were demonstrated using an example in advanced colorectal cancer where the surrogate relationships may vary across treatment classes.[10] To assess models' performance and compare them with subgroup analysis we carried out a simulation study. In the remainder of this paper, we present the standard model in Section 2 , the two proposed models are introduced in Section 3, the results of the simulation study are demonstrated in Section 5 and the illustrative example as well as the results from its analysis are presented in Section 6. The paper concludes with a discussion in Section 7.

2 | STANDARD SURROGACY MODEL

To investigate surrogate relationships within treatment classes using aggregate data, we performed subgroup analysis adopting a standard surrogacy model that was introduced by Daniels and Hughes[4] for the study-level evaluation of potential surrogate markers. Equation (1) corresponds to the within-study model where Y_{1i}, Y_{2i} are the estimates of treatment effects on surrogate endpoint and on the final outcome (eg, log odds ratios for TR and log hazard ratio for OS). These effects follow a bivariate normal distribution with μ_{1i} and μ_{2i} corresponding to the true treatment effects on the surrogate and the final clinical outcome, respectively, while, σ_{1i}, σ_{2i}, and ρ_{wi} are the within-study SDs for both outcomes and the within-study correlations between the treatment effects on the two outcomes for each study i.

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \sigma_{1i}\sigma_{2i}\rho_{wi} \\ \sigma_{1i}\sigma_{2i}\rho_{wi} & \sigma_{2i}^2 \end{pmatrix} \right). \tag{1}$$

$$\mu_{2i}|\mu_{1i} \sim N(\lambda_0 + \lambda_1\mu_{1i}, \psi^2). \tag{2}$$

At the between-studies level (2), the true effects on the surrogate endpoint μ_{1i} are modeled as fixed effects, and the true effects on the final outcome μ_{2i} have linear relationship with the true effects on the surrogate μ_{1i}. This relationship plays a very important role as it can be used to predict μ_{2i} from known μ_{1i} in a new study i. The parameters λ_0, λ_1, and ψ^2 correspond to the intercept, the slope, and the conditional variance of the linear model and measure the shape of the relationship and the strength of association between the treatment effects on the surrogate endpoint and the effects on the final outcome.

In the Bayesian framework, the Daniels and Hughes model was implemented by assuming no prior knowledge about surrogate relationship by using vague prior distributions. This allows the data to dominate the posterior distribution even if the dataset is relatively small. The following prior distributions can be used: $\mu_{1i} \sim N(0, a)$, $\lambda_0 \sim N(0, a)$, $\lambda_1 \sim N(0, a)$, $\psi \sim N(0, b)I(0,)$, where $N(0, b)I(0,)$ denotes a normal distribution truncated[18] at the mean $\mu = 0$ with SD $s = b$. The parameters a, b are chosen to be sufficiently large and depend on the scale of data.

By adapting this method in our research, we applied this standard model to subsets of data that consist of only one class of treatment examining the surrogate relationship of each subgroup separately, taking motivation from similar analyses in clinical trials.[19,20] This kind of analysis is very practical when association patterns in a given disease area are different and the treatment classes consist of many studies. By performing subgroup analysis using the standard model, we explored potential differences in the association patterns across treatment classes and use them as a reference for results obtained with the newly developed methods.

2.1 | Criteria for surrogacy

As we mentioned previously, the parameters λ_0, λ_1, ψ^2 play a very important role, as they are used to evaluate surrogacy. A good surrogate relationship should imply that $\lambda_1 \neq 0$ as slope establishes the association between treatment effects on the surrogate and the final outcome. Subsequently, having $\psi^2 = 0$ implies that μ_{2i} could be perfectly predicted given μ_{1i}. The parameter λ_0 corresponds to the intercept and is expected to be zero for a good surrogate relationship. This ensures that no treatment effect on the surrogate endpoint will imply no effect on the final outcome. These three criteria proposed by Daniels & Hughes,[4] will be referred to as surrogacy criteria in the remainder of this paper. A simple way to examine these surrogacy criteria is to check whether or not zero is included in the 95% credible intervals (CrIs) of λ_0, λ_1 and to compute the Bayes factor for the hypothesis H_1: $\psi^2 = 0$. The model with $\psi^2 = 0$ is a nested model within the standard model,[21] so in order to compare these models, Bayes factors can be computed using the Savage Dickey density ratio.[22] To implement the Savage Dickey density ratio, proper prior distributions for ψ are needed. In our research a moderately informative half normal prior distribution $N(0, 2)I(0,)$ was used for the conditional SD. A strong association pattern requires zero to be included in the CrI of λ_0, zero not to be included in the CrI of λ_1 and the Bayes factor of ψ^2 to be greater than 3.3.[23] In this paper we used the evaluation framework proposed by Daniels and Hughes. However, there are other definitions and criteria for surrogacy in the literature. Detail review of other evaluation frameworks can be found in Lassere et al.[7]

2.2 | Cross-validation

One of the main aims of this paper was to explore whether the two hierarchical methods, that we propose in the next section, improve the predictions of treatment effect on the final outcome (by reducing bias and/or uncertainty) compared to subgroup analysis using the standard model. To evaluate this, a cross-validation procedure was carried out. It is a similar to the "leave-one-out" procedure described by Daniels and Hughes[4] and it is repeated as many times as the number of studies in the dataset. In a simulated data scenario, this can be used to draw inferences about predicting the true effect on the final endpoint μ_{2i} in a "new" study i; however, in a real-data scenario true effects are unknown and therefore, we can only compare the observed values Y_{2i} with their predicted intervals. For each study i ($i = 1, \dots , N$), treatment effect on the final endpoint Y_{2i} is omitted and assumed unknown. This effect is then predicted from the observed effect on the surrogate endpoint Y_{1i} and by taking into account the treatment effects on both outcomes from the remaining studies. In a Bayesian framework it can be achieved by performing Markov chain Monte Carlo (MCMC) simulation. The mean predicted effect is equal to the true effect $\hat{\mu}_{2i}$ predicted by MCMC simulation and the variance of the predicted effect is equal to $\sigma^2_{2i} + var(\hat{\mu}_{2i}|Y_{1i}, \sigma_{1i}, Y_{1(-i)}, Y_{2(-i)})$ where $Y_{1,2(-i)}$ denote the observed treatment effects from the remaining studies without the study that is omitted in ith iteration.[4] We then checked whether the 95% predictive interval (constructed using the variance) included the observed value of the treatment difference on the final outcome.

3 | METHODS FOR SURROGATE ENDPOINT EVALUATION INCORPORATING AGGREGATE DATA FROM DIFFERENT TREATMENT CLASSES

When subgroup analysis is used to investigate the study-level surrogate relationships within treatment classes the validation process may fail due to limited data resulting in estimates of the parameters describing surrogate relationships obtained with considerable uncertainty.[19] We propose two hierarchical models to investigate surrogate relationships within treatment classes allowing different degrees of borrowing of information about the parameters of interest, as alternative approaches to subgroup analysis with the standard model. These models were developed to investigate the study-level association and therefore they can only be applied to aggregate data (eg, logHR or logOR). They allow for the association patterns to vary across classes taking advantage of the attractive statistical properties of exchangeability. [14-16]

3.1 | Hierarchical model with full exchangeability

Our first approach extends the standard model accounting for differences in study-level surrogacy across different treatment classes.[24-26] Similarly as in the standard model, at the within-study level we assume that correlated and normally distributed observed treatment effects Y_{1ij} and Y_{2ij} (eg, logHR or logOR) in each study i estimate the true treatment effects μ_{1ij} and μ_{2ij} on the surrogate and final outcomes, respectively. In addition, by introducing index j we account for the differences between the classes.

$$\begin{gathered}\begin{pmatrix} Y_{1ij} \\ Y_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1ij} \\ \mu_{2ij} \end{pmatrix}, \begin{pmatrix} \sigma^2_{1ij} & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma^2_{2ij} \end{pmatrix} \right) \\ \mu_{2ij}|\mu_{1ij} \sim N(\lambda_{0j} + \lambda_{1j}\mu_{1ij}, \psi^2_j) \\ \lambda_{0j} \sim N(\beta_0, \xi^2_0), \lambda_{1j} \sim N(\beta_1, \xi^2_1). \end{gathered} \tag{3}$$

The parameters σ_{1ij}, σ_{2ij}, and ρ_{wij} correspond to the within-study SDs and within-study correlations for each study i in treatment class j. The observed estimates Y_{1ij}, Y_{2ij}, σ_{1ij}, σ_{2ij} are aggregate data extracted from systematic review RCTs while, the within-study correlations ρ_{wij} can be calculated using a bootstrapping method from individual patient data (IPD). Similarly as in the standard model, the true effects μ_{1ij} on the surrogate endpoint are modeled as fixed effects.

In contrast to the standard surrogacy model, this method assumes unique surrogate relationships between true treatment effects on the surrogate endpoint and the final outcome across treatment classes in a single model, allowing for borrowing of information across them. Each relationship between the true effects on the surrogate endpoint μ_{1ij} and the final outcome μ_{2ij} is described by a linear model where, λ_{0j} denotes the intercept of the jth treatment class and λ_{1j} establishes the relationship between treatment effects on surrogate and final outcomes within the treatment class j. To evaluate whether a candidate endpoint is considered a valid surrogate endpoint in a given treatment class, all three surrogacy criteria need to be met for this particular class. Implementing this model in the Bayesian framework, we place noninformative prior distributions on the model parameters such as: $\beta_0, \beta_1 \sim N(0, a)$ and $\xi_0, \xi_1 \sim N(0, b)I(0,)$, $\mu_{1ij} \sim N(0, a)$ and $\psi_j \sim N(0, b)I(0,)$. Similarly as in the standard model a, b are chosen to be sufficiently large and depend on the scale of data.

