BAYESIAN MULTI-REGIONAL CLINICAL TRIALS USING MODEL AVERAGING

Nathan W. Bean

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Joseph Ibrahim

Matthew Psioda

Lisa LaVange

Fang Chen

Lisa Carey

ABSTRACT

Nathan W. Bean: Bayesian Multi-Regional Clinical Trials Using Model Averaging
(Under the direction of Joseph Ibrahim and Matthew Psioda)

Multi-regional clinical trials (MRCTs) provide the benefit of more rapidly introducing drugs to the global market; however, small regional sample sizes can lead to poor estimation quality of region-specific effects when using current statistical methods. With the publication of the International Conference for Harmonisation E17 guideline in 2017, the MRCT design is recognized as a viable strategy that can be accepted by regional regulatory authorities, necessitating new statistical methods that improve the quality of region-specific inference. We develop novel methodology using Bayesian model averaging (BMA) to estimate region-specific and global treatment effects for MRCTs that compare two treatment groups with respect to a continuous outcome, a time-to-event (TTE) outcome, or both a TTE outcome and a longitudinal marker jointly. This approach accounts for the possibility of heterogeneous treatment effects between regions, and we discuss how to assess the consistency of these effects using posterior model probabilities.

In the case of a continuous or TTE endpoint, we show through simulation studies that the proposed modeling approaches estimate region-specific treatment effects with lower mean squared error than commonly used models (e.g., fixed effects linear regression models, Cox proportional hazards models) while resulting in higher rejection rates of the global treatment effect compared to Bayesian hierarchical models. For both types of endpoints, we further develop three measures to evaluate the consistency of the treatment effect across regions. These three approaches quantify the strength of evidence that a lack of clinically relevant differences exists between treatment effects (1) among regions overall, (2) for any two regions, and (3) for any given region versus all other regions together.

When jointly modeling a TTE endpoint and an associated longitudinal marker, we show that the BMA approach can result in an increase in the global rejection rate compared to survival models that account for only the TTE endpoint. We then apply both the survival and joint model variations of the BMA approach to data from the LEADER trial, an MRCT designed to evaluate the cardiovascular safety of an anti-diabetic treatment.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors, Dr. Joseph Ibrahim and Dr. Matthew Psioda, for their continual guidance, support, and patience over the past four years. In addition to their pivotal role in my thesis work, they have helped me to set high expectations for myself and provided me with numerous opportunities to further develop a skill set that will aid me throughout my future. I would also like to thank Dr. Lisa LaVange, Dr. Fang Chen, and Dr. Lisa Carey for being willing to serve on my doctoral committee and for their feedback in helping to shape my thesis work.

I especially would like to thank my parents and several close friends, each of whom have provided tremendous support through all stages of my schooling career. They were always willing to share both my stress and my excitement at all times, even if it meant having to tolerate and attempt to understand lengthy explanations of my thesis research.

Furthermore, I would like to express my deepest gratitude for friends and fellow students in the Department of Biostatistics. Specifically, I thank Ethan Alt, Hillary Heiling, Bonnie Shook-Sa, and Kimi Enders who were crucial in helping me survive the core classes. I would also like to thank John Kidd and his wife, Cassie, who welcomed me from the beginning and provided both support and encouragement in the form of pep talks and homemade dinners.

Lastly, I would like to thank all who have helped make this dissertation possible: the National Institute of Environmental Health Sciences for providing four years of funding (NIH Grant T32ES007018), managers at my internships with Precision BioSciences and Dova Pharmaceuticals for giving me practical experience, and Novo Nordisk for providing access to data from the LEADER trial which serves as the primary motivation for this thesis work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BHM	Bayesian hierarchical model
BMA	Bayesian model averaging
BMA-JM	BMA joint modeling approach
BMA-S	BMA approach for survival data only
COPD	chronic obstructive pulmonary disease
CPHM	Cox proportional hazards model
FELM	fixed effects linear model
FPR	false positive rate
HbA1c	glycated hemoglobin
HR	hazard ratio
ICH	International Council for Harmonisation
LEADER	Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results
MACE	major adverse cardiovascular events
MCIRD	minimal clinically important regional difference
MRCT	multi-regional clinical trial
MSE	mean squared error
PMP	posterior model probability
TPR	true positive rate
TTE	time to event

CHAPTER 1: INTRODUCTION

In recent years, multi-regional clinical trials (MRCTs), or trials with multiple geographic regions included in the same study protocol, have increased in popularity in the pharmaceutical industry due to their ability to accelerate the global drug development process. Sponsors can simultaneously collect data from different regions and seek approval for an investigational treatment from the corresponding regulatory authorities. While regional differences in the treatment effect may exist due to either intrinsic or extrinsic factors, commonly used statistical models typically estimate only a global treatment effect without accounting for differences between regions in the primary analysis. Instead, region-specific treatment effects are often estimated in subgroup analyses using only data from the respective region, and these estimates can be susceptible to high variation due to small regional sample sizes.