F-EX model extends the standard model (described in Section 2) by including an additional layer of hierarchy to the linear relationship between true effects on the surrogate and the final outcome, assuming that slopes and intercepts are fully exchangeable across treatment classes. This can be implemented by placing common normal distributions on λ_{0j} and λ_{1j} with means and variances β_0, ξ_0^2 and β_1, ξ_1^2, leading to borrowing of information across treatment classes. Hierarchical models have desirable statistical properties that allow us to improve our inferences taking advantage of borrowing of information from other treatment classes. The exchangeable estimates, however, are shrunk toward the means β_0, β_1 and the amount of shrinkage depends on the number of studies within each class, the between treatment class heterogeneity[17] and the number of treatment classes. Although these statistical properties are very attractive in terms of potential reduction of uncertainty around the parameters of interest, they are advantageous only when the assumption of exchangeability is reasonable, otherwise there is a danger of excessive shrinkage.

3.2 | Hierarchical model with partial exchangeability

F-EX method can be extended into a method with partial exchangeability (P-EX) similar to the method proposed by Neuenschwander et al.[17] This model is able to relax the assumption of exchangeability allowing the parameters of interest for each class to be either exchangeable with all or some of the parameters from other treatment classes or nonexchangeable. The proposed method is more flexible compared to F-EX model, in particular in data scenarios where the assumption of exchangeability is not reasonable for some of the treatment classes.

The within study and the between studies levels of this model are exactly the same as in the method with full exchangeability (F-EX) where, Y_{1ij}, Y_{2ij} are the treatment effects on the surrogate and final clinical outcomes and they follow a bivariate normal distribution with mean values corresponding to the true treatment effects μ_{1ij} and μ_{2ij} on the two outcomes.

$$\begin{aligned}\begin{pmatrix} Y_{1ij} \\ Y_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1ij} \\ \mu_{2ij} \end{pmatrix}, \begin{pmatrix} \sigma_{1ij}^2 & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma_{2ij}^2 \end{pmatrix}\right) \\ \mu_{2ij}|\mu_{1ij} \sim N(\lambda_{0j} + \lambda_{1j}\mu_{1ij}, \psi_j^2) \\ \lambda_{0j} \sim N(\beta_0, \xi_0^2) \\ \lambda_{1j} = \begin{cases} \lambda_{1j} \sim N(\beta_1, \xi_1^2) & \text{if } p_j = 1 \\ \lambda_{1j} \sim N(0, b) & \text{if } p_j = 0 \end{cases}\end{aligned} \tag{4}$$

However, the parameters of slopes are modeled in a different way compared to those of F-EX model. In this approach two possibilities arise for these parameters for each treatment class j. When $p_j = 1$ the parameter λ_{1j} can be exchangeable with some or all the parameters of the slopes from the other treatment classes via an exchangeable component. It follows a common normal distribution with other slopes as in F-EX model. On the other hand, when $p_j = 0$ the slope can be nonexchangeable with any slopes from the other treatment classes. In this case a vague prior distribution can be placed on the parameter, as in the standard model. The method evaluates the degree of borrowing of information of each parameter λ_{1j} by using these two components with respective mixture weights.

The main advantage of this method is that it allows the degree of exchangeability to be inferred from the data. In each MCMC iteration, the sampler chooses between the two components by using a Bernoulli distribution $p_j \sim Bernoulli(\pi_j)$. By calculating the posterior mean of this Bernoulli distribution we derive the mixture weights of each treatment class. The hyper-parameters π_j of the Bernoulli prior distribution can be either fixed or, in a fully Bayesian framework, they can follow a prior distribution, for example, a Beta distribution $\pi_j \sim Beta(1, 1)$. We have used fixed π_j, since placing a prior distribution required longer chains to converge and provided almost the same results.

In a special case where $p_j = 1$ for all treatment classes, P-EX model reduces to full exchangeability model as it uses only the exchangeable component. Having $p_j = 0$ for all treatment classes makes the P-EX model equivalent to subgroup analysis using the standard model as only the nonexchangeable component is used to estimate λ_{1j} in this case. In a Bayesian framework vague prior distributions can be placed on the parameters β_0, β_1, ξ_0, ξ_1, μ_{1ij} as in F-EX model.

4 | SOFTWARE IMPLEMENTATION AND COMPUTING

All models were implemented in OpenBUGS[27] where posterior estimates were obtained using MCMC simulations performing 50 000 iterations (after discarding 20 000 iterations as burn-in period). The OpenBUGS code of F-EX and P-EX models can be found in Appendix S1 (Sections D3, D4). Convergence was assessed visually by checking the history, chains and autocorrelation plots using graphical tools in OpenBUGS and R. All estimates are presented as means with corresponding 95% CrIs. The median was used only for the estimates of the conditional variances as a measure of central tendency since their posterior distributions were very skewed. The cross-validation procedure was performed in R using R2OpenBUGS[27] package to execute OpenBUGS code multiple times.

5 | SIMULATION STUDY

The proposed hierarchical methods allow different levels of borrowing of information for the parameters of interest. F-EX model assumes exchangeability of slopes while, the P-EX model allows for partial exchangeability of these parameters. We carried out a simulation study to assess the performance of the hierarchical methods and to compare them with subgroup analysis conducted using the standard model. We evaluated the performance of the methods in distinct data scenarios generated assuming different strengths of association within classes, different levels of similarity of the association patterns across classes and different number of studies per class. We evaluated the models' ability to identify treatment classes with strong association patterns and to make predictions of the treatment effect on the final outcome in a new study from a treatment effect measured on the surrogate endpoint.

5.1 | Data generation process and scenarios

We simulated data under nine different scenarios generating 1000 replications for each scenario. Each replication included average treatment effects on the surrogate and the final outcome (and corresponding SEs and within-study correlations) from a number of studies of treatments belonging to five treatment classes. We assumed that the data in each treatment class had a different heterogeneity pattern. Therefore, to have a control over such heterogeneity patterns when simulating the data we needed to make an assumption about the distribution of the true effects both on the surrogate and the final endpoints. The standard model by Daniels and Hughes assumes fixed effect for the true effects on the surrogate endpoint (no common distribution) making difficult to control the heterogeneity patterns when simulating the data. To avoid this issue, we simulated data using a product normal formulation of bivariate random effect meta-analysis (BRMA) (Equation (5)), assuming normal random effects on the surrogate endpoint. Apart from this assumption, this method is the same as Daniels and Hughes model using a bivariate normal distribution to describe the within-study variability and a linear relationship to model the association between the surrogate and the final outcome.

$$
\begin{aligned}
\begin{pmatrix} Y_{1ij} \\ Y_{2ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{1ij} \\ \mu_{2ij} \end{pmatrix}, \begin{pmatrix} \sigma_{1ij}^2 & \sigma_{1ij}\sigma_{2ij}\rho_{wij} \\ \sigma_{1ij}\sigma_{2ij}\rho_{wij} & \sigma_{2ij}^2 \end{pmatrix}\right) \\
\mu_{1ij} \sim N(\eta_{1j}, \psi_{1j}^2) \\
\mu_{2ij} | \mu_{1ij} \sim N(\eta_{2ij}, \psi_{2j}^2) \\
\eta_{2ij} = \lambda_{0j} + \lambda_{1j}\mu_{1ij} \\
\psi_{1j} = \frac{\psi_{2j}}{|\lambda_{1j}|\sqrt{(1/\rho_{bj}^2) - 1}}.
\end{aligned}
\tag{5}
$$

Simulating data from this model, however, can lead to results obtained with increased uncertainty, as the models used to analyze the data make fewer distributional assumptions.

To generate the data, we pursued the following steps:

1. Set the number of classes $N = 5$.
2. Simulate the data for each class separately using BRMA model (Equation 5) under three main designs.
3. Create three sets of scenarios: two with fixed number of studies ($n_j = 16$ and $n_j = 8, \mathrm{j} = 1, \ldots, 5$) per treatment class and one with unbalanced classes ($n_1 = 4, n_2 = 8, n_3 = 6, n_4 = 10, n_5 = 7$). We applied the three sets of scenarios to each design. In total, we have nine scenarios (3 designs × 3 sets = 9 scenarios).
4. Simulate the true effects using model (Equation 5)

The values of the parameters are listed in Table 1 and a short description of each design can be found below:

Design 1:

In the first design, our aim was to illustrate the properties of exchangeability. We simulated data in five treatment classes assuming high degree of similarity for their slopes and intercepts. The data in each treatment class were simulated assuming strong association (see surrogacy criteria in Section 2.1) for each individual class but weak overall.

Design 2:

The second design illustrates the case where there is a treatment class with very different association pattern (slope) compared to the other classes. This implies that the assumption of exchangeability is in doubt for this parameter in this particular class. Similarly as in the first scenario, we assumed strong association for each individual class.

Design 3:

The last design focuses on the association patterns of strengths that vary across treatment classes, investigating whether the proposed methods can estimate a strong association pattern better compared to subgroup analysis with the standard model and whether they can distinguish between the different association patterns despite borrowing of

TABLE 1 Simulation designs

First Design	Second Design	Third Design
$\lambda_{11} = 0.40, \rho_{b1} = 0.89$	$\lambda_{11} = 0.60, \rho_{b1} = 0.93$	$\lambda_{11} = 0.40, \rho_{b1} = 0.90$
$\lambda_{12} = 0.45, \rho_{b2} = 0.90$	$\lambda_{12} = 1.55, \rho_{b2} = 0.99$	$\lambda_{12} = 0.50, \rho_{b2} = 0.70$
$\lambda_{13} = 0.50, \rho_{b3} = 0.91$	$\lambda_{13} = 1.60, \rho_{b3} = 0.99$	$\lambda_{13} = 0.60, \rho_{b3} = 0.93$
$\lambda_{14} = 0.55, \rho_{b4} = 0.92$	$\lambda_{14} = 1.65, \rho_{b4} = 0.99$	$\lambda_{14} = 0.70, \rho_{b4} = 0.75$
$\lambda_{15} = 0.60, \rho_{b5} = 0.93$	$\lambda_{15} = 1.70, \rho_{b5} = 0.99$	$\lambda_{15} = 0.80, \rho_{b5} = 0.95$
$\lambda_{0j} = 0$	$\lambda_{0j} = 0$	$\lambda_{0j} = 0$
$\sigma_{1ij,2ij} = 0.1$	$\sigma_{1ij,2ij} = 0.1$	$\sigma_{1ij,2ij} = 0.1$
$\rho_{wij} = 0.4$	$\rho_{wij} = 0.4$	$\rho_{wij} = 0.4$
$\psi_{2j} = 0.08$	$\psi_{2j} = 0.08$	$\psi_{21,23,25} = 0.08$
		$\psi_{22,24} = 0.30$
$\eta_{1j} = 0.3$	$\eta_{1j} = 0.3$	$\eta_{1j} = 0.3$

information across treatment classes. To achieve this, we generated three out of five treatment classes with strong association and the remaining two classes with a weak association.

5.2 | Performance measures

To evaluate the goodness of fit of the models, we calculated the coverage probability of the 95% CrIs of λ_{1j} and the 95% predictive intervals of μ_{2ij}. The absolute bias and the root mean square error (RMSE) of $\hat{\lambda}_{1j}$ and $\hat{\mu}_{2ij}$ were also monitored and reported in the tables. In order to investigate potential decrease in the degree of uncertainty of the estimates as a result of borrowing of information across treatment classes, we calculated ratios of the width of the 95% CrIs. The width ratio $w_{\lambda_{1j}^{FEX,(PEX)}}/w_{\lambda_{1j}^{subgr}}$ was defined as the ratio of the widths of the CrIs of λ_{1j} from F-EX or P-EX to the width of the CrIs of λ_{1j} from subgroup analysis using the standard model. Similarly, the width ratio $w_{\mu_{2ij}^{\mathrm{FEX,(PEX)}}}/w_{\mu_{2ij}^{\mathrm{subgr}}}$ was the ratio of the 95% predictive intervals of the true effects μ_{2ij} from F-EX or P-EX to the width of the predictive intervals of μ_{2ij} from subgroup analysis using the standard model. We also monitored the largest Monte Carlo error (MCE) of the simulations as an index of accuracy of the Monte Carlo samples.

Furthermore, a cross-validation procedure was applied to each method across the simulated data scenarios. In the simulation study, the true effect on the final endpoint μ_{2ij} was known, since it had been simulated ,therefore the cross-validation procedure was applied on the true effects (in real data scenarios we compare the predicted effect with the observed effect) by checking whether the simulated value of the true effect $\hat{\mu}_{2ij}$ was included in the predictive interval of μ_{2ij}.

5.3 | Results

All the tables in the results section list the performance of the posterior means of $\hat{\lambda}_{1j}$, the performance of the posterior means of $\hat{\mu}_{2ij}$ as well as the probabilities of estimating a strong association pattern (see definition in Section 2.1) for each class across methods. The following section presents the results of the analysis by reporting the coverage probabilities of the CrIs of λ_{1j} and μ_{2ij} for each scenario (by taking the mean of coverage probabilities across classes), the overall absolute bias and RMSE of $\hat{\lambda}_{1j}$ and $\hat{\mu}_{2ij}$, the width ratios of λ_{1j} and μ_{2ij} for each scenario (by calculating the mean of the width ratios of λ_{1j} across classes and the mean of the width ratios of μ_{2ij} across studies and classes), the MCE and the probability to estimate a strong association pattern by fitting each model. Detailed results for the performance of $\hat{\lambda}_{1j}$ and $\hat{\mu}_{2ij}$ for each class separately and across methods are listed in Appendix S1 (see Sections B and C).

5.3.1 | Performance of the estimates $\hat{\lambda}_{1j}$

Table 2 presents the results across the nine scenarios reporting averages of the measures we monitored for $\hat{\lambda}_{1j}$ over the five classes of treatment. The performance of the models varied in terms of the coverage probability of the 95% CrIs of λ_{1j} across scenarios. In the scenarios 1, 4, and 7 where the number of studies per class was relatively high, the models achieved 95% coverage probabilities. However, in the scenarios where the number of studies was smaller the coverage probability was higher due to increased uncertainty and likely to the fact that the model we used in the generation process was slightly different from models used to fit the data. MCEs were small across most of the scenarios implying good accuracy of the Monte Carlo samples and that convergence was achieved in those scenarios across all the methods. However, in scenarios where the data were limited (scenarios 3, 6, and 9) subgroup analysis with the standard model yielded larger MCEs. This implies that subgroup analysis requires longer chains to achieve the same level of convergence as the other two models.

In the first three scenarios (first design), where the treatment classes were very similar in terms of patterns (similar slopes), F-EX and P-EX were superior compared to subgroup analysis as they gave posterior means of slopes with lower absolute bias, RMSE and reduced uncertainty (narrower 95% CrIs) due to borrowing of information across classes. P-EX model achieved almost the same level of borrowing of information as F-EX model, with mixtures weights were very close to 1 across treatment classes (see details in the Section D1 in Appendix S1 where the mixture weights are listed). Overall, the proposed hierarchical models performed better compared to subgroup analysis but the difference

TABLE 2 Performance of $\hat{\lambda}_{1j}$

Scenario	Number of Studies Across Classes	Methods	Coverage (Mean)	Absolute Bias (Mean)	RMSE	Width Ratio (Mean)	Monte Carlo Error	Probability of strong association (Mean)
	First design							
First	Fixed ($n_j = 16$)	Subgroup analysis	0.95	0.08	0.10		0.003	0.81
		F-EX model	0.95	0.06	0.07	0.72	0.002	0.85
		P-EX model	0.96	0.06	0.07	0.72	0.002	0.85
Second	Fixed ($n_j = 8$)	Subgroup analysis	0.98	0.11	0.15		0.005	0.71
		F-EX model	0.97	0.07	0.09	0.60	0.003	0.89
		P-EX model	0.97	0.07	0.09	0.61	0.003	0.90
Third	Unbalanced	Subgroup analysis	0.99	0.13	0.18		0.017	0.56
	($n_1 = 4, n_2 = 8, n_3 = 6,$	F-EX model	0.99	0.07	0.09	0.52	0.003	0.89
	$n_4 = 10, n_5 = 7$)	P-EX model	0.99	0.07	0.09	0.53	0.004	0.88
	Second design							
Fourth	Fixed ($n_j = 16$)	Subgroup analysis	0.95	0.09	0.11		0.007	0.89
		F-EX model	0.94	0.08	0.10	0.90	0.005	0.91
		P-EX model	0.94	0.07	0.09	0.86	0.004	0.91
Fifth	Fixed ($n_j = 8$)	Subgroup analysis	0.97	0.14	0.17		0.007	0.88
		F-EX model	0.96	0.12	0.15	0.86	0.005	0.92
		P-EX model	0.97	0.10	0.12	0.78	0.005	0.92
Sixth	Unbalanced	Subgroup analysis	0.98	0.15	0.20		0.025	0.72
	($n_1 = 4, n_2 = 8, n_3 = 6,$	F-EX model	0.96	0.17	0.21	0.70	0.011	0.88
	$n_4 = 10, n_5 = 7$)	P-EX model	0.97	0.14	0.18	0.70	0.011	0.87
	Third Design							
Seventh	Fixed ($n_j = 16$)	Subgroup analysis	0.95	0.11	0.14		0.003	
		F-EX model	0.95	0.09	0.11	0.79	0.002	
		P-EX model	0.95	0.09	0.11	0.79	0.003	
Eighth	Fixed ($n_j = 8$)	Subgroup analysis	0.97	0.17	0.22		0.006	
		F-EX model	0.96	0.11	0.14	0.67	0.004	
		P-EX model	0.96	0.11	0.14	0.67	0.004	
Ninth	Unbalanced	Subgroup Analysis	0.98	0.19	0.25		0.021	
	($n_1 = 4, n_2 = 8, n_3 = 6,$	F-EX model	0.97	0.12	0.15	0.56	0.005	
	$n_4 = 10, n_5 = 7$)	P-EX model	0.97	0.12	0.15	0.57	0.005	

TABLE 3 Probabilities of estimating a strong association pattern per class in the third design

Scenario	Number of Studies Across Classes	Treatment Classes	Subgroup Analysis	F-EX Model	P-EX Model
Seventh	Fixed ($n_j = 16$)	First class	0.82	0.84	0.84
		Second class[a]	0.00	0.00	0.00
		Third class	0.83	0.85	0.85
		Fourth class[a]	0.00	0.00	0.00
		Fifth class	0.80	0.80	0.80
Eighth	Fixed ($n_j = 8$)	First class	0.78	0.89	0.89
		Second class[a]	0.04	0.05	0.05
		Third class	0.80	0.90	0.90
		Fourth class[a]	0.06	0.06	0.06
		Fifth class	0.85	0.87	0.86
Ninth	Unbalanced	First class	0.06	0.82	0.80
	($n_1 = 4, n_2 = 8,$	Second class[a]	0.06	0.07	0.07
	$n_3 = 6, n_4 = 10,$	Third class	0.65	0.91	0.91
	$n_5 = 7$)	Fourth class[a]	0.03	0.03	0.03
		Fifth class	0.82	0.89	0.89

Abbreviations: F-EX, full exchangeability; P-EX, partial exchangeability.