To address potential challenges with MRCTs, the International Council for Harmonisation released the E17 guidance document which suggests the use of statistical methods that utilize information borrowing across regions if regional sample sizes are small (ICH, 2017). We address this guideline by developing an approach that allows for information borrowing using Bayesian model averaging (BMA) to analyze data from MRCTs with two treatment arms. Additionally, we develop three methods to quantify evidence in favor of consistency of the treatment effects across regions, which is defined as the lack of clinically meaningful differences between region-specific treatment effects.

In Chapter 2, we provide a literature review on models that have previously been proposed to analyze data from MRCTs. We also discuss examples from the vast literature concerning the evaluation of consistency.

In Chapter 3, we first detail the methodology of the proposed BMA approach in the context of MRCTs with a continuous endpoint. Through simulation studies, we show that the BMA approach results in better estimation quality of the region-specific treatment effects compared to a fixed effects linear model. Unlike the Bayesian hierarchical model, which also incorporates information borrowing across regions, we demonstrate the robustness of the prior elicitation for the BMA approach with respect to the global rejection rate.

In Chapter 4, we extend the BMA approach to MRCTs with a time-to-event (TTE) endpoint. We obtain posterior samples of the treatment effects using a Laplace approximation, and we again demonstrate through simulation studies that the proposed modeling approach estimates region-specific treatment effects with lower mean squared error than a Cox proportional hazards model while resulting in a similar rejection rate of the global treatment effect. We then apply the BMA approach to data from the LEADER trial, an MRCT designed to evaluate the cardiovascular safety of an anti-diabetic treatment.

In Chapter 5, we further develop the BMA approach in the context of a joint analysis of survival and longitudinal data from MRCTs. In this novel application of joint models to MRCTs, we use Laplace's method to integrate over subject-specific random effects and to approximate posterior distributions for region-specific treatment effects on the time-to-event outcome. We conduct simulation studies to compare this joint modeling approach to methods that analyze survival data alone, and we demonstrate that the joint modeling approach can result in an increased rejection rate when testing the global treatment effect. We again apply the proposed approach to data from the LEADER trial by jointly analyzing repeated HbA1c measurements with the time-to-event primary endpoint.

CHAPTER 2: LITERATURE REVIEW

2.1 Proposed Models for MRCTs

Generally, treatment effects on the outcome of interest for multi-regional clinical trials (MRCTs) are tested and estimated using statistical models with fixed effects. Regardless of the type of endpoint, the global treatment effect is typically estimated in the primary analysis without accounting for potential differences between regions, and region-specific treatment effects are estimated as part of subgroup analyses. The total sample size of an MRCT is calculated to achieve a desired level of power with respect to the global treatment effect, but the small regional sample sizes can result in extreme estimates of the region-specific treatment effects due to chance. Such behavior has been observed in high profile MRCTs, including the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results (LEADER) trial (Marso *et al.*, 2016) and the Platelet Inhibition and Patient Outcomes (PLATO) trial (Mahaffey *et al.*, 2011). Both MRCTs were cardiovascular outcomes trials that reported statistically significant results when testing the global treatment hazard ratio, however, both trials found that the region-specific hazard ratio for North America (and the U.S. in particular) had a reversal of direction when compared to the global effect (i.e., greater than 1.0). While further investigations and post hoc analyses suggest that these extreme results for North America were likely due to chance (Yusuf and Wittes (2016)), these examples illustrate the challenges that can arise when using fixed effects models to estimate region-specific treatment effects.

In this section, we provide further details on some of the fixed effects models commonly used to analyze data from MRCTs with different endpoints. Additionally, we discuss other models that have been proposed for MRCTs.