[a] Treatment classes with weak association pattern.

was more pronounced in the scenarios with small number of studies. In the second design (scenarios 4, 5, and 6), where the exchangeability assumption was not reasonable for one of the classes, P-EX model yielded the most robust results. The model resulted in the posterior means with the smallest absolue bias and RMSE, reducing the degree of borrowing of information for the class with the distinctly different (the mixture weights in this class were $p_1 = 0.56$, $p_1 = 0.31$ and $p_1 = 0.80$ respectively) while it still borrowed almost the same level of information across the remaining classes as F-EX model ($p_2, p_3, p_4, p_5 \approx 0.97$). On the other hand, F-EX performed poorer compared to the other methods in scenario 6 with unbalanced and relatively small number of studies per class, leading to more biased results. This indicates that F-EX model is not appropriate when the assumption of exchangeability is not reasonable. Subgroup analysis using the standard model achieved decent performance only in the forth scenario where there were sufficient data. In the third design (scenarios 7, 8, and 9) the proposed models achieved superior performance compared to subgroup analysis for the estimates of λ_{1j}, similarly as in the first three scenarios.

The last column of Table 2 shows the probabilities of estimating a strong association pattern across the data scenarios and models. F-EX and P-EX methods estimated the surrogacy (based on the three surrogacy criteria) better compared to subgroup analysis across all scenarios. In the first design (scenarios 1, 2, and 3) where the association was designed to be strong for all the classes, F-EX and P-EX models predicted a strong association pattern in more than 85% of the simulations. Subgroup analysis predicted the 81% of them in the first scenario but its performance reduced noticeably in the second and third scenario where the data were more sparse. In the second design (scenarios 4, 5, and 6) with strong association patterns across all classes, P-EX and F-EX estimated more than 87% of the association patterns across these three scenarios. Subgroup analysis performed well only in the fourth scenario predicting the 89% of the association patterns but its performance gradually reduced as the number of studies was decreased in scenario 5 and 6.

Table 3 presents the results from the last three scenarios (third design), where the surrogate relationships varied across classes. F-EX and P-EX methods were able to estimate a strong association pattern with higher probability compared to subgroup analysis in the classes where the association was designed to be strong. At the same time, the methods successfully identified classes with strong association patterns from a mixture of classes with weak and strong association patterns, even for the scenarios with relatively few studies per class where subgroup analysis failed almost completely to identify. The probabilities of estimating a strong association per class in designs 1 and 2 are presented in Appendix S1 (see sections B1, B2, B3, B4, B5, B6).

TABLE 4 Performance of $\hat{\mu}_{2ij}$

Scenario	Number of Studies Across Classes	Methods	Coverage (Mean)	Absolute Bias (Mean)	RMSE	Width Ratio (Mean)	MCE
First design							
First	Fixed ($n_j = 16$)	Subgroup analysis	0.95	0.09	0.11		0.003
		F-EX model	0.95	0.08	0.10	0.93	0.002
		P-EX model	0.95	0.08	0.10	0.93	0.002
Second	Fixed ($n_j = 8$)	Subgroup Analysis	0.98	0.11	0.13		0.010
		F-EX model	0.98	0.08	0.10	0.80	0.004
		P-EX model	0.98	0.08	0.10	0.80	0.004
Third	Unbalanced	Subgroup analysis	0.99	0.12	0.18		0.023
	($n_1 = 4, n_2 = 8, n_3 = 6,$	F-EX model	0.99	0.08	0.11	0.67	0.005
	$n_4 = 10, n_5 = 7$)	P-EX model	0.99	0.09	0.11	0.68	0.008
Second design							
Fourth	Fixed ($n_j = 16$)	Subgroup analysis	0.95	0.13	0.18		0.009
		F-EX model	0.95	0.13	0.18	0.97	0.008
		P-EX model	0.96	0.12	0.17	0.96	0.008
Fifth	Fixed ($n_j = 8$)	Subgroup analysis	0.99	0.16	0.20		0.015
		F-EX model	0.98	0.15	0.19	0.92	0.009
		P-EX model	0.98	0.14	0.18	0.87	0.008
Sixth	Unbalanced	Subgroup analysis	0.99	0.18	0.23		0.021
	($n_1 = 4, n_2 = 8, n_3 = 6,$	F-EX model	0.99	0.18	0.22	0.80	0.009
	$n_4 = 10, n_5 = 7$)	P-EX model	0.99	0.15	0.19	0.77	0.010
Third Design							
Seventh	Fixed ($n_j = 16$)	Subgroup analysis	0.95	0.16	0.23		0.006
		F-EX model	0.95	0.16	0.22	0.96	0.004
		P-EX model	0.95	0.16	0.22	0.96	0.004
Eighth	Fixed ($n_j = 8$)	Subgroup analysis	0.98	0.18	0.26		0.017
		F-EX model	0.97	0.16	0.22	0.85	0.006
		P-EX model	0.97	0.16	0.22	0.85	0.006
Ninth	Unbalanced	Subgroup analysis	0.98	0.20	0.28		0.027
	($n_1 = 4, n_2 = 8, n_3 = 6,$	F-EX model	0.97	0.17	0.21	0.72	0.008
	$n_4 = 10, n_5 = 7$)	P-EX model	0.97	0.17	0.21	0.72	0.009

Abbreviations: F-EX, full exchangeability; MCE, Monte Carlo errors; P-EX, partial exchangeability; RMSE, root mean square error.

5.3.2 | Performance of predictions $\hat{\mu}_{2ij}$

Table 4 shows the results from cross-validation procedure which resulted in the posterior means ($\hat{\mu}_{2ij}$) and 95% predictive intervals of the true effects μ_{2ij}. It presents the same measures as Table 2 averaged over the five classes. In scenarios 1, 4, and 7, the models achieved 95% coverage due to the large amount of data, however, in the remaining scenarios where the number of studies was smaller the models yielded higher coverages probabilities. F-EX and P-EX had small MCEs across all scenarios, however, subgroup analysis gave on average significantly larger MCEs compared to the proposed methods in scenarios 3, 6, and 9 (see details in Sections C3, C6, C9 in Appendix S1). This indicates that subgroup analysis with the standard model requires longer chains for its posteriors to achieve the same level of convergence as the other two methods.

In the first three scenarios, F-EX and P-EX models outperformed subgroup analysis in terms of the absolute bias, RMSE and the uncertainty of $\hat{\mu}_{2ij}$. However, there was no winner between them as both methods had almost the same degree of borrowing of information resulting in 7%, 20%, and 33% narrower predictive intervals compared to subgroup analysis across these three scenarios, respectively. In the scenarios 4 ,5, and 6 (second design), P-EX yielded posterior means with the smallest absolute bias, RMSE and CrIs with the smallest width ratio across classes. Furthermore, P-EX method gave the most robust results for the ”extreme” treatment class reducing by 44%, 69%, and 20% the borrowing of information in this class across the scenarios (see the mixture weights in Section D1 in Appendix S1). In the sixth scenario F-EX performed poorer compared to P-EX model leading to biased results especially for the treatment class where the surrogacy was different and the exchangeability assumption unreasonable (first class in the Section C6 of Appendix S1). Subgroup analysis performed almost equally well as the P-EX model in the 4th scenario where the number of studies per class was relatively large. The last three scenarios (third design) gave similar results as the first three in terms of the uncertainty, the absolute bias and the RMSE of $\hat{\mu}_{2ij}$. F-EX P-EX models performed equally well, while subgroup analysis with the standard model was the worst approach resulting in inflated predictive intervals, larger RMSE and worse MCE in all cases.

5.4 | Discussion of the results

The aim of the simulation study was to illustrate and assess the performance of the methods under different scenarios. The models gave 95% coverage probabilities in the scenarios 1, 4, and 7 where the number of studies was sufficiently large (16 for each class). However, in the remaining scenarios the coverage probabilities were higher than 95%, which means that the methods derived more conservative CrIs of parameters than expected. This is largely due to the sparsity of the data in these scenarios but may also be partly due to different models being used to simulate and analyze the data as explained in Section 5.1. In the first design (scenarios 1, 2, and 3) where the assumption of exchangeability was reasonable, F-EX and P-EX models performed better than the subgroup analysis giving on average narrower 95% CrIs of λ_{1j} and 95% predictive intervals of μ_{2ij}. This indicates that P-EX model successfully identified the correct level of borrowing of information inferring that the mixture weights should be very close to 1. P-EX model was the best choice in all the scenarios of the second design (scenarios 4, 5, and 6) where there was a treatment class with distinctly different slope. It reduced the degree of borrowing of information for the ”extreme” treatment class, giving the most accurate posterior means of the slopes. Moreover, P-EX model was the best choice in terms of predictions of the true effect on the final endpoint, reducing the width of predictive intervals by 4%, 13%, and 23% compared to subgroup analysis in each scenario, respectively. Last but not least, the proposed methods estimated the strong association patterns better compared to subgroup analysis across all data scenarios. In particular, in scenarios 3, 6, and 9, where the data were sparse, the proposed hierarchical methods were able to estimate surrogacy significantly better compared to the subgroup analysis. This illustrates well the benefits of using hierarchical methods when data are limited. Furthermore, as illustrated by scenarios 7, 8, and 9, F-EX and P-EX could easily distinguish between the different association patterns as they identified treatment classes with strong association patterns and at the same time did not overestimate the strength of the association in the classes where the association was designed to be weak.

6 | APPLICATION: ADVANCED COLORECTAL CANCER

6.1 | Data

We illustrate the proposed methodology in an example in aCRC. The data were obtained from a systematic review conducted by Ciani et al[10] which included 101 RCTs published between 2003 and 2013, evaluating multiple interventions in aCRC. The review consist of trials that report treatment effects on OS or/and on alternative endpoints such as PFS, TR. OS was defined as the time from randomization to time of death, PFS was set as the time from randomization to tumor progression or death from any cause. TR was estimated using objective tumor measurements which are measured using imaging methods and determined according to the Response Evaluation Criteria in Solid Tumors guidelines[28] or the World Health Organization recommendations.[29] The RCTs in the systematic review contain five treatment classes: the class of chemotherapies, the anti-epidermal growth factor receptor (Anti-EGFR) monoclonal antibodies class, angiogenesis inhibitors, other molecular-targeted agents and intrahepatic arterial chemotherapies .

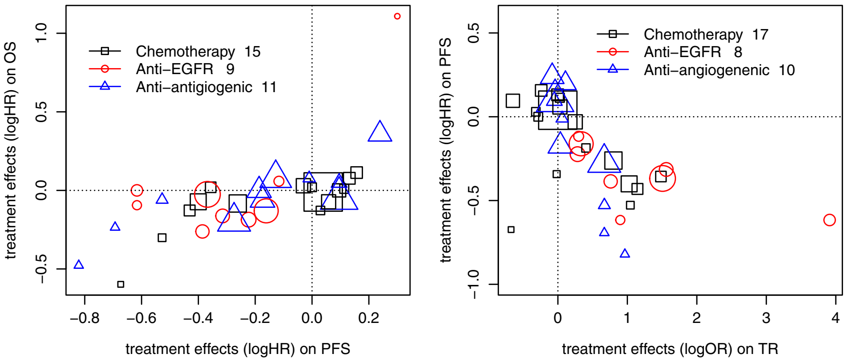


FIGURE 1 Scatterplots of treatment effects on progression-free survival-overall survival and tumor response-progression-free survival [Color figure can be viewed at wileyonlinelibrary.com]

Ciani et al[10] investigated surrogate relationships between treatment effects on potential surrogate endpoints (TR and PFS) and on the final clinical outcome (OS). They found that the surrogate relationships between treatment effects on these endpoints were suboptimal. Furthermore, they stated that PFS was an acceptable surrogate endpoint for OS, whereas TR should not be used as a surrogate endpoint for this final outcome. They concluded that good surrogacy observed in previous studies, that included traditional chemotherapy trials in aCRC may not apply directly across other classes of treatments. More details about the studies and how the systematic review was designed can be found in Ciani et al.[10] We refer these data as ”Ciani data” in the remainder of this paper.

In our example, we focused on a subset of these data examining the surrogacy between treatment effects on TR and PFS and treatment effects on PFS and OS including data from three treatment classes. We obtained data from 35 studies reporting treatment effect on PFS and OS where, 15 of them belonged to the chemotherapy treatment class, nine of them investigated anti-EGFR therapies and 11 anti-angiogenic treatments. To investigate surrogate relationships between treatment effects on TR and PFS we used data from 35 studies reporting treatment effects on these endpoints; 17 of them investigated chemotherapies, 8 and 10 studies anti-EGFR and anti-angiogenic treatments respectively. TR can be evaluated as a surrogate endpoint to treatment effect on PFS, as treatment effects on TR is typically measured earlier compared to treatment effects on PFS.

Figure 1 provides a graphical representation of the dataset we used. It illustrates the association patterns between the treatment effects across classes on each pair of outcomes.

IPD were available from four RCTs,[30-33] which were used to estimate the within-study correlations. By applying a bootstrap method (see Section A in Appendix S1) we estimated two sets of within-study correlations: for each of the two pairs of outcomes one correlation corresponding to each treatment class. We assumed that within treatment classes the within-study correlations are the same across studies.

6.2 | Scale of the outcomes

The treatment effects on OS and PFS were modeled on the log hazard ratio scale $logHR(OS)$, $logHR(PFS)$, whereas the treatment effects on TR were modeled on log odds ratio $logOR(TR)$ scale. We retrieved the corresponding SEs of $logHR(PFS)$ and $logHR(OS)$ on PFS and OS from the 95% confidence intervals and by using the standard formulae for the SEs of $logOR(TR)$.

6.3 | Results of data analysis

The first aim of our analysis was to explore potential differences in association patterns across treatment classes. To investigate this, we applied the two proposed models and subgroup analysis using standard model to the data and derived posterior distributions for the parameters of the surrogate relationships for each treatment class. We obtained the posteriors mean of the intercepts $\hat{\lambda}_{0j}$, the slopes $\hat{\lambda}_{1j}$ and posterior median of conditional variances $\hat{\psi}_j^2$ with corresponding 95% CrIs across treatment classes. By checking the surrogacy criteria (described in Section 2.1) we were able to infer whether or not a candidate endpoint is a valid surrogate in each treatment class. We carried out a cross-validation procedure (Section 2.2) to investigate how well the models predict the true treatment effect on the final clinical outcome. The measures we monitored were the absolute error of the predictions, the ratios of the width of the 95% predictive intervals from P-EX or F-EX to the width of the 95% predicted interval obtained from subgroup analysis and the largest MCE.

6.3.1 | Results across models and treatment classes

Subgroup analysis with the standard model

The results of subgroup analysis presented in the first two columns of Table 5 showed strong association between the treatment effects on PFS and the effects on OS in the class of chemotherapies and the anti-angiogenic treatment class with all three criteria for surrogacy satisfied (the 95% CrIs of λ_{01} and λ_{03} included zero, the 95% CrIs of λ_{11} and λ_{13} did not contain zero and there was substantial evidence using Bayes factors in favor of the hypotheses $H_1 : \psi_1^2 = 0$, and $H_1 : \psi_3^2 = 0$ (see details about Bayes factors Section D2 in Appendix S1)). In contrast, we can infer that the surrogate relationship between treatment effects on PFS and the effects on OS in the anti-EGFR treatment class was weak, as the 95% CrI of the posterior distribution of the slope included zero. Investigating the surrogacy on TR-PFS pair we found a similar pattern, thus we can infer that there was an acceptable surrogate relationship between treatment effects on TR and PFS in the chemotherapy and the anti-angiogenic classes. The relationship was negative overall, since the slopes were negative across classes. On the other hand, the surrogacy criteria indicated poor surrogacy between the treatment effects on TR and the treatment effects on PFS for anti-EGFR class, since the 95% CrI of the slope λ_{12} included zero.

F-EX model

The results of F-EX model are presented in columns 3 and 4 of Table 5. For the PFS-OS pair of outcomes, the association patterns were very similar in the anti-angiogenic and chemotherapy treatment classes as both classes satisfied the surrogacy criteria and the slopes were of similar magnitude. The 95% CrIs of the intercepts λ_{01} and λ_{03} included zero indicating that zero treatment effect on the surrogate implies zero treatment effect on the final outcome for these two classes. The intervals of the slopes λ_{11} and λ_{13} did not contain zero indicating positive association as the two slopes were positive. The conditional variances in these two classes were small indicating strong association which was supported by the analysis using Bayes factors (see details about the Bayes factors in Section D2 in Appendix S1). On the other hand, the association was weak in the anti-EGFR treatment class failing to meet one of the criteria, as the 95% CrI of the slope λ_{12} included zero. On the contrary, for TR-PFS pair of outcomes all three surrogacy criteria were satisfied across all the treatment classes taking advantage of the assumption of exchangeability of the parameters λ_{0j} and λ_{1j}. This implies that TR was an acceptable surrogate endpoint for PFS across treatment classes in this data set.

P-EX model

P-EX model allows the parameters of slope of each treatment class to be either exchangeable or nonexchangeable with parameters of slopes from other classes yielding parameters with partial exchangeability. For both pairs of outcomes, fixed values for the hyper-parameters $\pi_j = (0.5, 0.5, 0.5)$ were chosen assuming that exchangeability and nonexchangeability were a priori equally likely.

As in the case of F-EX model, the surrogacy criteria were estimated for each class separately and then a cross-validation procedure followed, however, for this model we also monitored the mixture weights by calculating the posterior means of p_j in order to measure the degree of borrowing of information across classes (Table 5 columns 5, 6). For the PFS-OS pair, the weights increased from their prior values ($\pi_j = 0.5$) to 0.968 in the class of chemotherapy, to 0.965 in the anti-EGFR class and to 0.966 in the anti-angiogenic treatment class indicating that borrowing of information was reduced approximately 3.5% for each class compared to F-EX model. Looking at the results from P-EX model we drew the same inferences as from F-EX model, inferring that the association patterns were strong in the anti-angiogenic and the chemotherapy classes, but weak in the anti-EGFR treatment class where the 95% CrI of the slope λ_{12} included zero. In contrast to this, for TR-PFS pair the mixture weights were smaller than on PFS-OS pair due to the slightly larger between treatment class heterogeneity. There was 7.1% reduction in borrowing of information in anti-angiogenic class compared to F-EX models, while the weights for the chemotherapies and anti-EGFR agents were 0.944 and 0.95, respectively. All three surrogacy criteria were fulfilled across treatment classes despite the decrease in levels of borrowing of information, indicating that TR was an acceptable surrogate for PFS across treatment classes in the Ciani data.