2.1.1 Models for Continuous Outcomes

While fixed effects models are commonly used in practice, the use of random effects models where region-specific treatment effects are treated as random variables has also been suggested (Hung *et al.*, 2010). This approach overcomes limitations of fixed effects models by accounting for possible regional differences in the treatment effect. Let N be the total sample size and $n_i = r_i N$ the sample size of the ith region where $0 < r_i < 1$, $i = 1, \ldots, S$, and $\sum_{i=1}^{S} r_i = 1$ (and thus $\sum_{i=1}^{S} n_i = N$). Additionally, let γ denote the common effect size, and let γ_i and $\hat{\gamma}_i$ be the true and the estimated effect sizes, respectively, for the ith region. Assume that each region has an equal variance, $\sigma^2/4$, in both treatment groups. Under a normal random effects model, we assume

$$\hat{\gamma}_i | \gamma_i \sim N\left(\gamma_i, \frac{\sigma^2}{n_i}\right),$$

$$\gamma_i \sim N\left(\gamma, \sigma_\gamma^2\right),$$

where σ_γ is the between-region standard deviation of the regional effect sizes. If σ_γ is large, then γ may not be interpretable, and instead γ_i is applicable to the ith region. Consider the case when σ_γ is too large to be ignored, and we want to detect a treatment effect size of $\gamma = \delta > 0$ at significance level α and power $1 - \beta$. To do so, it must hold that $\sigma_\gamma / \delta < \left\{(z_\alpha + z_\beta)\sqrt{\sum_{i=1}^{S} r_i^2}\right\}^{-1}$, where z_α is the $(1-\alpha)$th percentile of the standard normal distribution. Under this condition, the total sample size should be calculated as

$$N = \left[\left(\frac{\delta}{\sigma(z_\alpha + z_\beta)}\right)^2 - \left(\frac{\sigma_\gamma}{\sigma}\right)^2 \sum_{i=1}^{S} r_i^2\right]^{-1}.$$

If we assume no regional differences (i.e., $\sigma_\gamma = 0$), then the resulting total sample size may be greatly underestimated if the between-region variance is in fact large.

Rothmann (2021) also suggested treating region-specific effects as random with the use of a Bayesian hierarchical model (BHM). The BHM allows for information borrowing across regions by shrinking region-specific effects towards the hierarchical mean, which can be used as

a logical estimator for the global treatment effect. Unlike fixed effects models, the BHM often avoids extreme estimates of the region-specific treatment effects, however, this model requires the possibly unrealistic assumption that these effects are a random sample from some underlying normal distribution.

In contrast to traditional random effects models, Wu *et al.* (2014) proposed a variation of the model in which they assume that the shift in variance across regions is random and that the mean response is fixed. Lan and Pinheiro (2012) develop a discrete random effects model (DREM) for MRCTs with a continuous outcome, and Lan *et al.* (2014) extend this method to MRCTs with either TTE or binary endpoints. They assume that regional sample sizes are a random sample from an underlying multinomial distribution, however, this assumption may not be practical considering that regional sample sizes are often determined in part by minimum sample size requirements set by regulatory authorities.

Chiang and Hsiao (2019) question altogether the use of random effects models in which regional differences are treated as random effects, arguing that regional intrinsic and extrinsic factors are generally known and should be considered as fixed. Further, random effects models poorly estimate the variability between levels of the grouping variable when the number of levels is small, as is often the case with the number of regions in MRCTs.

2.1.2 Models for Time-to-Event Outcomes

The proportional hazards model is usually used for MRCTs with time-to-event (TTE) endpoints (Quan *et al.*, 2010b). First introduced by Cox (1972), the proportional hazards model that compares an active treatment to a control is written as

$$\lambda_1(t) = \lambda_0(t)\exp(\gamma),$$

where $\lambda_1(t)$ and $\lambda_0(t)$ are the baseline hazards for the treatment group and control group, respectively, and $\exp(\gamma)$ is the hazard ratio between the treatment group and control group.

As described by Quan *et al.* (2010b), the power calculations are often based on the log-rank test. Let E be the expected total number of events from both groups. For a two-sided significance level α and power $1-\beta$, the expected total number of events are calculated as

$$E = \frac{4(z_{1-\alpha/2} - z_{1-\beta})^2}{\gamma^2},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$th percentile of the standard normal distribution.

Consider the case with S regions, and let γ_i be the region-specific treatment effect for the ith region, $i = 1, \ldots, S$. The estimated region-specific treatment effect $\hat{\gamma}_i$ is asymptotically distributed as $\hat{\gamma}_i \sim N(\gamma_i, 4/E_i)$, were E_i is the expected number of events in the ith region. The estimated overall treatment effect is defined as

$$\hat{\gamma} = \sum_{i=1}^{S} \frac{E_i}{E} \hat{\gamma}_i, \tag{2.1}$$

with an asymptotic distribution of $\hat{\gamma} \sim N\left(\sum_{i=1}^{s} \frac{E_i}{E}\gamma_i, 4/E\right)$.

2.1.3 Models for a Joint Time-to-Event Outcome and Longitudinal Marker

The use of statistical models that allow for information borrowing across regions can be particularly beneficial for MRCTs with both a TTE endpoint and some longitudinal marker, as is often the case with phase III oncology trials. Due to high mortality rates and unmet medical needs in oncology drug development, drug sponsors often rely on MRCTs to introduce cancer treatments more rapidly into the global market. Song *et al.* (2019) assessed all oncology clinical trials by the top ten pharmaceutical companies between 1 January 2008 and 31 December 2017, and they found that the 65.0% of phase II and 81.8% of phase III trials were MRCTs. Most phase III oncology MRCTs are powered using results with surrogate endpoints (e.g., response rate) from phase 1b or 2a studies which may not be representative of primary TTE endpoints (e.g., overall survival, progression-free survival). This, along with faster speeds common in oncology drug development, can result in phase III trials being underpowered. Wong *et al.* (2019)

estimated the probability of success for phase III oncology trials to be only 35.5%. The high failure rate for phase III oncology trials makes clear the need for new statistical methods that can increase power when testing the global treatment effect with respect to a TTE primary endpoint in MRCTs, and such an objective can be achieved by using joint models to account for possible associations between survival data and longitudinal data (e.g., quality of life assessments).

To the best of the author's knowledge, the application of joint models to MRCTs has not yet been proposed. However, the methodology of joint models continues to be developed in other contexts. In a joint model, TTE data and longitudinal data are typically linked together using one of three methods: (1) including the observed longitudinal outcomes as covariates in the survival submodel, (2) including the fitted values from a longitudinal submodel as covariates in the survival submodel, or (3) linking the longitudinal and survival submodels with shared parameters. The majority of developments in the research focus on the third method in which both submodels share common subject-specific random effects, and it has been shown that treatment effects are generally estimated with less bias when using this approach compared to the first two methods (Sweeting and Thompson, 2011). When a TTE outcome and a longitudinal marker are strongly correlated, joint models can identify significant treatment effects with greater sensitivity compared to survival models that ignore this underlying association (Gould *et al.*, 2015). Further, joint models can result in higher power and lower sample sizes when testing treatment effects on either the TTE or longitudinal outcomes (Ibrahim *et al.*, 2010).

Typically, a linear mixed model is used to analyze the longitudinal data. If we define $X_i(t)$ to be the observed longitudinal outcome for the ith subject at time t and $X_i^*(t)$ to be a trajectory function that depends on subject-specific random effects $\boldsymbol{b}_i$, then we can formulate the longitudinal submodel as

$$X_i(t) = X_i^*(t) + \epsilon_i(t),$$

where $\epsilon_i(t) \sim N(0, \sigma^2)$. We assume the errors $\epsilon_i(t)$ are independent from one another and from $\boldsymbol{b}_i$, and we also assume the random effects are normally distributed.

For the survival submodel, we fit the TTE data using a proportional hazards model. At time t, we write this submodel as

$$h\left(t|\, X_i^*, \boldsymbol{w}_i\right) = h_0(t) \exp\left\{g\left(\boldsymbol{\alpha}, X_i^*\right) + \boldsymbol{w}_i'\boldsymbol{\theta}\right\},$$

where $h_0(t)$ is the baseline hazard function, $g(\cdot)$ is a function that defines the association structure, $\boldsymbol{\alpha}$ measures the association between the longitudinal marker and the TTE endpoint, and $\boldsymbol{w}_i$ is a vector of covariates for the ith subject with corresponding effects $\boldsymbol{\theta}$. Different association structures have been proposed to connect the longitudinal and survival submodels, including the use of the entire trajectory function (i.e., $g\left(\boldsymbol{\alpha}, X_i^*\right) = \alpha X_i^*$) or only shared random effects (i.e., $g\left(\boldsymbol{\alpha}, X_i^*\right) = \boldsymbol{\alpha}'\boldsymbol{b}_i$).

Under the Bayesian framework, Faucett and Thomas (1996) propose a joint model with a piecewise constant baseline hazard, and they link the two submodels by including the longitudinal trajectory function in the linear predictor of the survival submodel. This model has been extended in several ways, such as to allow for greater flexibility in the structure of the longitudinal submodel (Wang and Taylor, 2001; Brown and Ibrahim, 2003) or to accommodate for multivariate longitudinal and survival data (Ibrahim *et al.*, 2004; Chi and Ibrahim, 2006). In practical contexts, the application of Bayesian joint models has become more accessible with the development of R packages such as `JMbayes` (Rizopoulos, 2020) and `rstanarm` (Gabry *et al.*, 2022)

2.2 Theory and Application of Bayesian Model Averaging

Bayesian model averaging (BMA) is typically used as an alternative method to model selection. Rather than use some criterion to select a single model with a subset of covariates of interest, BMA accounts for model uncertainty by averaging together results from all models in the model space, each of which fits a different sets of covariates (Hoeting *et al.*, 1999). We discuss the general theory of BMA, and we detail a novel application of BMA to basket trials.