TABLE 5 Estimates of the parameters defining the surrogacy criteria

	Standard Model		F-EX		P-EX	
Treatment Classes	**PFS-OS**	**TR-PFS**	**PFS-OS**	**TR-PFS**	**PFS-OS**	**TR-PFS**
Chemotherapy	N=15	N=17	N=15	N=17	N=15	N=17
p_1	–	–	–	–	0.968	0.944
λ_{01}	−0.002 (−0.059, 0.053)	−0.050 (−0.164, 0.033)	0.003 (−0.050, 0.054)	−0.051 (−0.154, 0.033)	0.003 (−0.050, 0.054)	−0.051 (−0.155, 0.033)
λ_{11}	0.322 (0.089, 0.548)	−0.261 (−0.402, −0.097)	0.334 (0.124, 0.533)	−0.267 (−0.406, −0.111)	0.334 (0.124, 0.535)	−0.266 (−0.404, −0.109)
ψ_1^2	0.001 ($5{\cdot}10^{-6}$, 0.009)	0.016 ($4{\cdot}10^{-4}$, 0.072)	0.001 ($2{\cdot}10^{-6}$, 0.009)	0.016 ($4{\cdot}10^{-4}$, 0.069)	0.001 ($2{\cdot}10^{-6}$, 0.009)	0.016 ($3{\cdot}10^{-4}$, 0.069)
Anti-EGFR	N=9	N=8	N=9	N=8	N=9	N=8
p_2	–	–	–	–	0.965	0.950
λ_{02}	−0.048 (−0.292, 0.296)	−0.195 (−0.415, 0.033)	0.001 (−0.153, 0.146)	−0.138 (−0.338, 0.059)	−0.001 (−0.160, 0.149)	−0.140 (−0.341, 0.058)
λ_{12}	0.126 (−0.544, 1.031)	−0.140 (−0.366, 0.019)	0.274 (−0.157, 0.640)	−0.187 (−0.421, −0.027)	0.268 (−0.182, 0.648)	−0.184 (−0.418, −0.026)
ψ_2^2	0.008 ($2{\cdot}10^{-5}$, 0.103)	0.013 ($7{\cdot}10^{-5}$, 0.131)	0.010 ($8{\cdot}10^{-5}$, 0.078)	0.014 ($8{\cdot}10^{-5}$, 0.128)	0.010 ($8{\cdot}10^{-5}$, 0.079)	0.014 ($7{\cdot}10^{-5}$, 0.127)
Anti-angiogenic	N=11	N=10	N=11	N=10	N=11	N=10
p_3	–	–	–	–	0.966	0.929
λ_{03}	0.052 (−0.038, 0.149)	0.074 (−0.079, 0.225)	0.031 (−0.041, 0.113)	0.030 (−0.131, 0.178)	0.032 (−0.041, 0.115)	0.031 (−0.132, 0.180)
λ_{13}	0.481 (0.174, 0.797)	−0.786 (−1.197, −0.455)	0.411 (0.158, 0.685)	−0.674 (−1.060, −0.271)	0.413 (0.160, 0.694)	−0.686 (−1.075, −0.280)
ψ_3^2	0.006 ($1{\cdot}10^{-4}$, 0.040)	0.011 ($4{\cdot}10^{-5}$, 0.092)	0.006 ($6{\cdot}10^{-5}$, 0.036)	0.015 ($1{\cdot}10^{-4}$, 0.115)	0.006 ($8{\cdot}10^{-5}$, 0.036)	0.015 ($6{\cdot}10^{-5}$, 0.114)

Abbreviations: F-EX, full exchangeability; OS, overall survival; P-EX, partial exchangeability; PFS, progression-free survival; TR, tumor response.

TABLE 6 Predictions of μ_{2ij} across treatments and models

		Chemotherapy		Anti-EGFR		Anti-angiogenic		Overall	
Models	**Measures**	**PFS-OS**	**TR-PFS**	**PFS-OS**	**TR-PFS**	**PFS-OS**	**TR-PFS**	**PFS-OS**	**TR-PFS**
Standard model	Performance of 95% predictive intervals	1.000	0.941	0.888	1.000	1.000	1.000	0.971	0.971
	Absolute error (median)	0.047	0.108	0.140	0.132	0.099	0.145	0.090	0.123
	MCE (max)	0.002	0.004	0.006	0.004	0.003	0.004	0.003	0.004
F-EX	Performance of 95% predictive intervals	1.000	0.941	1.000	1.000	1.000	1.000	1.000	0.971
	Absolute error (median)	0.041	0.104	0.102	0.112	0.123	0.206	0.089	0.128
	Width ratio (median)	0.988	0.985	0.862	0.968	0.930	0.997	0.950	0.987
	MCE (max)	0.002	0.004	0.005	0.005	0.003	0.003	0.003	0.004
P-EX	Performance of 95% predictive intervals	1.000	0.941	1.000	1.000	1.000	1.000	1.000	0.971
	Absolute error (median)	0.041	0.104	0.126	0.114	0.109	0.206	0.092	0.128
	Width ratio (median)	0.989	0.989	0.864	0.975	0.931	0.999	0.957	0.990
	MCE (max)	0.002	0.004	0.005	0.005	0.003	0.003	0.003	0.004

Abbreviations: F-EX, full exchangeability; MCE, Monte Carlo error; OS, overall survival; P-EX, partial exchangeability; PFS, progression-free survival; TR, tumor response.

6.3.2 | Results of the cross-validation procedure

After estimating the surrogacy criteria across treatment classes, we carried out cross-validation procedure to predict the treatment effects μ_{2i} on the final outcome. The results in Table 6 showed that the cross-validation procedure of subgroup analysis with the standard model gave predictive intervals of the effects on the final outcome containing the corresponding observed estimates Y_{2i} in the 97% of the studies for both pairs of outcomes confirming good fit of the model. The cross-validation procedure yielded the most accurate posterior means of the true effects on the final endpoint (small absolute error) in the treatment class of chemotherapies, where the number of the available studies was large and performed poorly in terms of accuracy of predictions in the anti-EGFR class (large absolute error) where the surrogacy was weak and the number of studies small. Similarly, subgroup analysis with the standard model was less accurate in targeted treatment classes for the TR-PFS pair of outcomes where the number of studies was smaller.

The results from the cross-validation procedure of F-EX model showed that the method fitted the data well. All of the predicted intervals of μ_{2ij} contained the observed values of the treatment effects on the final outcome on PFS-OS pair and all but one on TR-PFS pair. The cross-validation procedure yielded the posterior means of μ_{2ij} with the smallest absolute error in chemotherapy treatment class on PFS-OS pair and performed equally well in terms of its accuracy in the other two classes. In contrast to this, higher absolute error were observed in the anti-angiogenic class on TR-PFS pair indicating that the assumption of exchangeability of the parameters describing the surrogate relationships was fairly strong and it was likely to cause ”overshrinkage” in this particular class. The results obtained for the width ratios imply that F-EX method gave intervals of the true effect on the final endpoint with smaller degree of uncertainty compared to subgroup analysis. There was a small decrease in the uncertainty of the predictions of μ_{2ij} on PFS-OS pair for the chemotherapy treatment class, as the cross-validation procedure of F-EX model yielded 1.2% narrower intervals compared to subgroup analysis. Furthermore, significantly reduced uncertainty was observed in the other two treatment classes for PFS-OS pair, 13.8% in the anti-EGFR treatment class and 7% in the anti-angiogenic, where the number of studies was smaller. On the contrary, very limited decrease in the degree of uncertainty was observed for the TR-PFS pair of outcomes across all classes. Overall on this pair, the predictive intervals were only 1.3% narrower compared to subgroup analysis. The benefit was small (3.2% reduction of the width of the predictive interval) even for the anti-EGFR treatment class where there were only 8 studies for this pair.

Focusing on the results from the cross-validation procedure using P-EX model, all the intervals of the predicted treatment effects on the final outcome contained the observed treatment effects on PFS-OS pair and all but one on the TR-PFS pair. The absolute error was smaller in chemotherapy treatment class for the PFS-OS pair where the number of studies was large and significantly higher in the other two classes. In contrast to this, the cross-validation procedure with P-EX gave almost equally accurate estimates in the anti-EGFR and the chemotherapy treatment classes on TR-PFS pair. However, the absolute error was higher in the anti-angiogenic treatment class where the association was much stronger compared to the other two classes indicating potential excessive borrowing of information from the other classes. This is likely due to the assumption of full exchangeability of the intercepts.

The method predicted the effects on the final outcome with reduced uncertainty giving more precise estimates ($\hat{\mu}_{2ij}$) compared to subgroup analysis in the anti-EGFR class on PFS-OS pair reducing the uncertainty by 13.6%. On the other hand, the predicted effects $\hat{\mu}_{2ij}$ had almost the same degree of uncertainty as those from subgroup analysis for TR-PFS pair. The intervals were only 1% narrower on average across all classes compared to the subgroup analysis.

6.4 | Comparison of the results from F-EX, P-EX, and those from subgroup analysis

Figure 2 presents 95% CrIs of the slopes λ_{1j} and intercepts λ_{0j} across the treatment classes and methods of estimation. Comparing the aforementioned methods in regards to the surrogacy criteria on the PFS-OS pair, we can conclude that F-EX model estimated the parameters of the surrogate relationships with reduced uncertainty compared to the subgroup analysis and P-EX model taking advantage of borrowing of information across classes. P-EX relaxes the assumption of exchangeability reducing the effect of borrowing of information on average by 3.6%. It gave narrower CrIs of the parameters of interest compared to subgroup analysis but slightly larger than those obtained form F-EX model. Furthermore, both F-EX and P-EX methods can distinguish between the different association patterns avoiding to give over-shrunk estimates of the slopes and the intercepts, although they allow different degrees of borrowing of information for the slopes. In particular, this pair of outcomes (PFS-OS) illustrates well the impact of number of studies per class on the degree of borrowing of information. In general, borrowing of information is determined by the number of studies within treatment classes, between treatment classes heterogeneity, as well as the number of treatment classes. In this case, the fewer studies we have within a treatment class, the bigger is the impact of borrowing of information resulting in higher reduction in uncertainty of the estimates of surrogate relationships. This effect was particularly strong for the anti-EGFR treatment class.

On the other hand, TR-PFS pair is a good example to illustrate the performance of the hierarchical methods when between treatment class heterogeneity is relatively large. In this case, subgroup analysis performed equally well as the proposed methods in terms of uncertainty of the CrIs of the paramaters describing the surrogate relationships. For instance by fitting F-EX and P-EX models, we did not observe any decrease in uncertainty around λ_{1j} and λ_{0j} across classes. This is because the between treatment classes heterogeneity was relatively large for TR-PFS pair and hence there was not much shrinkage. Furthermore, using subgroup analysis, the surrogacy criteria failed in the anti-EGFR class (zero was included in the 95% CrI of the slope) where only eight studies available). However, the 95% CrI in the anti-EGFR class just contains zero and overlaps substantially with the 95% CrI of the slope for chemotherapy treatment class. By applying P-EX and F-EX models, we were able to draw different inferences for the surrogacy in the anti-EGFR class as these methods

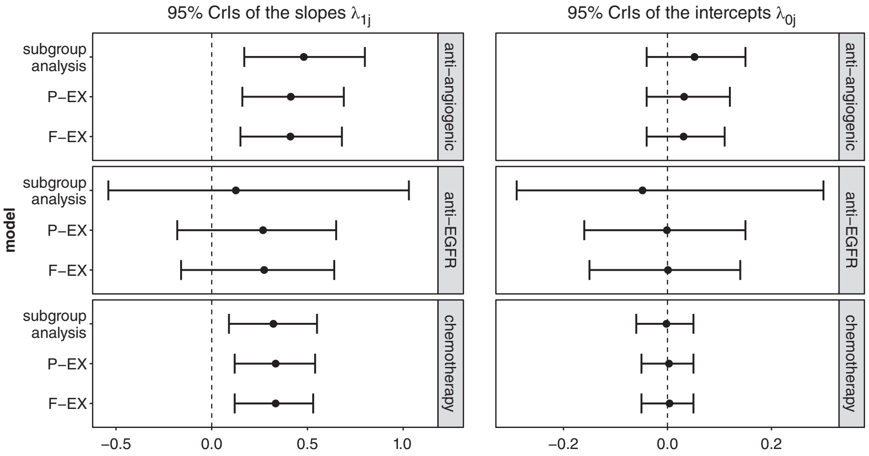


FIGURE 2 95% Credible intervals of λ_{1j} and λ_{0j} for the progression-free survival-overall survival pair of outcomes

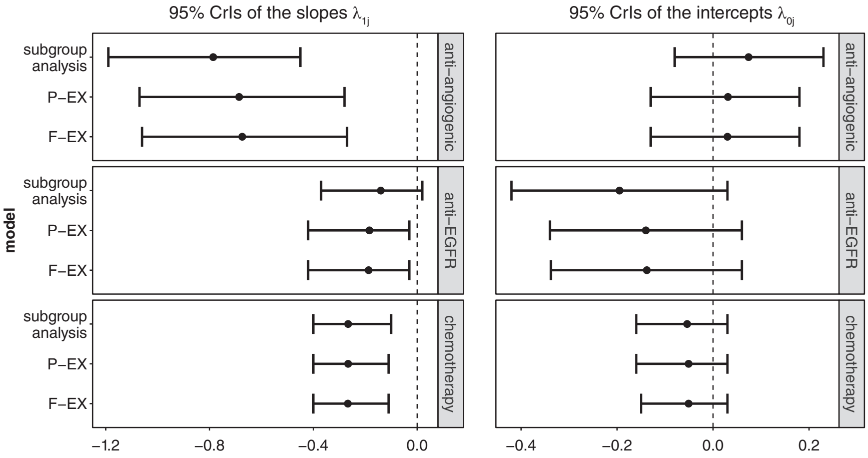


FIGURE 3 95% Credible intervals of λ_{1j} and λ_{0j} for the tumor response-progression-free survival pair of outcomes

allow for borrowing of information for the parameters describing the surrogate relationships from the other classes. As illustrated in Figure 3, both hierarchical models moved the 95% CrI of the slope in the direction of the CrIs of the other two classes resulting in the surrogacy criteria being satisfied across all treatment classes.

When carrying out cross-validation procedure, we wish to ensure that not only predictive intervals contain the observed values but also that they are sufficiently narrow. In general, adding a hierarchical structure to slopes and intercepts reduces the uncertainty and leads to more precise predictions compared to those obtained from subgroup analysis. For the PFS-OS pair of outcomes, the accuracy of the predictions was very similar across all methods (similar absolute error) but the uncertainty varied depending on the level of borrowing of information. F-EX model gave on average the most precise estimates ($\hat{\mu}_{2ij}$) having the narrowest 95% predictive intervals of the effect on the final outcome (smallest width ratio seen in Tables 3, 5, and 6) reducing the overall uncertainly by 5%. The benefit was smaller in the chemotherapy class where the number of studies was much larger compared to the anti-EGFR treatment class where we had only eight studies available. Overall, P-EX performed better than subgroup analysis and equally well with F-EX regarding the uncertainty of the predictions. This indicates that the assumption of exchangeability seems to be plausible for this pair of outcomes and P-EX model was able to identify this.

For the TR-PFS pair, subgroup analysis with the standard model was a robust approach in terms of the accuracy of its predictions. Although the overall absolute error was very similar across models, F-EX and P-EX yielded higher absolute error compared to subgroup analysis in the anti-angiogenic class. This implies that the posterior means of the true effects were to some extent ”overshrunk” due to excessive borrowing of information from the other classes. P-EX model was implemented allowing for partial exchangeability of the slopes only, this decision is likely to affect the performance of the model in terms of its predictions on TR-PFS pair of outcomes. However, the model can be extended allowing for partial exchangeability also of the intercepts or the conditional variances and different combinations of these assumptions can be explored and models compared using deviance information criterion (DIC). Similarly, there was no significant decrease in the degree of uncertainty of the estimates $\hat{\mu}_{2ij}$ of F-EX and P-EX models. The results indicate that the hierarchical methods performed slightly better compared to subgroup analysis in terms of uncertainty only in the class of chemotherapy and the anti-EGFR treatment class giving 1.5% and 3% narrower predictive intervals, respectively. This kind of behavior might be caused by the relatively large between treatment class heterogeneity and the assumption of full exchangeability of the intercepts.

7 | DISCUSSION

We developed two hierarchical models allowing to account for distinct treatment classes when examining the surrogate relationships. The proposed models may be particularly useful in surrogate endpoint evaluation in complex diseases where different treatment classes of different mechanism of action and potential different association patterns within those classes exist. These models investigate potential differences in study-level surrogacy across treatment classes in a particular disease area and can help to identify treatment classes with strong association patters, even when data are relatively sparse. F-EX model is somewhat restrictive, assuming full exchangeability of the parameters describing the surrogate relationships across treatment classes. In many situations the assumption of exchangeability may be too strong given the heterogeneity between treatment classes. In such circumstances, a more flexible model such as P-EX may be a better choice. P-EX model can infer an appropriate level of borrowing of information from the data, reducing the degree of borrowing of information through the mixture weights, thus relaxing the assumption of exchangeability when it is not fully reasonable. It evaluates whether the association pattern between treatment effects (logHR or logOR) on the surrogate and the final endpoint in a specific treatment class differs from the other patterns in other classes.

F-EX model is appropriate only when the degree of similarity of surrogate relationships is relatively high. It can offer substantial gains in precision, reduced RMSE of the posterior means of the parameters describing surrogate relationships and it can improve the predictions of the true effects on the final endpoint. For example, F-EX model gave posterior means of the slopes and predicted effects with reduced uncertainty (smaller CrIs) compared to subgroup analysis for the first simulated data scenario and for the illustrative example on PFS-OS pair where the parameters describing the surrogate relationship were similar and the assumption of full exchangeability was reasonable. These findings are consistent with the results from other hierarchical Bayesian methods which assume full exchangeability and were developed in other research areas.[24,25] However, P-EX model achieves the same degree of borrowing of information in such data scenarios making less assumptions compared to F-EX model. P-EX model regulates the degree of borrowing of information using its exchangeable and nonexchangeable components with respective mixture weights. For instance, when between treatment class heterogeneity is relatively large or there is a treatment class with distinctly different pattern, P-EX model has the advantage of avoiding the excessive borrowing of information, as illustrated in the second design of the simulation study. All the above illustrate the benefits of partial exchangeability, as described by Neuenschwander et al[17] in their work. Subgroup analysis using the standard model is a simple approach which performs well when there are sufficient data available for each treatment class, but it produces estimates with higher bias and uncertainty when data within a treatment class are limited.

Although the proposed methods provide additional robustness to the CrIs and the posterior means of the parameters describing the surrogate relationships compared to subgroup analysis, potential limitations should always be kept in mind. First, in real data scenarios it can be challenging to find datasets with sufficient number of treatment classes. The small number of treatment classes can affect the performance of hierarchical methods substantially[34] reducing the impact of borrowing of information. For instance, fitting P-EX model to the illustrative example (in aCRC with three treatment classes) led to a situation where in some of the MCMC iterations only one class was deemed exchangeable by the model which is not possible since there were no other classes to exchange information with. However, in our example it did not affect the performance of the model as it occurred only in the 0.5% of the MCMC iterations. On the other hand, there is no upper limit to the number of classes we can have. In general, the more classes the better it is for the models to borrow information across them.

Another limitation of the illustrative example is that treatment switching was applied in a subset of trials in this dataset. Patients were allowed to switch from the treatment that was initially assigned to them to the other treatment arm in the trial. Most commonly patients switched after progression from control to experimental arm in particular, if there was sufficient evidence during the trial that the experimental treatment was better than control.[35] Treatment switching has diminishing effect on the difference in treatment effects on OS when applying intention-to-treat analysis, and the effect is often obtained with larger uncertainty. This makes the estimation of surrogacy between treatment effects on the surrogate and treatment effects on the final outcome very challenging. Many adjustment methods have been proposed, however, their validity is often questionable.[35] Additionally, the evaluation of PFS as a surrogate endpoint is distinctive compared to other surrogate endpoints as PFS can be considered as nested outcome within OS outcome. These factors may explain the different findings for the two pairs of outcomes (PFS-OS and TR-OS).

Furthermore, as it was mentioned in Section 6, each treatment class consist of studies with multiple treatment comparisons. According to Daniels and Hughes[4] and Shanafelt et al[36] different treatment comparisons and the use of active or inactive control interventions may influence the surrogate relationship. This could potentially be resolved by classifying treatment according the treatment class comparison (eg, anti-angiogenic therapies versus chemotherapy) which potentially would lead to more treatment classes, but with reduced number of studies per class. To continue with the same issue, in this paper the treatment classes were defined according to the class of the experimental treatment regardless of the control. Alternatively, we could classify them according to the treatment contrasts taking into account the class of the control group, however, this could result in fewer studies per class. A network meta-analysis model was developed for this problem by Bujkiewicz et al.[37]

Additionally, the evaluation framework proposed by Daniels and Hughes (see Section 2.1) examine whether zero is contained in the CrIs of λ_1 and λ_0. However, the sparsity of data may lead to increased uncertainty around the intercept and slope. This increased uncertainty is also likely to manifest itself in increased conditional variance, thus invalidating the third criterion. Unsurprisingly, for sparse data it is unlikely that all the surrogacy criteria hold and this problem is more likely to occur in subgroup analyses. Our proposed methods alleviate this problem as shown in some of the scenarios of the simulation study. However, we used the criteria mainly for the purpose of model comparison. In real-life scenarios, when evaluating a potential surrogate endpoint for use in clinical trials or regulatory decision-making, the decision of whether the surrogate endpoint may be used to make the prediction of the clinical benefit should be based on the balance between the strength of the surrogate relationship and the need for the decision to be made about the effectiveness of the new treatment.[38] Moreover, the strength (or weakness) of the surrogate relationship will manifest itself in the width of the predicted interval of the treatment effect on the final outcome. A larger interval around the intercept and slope will result in a larger interval around the predicted effect and hence increased uncertainty about the regulatory or clinical decision made based on such prediction. The implication of this is that perhaps we do not need precise surrogacy criteria and instead we need only look at the predictions.[37] The quality of predictions can be evaluated through a cross-validation procedure (see Section 2.2).

A possible extension of these methods is to add another layer of hierarchy accounting for the different treatments within a treatment class. However, a relatively large number of studies for each treatment and number of treatments per class would be required to fit such model. As we mentioned in Section 6.4, P-EX model could also be extended by making additional partial-exchangeability assumptions about the intercepts and the conditional variances, however, this may lead to over-parameterizing the model. Furthermore, taking advantage of the setting proposed by Bujkiewicz et al,[39] both hierarchical models can be extended to allow for modeling multiple surrogate endpoints (or the same surrogate endpoint but reported at multiple time points) as joint predictors of treatment effect on the final outcome.

Further research is also needed to extend the proposed methodology to binomial data or to time to event data where the assumption of normality is not plausible. Moreover, to overcome the convergence issues caused by vague prior distributions on the hyper-parameter of the mixture weights (π_j), alternative prior distributions should be developed by extending the P-EX in a similar way as proposed by Kaizer et al.[40]

In summary, we developed hierarchical Bayesian methods for evaluating surrogate relationships within treatment classes while borrowing of information for surrogate relationships across treatment classes. We believe that the proposed methods have a lot of potential for improving the validation of surrogate endpoints in the era of personalized medicine, where the surrogacy may depend on the mechanism of action of specific targeted therapies.

ACKNOWLEDGEMENTS

This research used the ALICE/SPECTRE High Performance Computing Facility at the University of Leicester and was partly funded by the Medical Research Council, grant no. MR/L009854/1 awarded to Sylwia Bujkiewicz. Keith Abrams is partially supported as a UK National Institute for Health Research (NIHR) Senior Investigator Emeritus (NI-SI-0512-10159). We also thank the anonymous reviewers for their comments, which helped to improve the quality of the manuscript.

ORCID

Tasos Papanikos https://orcid.org/0000-0001-8971-6221
Nicolas Städler https://orcid.org/0000-0001-8212-7459
Oriana Ciani https://orcid.org/0000-0002-3607-0508
Sylwia Bujkiewicz https://orcid.org/0000-0002-3003-9403

REFERENCES

1. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. New York: Springer Science & Business Media; 2006.
2. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Stat Med*. 2012;31(25):2973-2984.
3. Phillips A, Haudiquet V. ICH E9 guideline Statistical principles for clinical trials: a case study. *Stat Med*. 2003;22(1):1-11.
4. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med*. 1997;16(17):1965-1982.
5. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000;1(1):49-67.

6. Bujkiewicz S, Thompson JR, Spata E, Abrams KR. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. *Stat Methods Med Res*. 2015;26(5):2287-2318.
7. Lassere MN. The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Stat Methods Med Res*. 2008;17(3):303-340.
8. Buyse M, Burzykowski T, Carroll K, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *J Clin Oncol*. 2007;25(33):5218-5224.
9. Giessen C, Laubender RP, Ankerst DP, et al. Progression-free survival as a surrogate endpoint for median overall survival in metastatic colorectal cancer: literature-based analysis from 50 randomized first-line trials. *Clin Cancer Res*. 2013;19(1):225-235.
10. Ciani O, Buyse M, Garside R, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. *J Clin Epidemiol*. 2015;68(7):833-842.
11. Chirila C, Odom D, Devercelli G, et al. Meta-analysis of the association between progression-free survival and overall survival in metastatic colorectal cancer. *Int J Color Dis*. 2012;27(5):623-634.
12. Macedo M, Melo FM, Ribeiro H, et al. KRAS mutation status is highly homogeneous between areas of the primary tumor and the corresponding metastasis of colorectal adenocarcinomas: One less problem in patient care. *Am J Cancer Res*. 2017;7:1978-1989.
13. Perez K, Walsh R, Brilliant KE, et al. Heterogeneity of colorectal cancer (CRC) in reference to KRAS proto-oncogene utilizing wave technology. *J Clin Oncol*. 2013;31(suppl 15):e14637-e14637.
14. Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. *J Am Stat Assoc*. 1975;70:311-319.
15. Louis TA. Estimating a population of parameter values using bayes and empirical bayes methods. *J Am Stat Assoc*. 1984;79(386):393-398.
16. Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. New York: Chapman and Hall/CRC; 2010
17. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm Stat*. 2016;15(2):123-134.
18. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160.
19. Berry DA. Subgroup analyses. *Biometrics*. 1990;46(4):1227-1230.
20. Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: statistical and regulatory issues. *J Biopharm Stat*. 2005;15(5):869-882. https://doi.org/10.1081/BIP-200067988.
21. Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Am Stat Assoc*. 1995;90(431):928-934.
22. Verdinelli I, Wasserman L. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J Am Stat Assoc*. 1995;90(430):614-618.
23. Jeffreys H. *The Theory of Probability*. Oxford, UK: Oxford University Press; 1998.
24. Berry SM, Broglio KR, Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clin Trials*. 2013;10(5):720-734. https://doi.org/10.1177/1740774513497539.
25. TP F, Kyle WJ, Nebiyou BB, CR E, BL H, BR S. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Stat Med*. 2003;22(5):763-780. https://doi.org/10.1002/sim.1399.
26. Chugh R, Wathen JK, Maki RG, et al. Phase II multicenter trial of imatinib in 10 histologic subtypes of sarcoma using a Bayesian hierarchical statistical model. *J Clin Oncol*. 2009;27(19):3148-3153.
27. Sturtz S, Ligges U, Gelman A. R2WinBUGS: a package for running WinBUGS from R. *J Stat Softw*. 2005;12(3):1-16.
28. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *JNCI*. 2000;92(3):205-216.
29. World Health Organization. Geneva: World Health Organization. https://apps.who.int/iris/handle/10665/37200 (accessed Jan 5, 2020)
30. Bennouna J, Sastre J, Arnold D, et al. Continuation of bevacizumab after first progression in metastatic colorectal cancer (ML18147): a randomised phase 3 trial. *Lancet Oncol*. 2013;14(1):29-37.
31. Rothenberg M, Cox J, Butts C, et al. Capecitabine plus oxaliplatin (XELOX) versus 5-fluorouracil/folinic acid plus oxaliplatin (FOLFOX-4) as second-line therapy in metastatic colorectal cancer: a randomized phase III noninferiority study. *Ann Oncol*. 2008;19(10):1720-1726.
32. Cassidy J, Clarke S, Dıaz-Rubio E, et al. XELOX vs FOLFOX-4 as first-line therapy for metastatic colorectal cancer: NO16966 updated results. *Br J Cancer*. 2011;105(1):58.
33. Hurwitz H, Fehrenbacher L, Novotny W, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *N Engl J Med*. 2004;350(23):2335-2342.
34. McNeish D, Stapleton LM. Modeling clustered data with very few clusters. *Multivar Behav Res*. 2016;51(4):495-518.
35. Latimer NR, Henshall C, Siebert U, Bell H. Treatment switching: statistical and decision-making challenges and approaches. *Int J Technol Assess Health Care*. 2016;32(3):160-166.
36. Shanafelt TD, Loprinzi C, Marks R, Novotny P, Sloan J. Are chemotherapy response rates related to treatment-induced survival prolongations in patients with advanced cancer. *J Clin Oncol*. 2004;22(10):1966-1974.
37. Bujkiewicz S, Jackson D, Thompson JR, et al. Bivariate network meta-analysis for surrogate endpoint evaluation. *Stat Med*. 2019;38(18):3322-3341.
38. Alonso A, Bigirumurame T, Burzykowski T, et al. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. New York: CRC Press; 2016.
39. Bujkiewicz S, Thompson JR, Riley RD, Abrams KR. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Stat Med*. 2015;35(7):1063-1089.

40. Kaizer AM, Koopmeiners JS, Hobbs BP. Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics*. 2017;19(2):169-184.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Papanikos T, Thompson JR, Abrams KR, et al. Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data. *Statistics in Medicine*. 2020;39:1103–1124. https://doi.org/10.1002/sim.8465