2.2.1 Theory of Bayesian Model Averaging

Let $\mathcal{M}_L$ denote a model space with L possible models, and let M_ℓ denote the ℓth model with prior model probability $p(M_\ell)$, $\ell = 1, \ldots, L$. Further, let $\mathbf{D}$ denote the observed data and $\boldsymbol{\theta}_\ell$ the vector of parameters for M_ℓ. Under this model, we write the likelihood of $\mathbf{D}$ and the prior for $\boldsymbol{\theta}_\ell$ as $p(\mathbf{D}|\boldsymbol{\theta}_\ell, M_\ell)$ and $p(\boldsymbol{\theta}_\ell|M_\ell)$, respectively, and the marginal likelihood of M_ℓ is given by $p(\mathbf{D}|M_\ell) = \int p(\mathbf{D}|\boldsymbol{\theta}_\ell, M_\ell)p(\boldsymbol{\theta}_\ell|M_\ell)d\boldsymbol{\theta}_\ell$. The posterior model probability (PMP) for M_ℓ is calculated as

$$p(M_\ell|\mathbf{D}) = \frac{p(\mathbf{D}|M_\ell)p(M_\ell)}{\sum_{\ell'=1}^{L} p(\mathbf{D}|M_{\ell'})p(M_{\ell'})}. \tag{2.2}$$

With the PMPs as weights, we obtain model averaged posterior quantities $p(\boldsymbol{\theta}|\mathbf{D})$ as

$$p(\boldsymbol{\theta}|\mathbf{D}) = \sum_{\ell=1}^{L} p(\boldsymbol{\theta}_\ell|M_\ell, \mathbf{D})p(M_\ell|\mathbf{D}). \tag{2.3}$$

2.2.2 Bayesian Model Averaging Applied to Basket Trials

Psioda *et al.* (2021) apply BMA to oncology basket trials with a binary endpoint. For a given number of baskets, the model space includes models that correspond to all possible partitions of the baskets into sets in which baskets within the same set are constrained to share a common binary response rate (i.e., the probability of having a favorable response to the treatment). Under this model fromulation, the two extremes include a model that assumes no differences between the baskets by pooling all baskets together and a model that assumes the response rates of all baskets differ. They then obtain posterior quantities of interest averaged over the models with the PMPs as weights.

Consider the scenario with K distinct baskets and binary response rates. The assumption is made that patients within a basket are independent and share a common response probability. For K distinct baskets, consider all possible ways of classifying the baskets into active and inactive groups, where active baskets have response rates that indicate a desirable treatment effect. We assign a model to each possible grouping, resulting in J models, and we denote the number of

distinct response rates for the jth model by P_j. Let $\pi_{(j,p)}$ be the pth distinct response rate for the jth possible model M_j, $p = 1, \ldots, P_j$, and let $\Omega_{j,p}$ be the set of basket labels corresponding to baskets having the pth distinct response rate for model M_j. Additionally, let $\mathbf{D} = \{y_k, n_k : k = 1, \ldots, K\}$ be the data at a given time of analysis where n_k and y_k are the number of patients and responders, respectively, in the kth basket at that time.

To apply BMA, prior distributions are required for the model space and model parameters. The proposed prior for model M_j is $p(M_j) \propto P_j^{\alpha}$, $j = 1, \ldots, J$, where $\alpha \geqslant 0$ is tuned to achieve the desired amount of information borrowing. For $\alpha = 0$, the prior model probabilities are uniform, resulting in a greater amount of information sharing and more variable true positive rates. As α increases, models with more parameters are given more weight *a priori*, leading to less borrowing and more stable true positive rates. Based on extensive simulation studies, the authors recommend the choice of $\alpha = 2$ to achieve balance in the amount of information sharing.

For the priors on the response rates, let $\pi_{(j,p)}|M_j \sim \text{Beta}(a_0, b_0)$ for each $j = 1, \ldots, J$ and $p = 1, \ldots, P_j$. The authors provide default recommendations for choosing initial values of a_0 and b_0 by setting

$$\frac{a_0}{a_0 + b_0} = \pi_A,$$

where π_A is a hypothesized plausible response rate associated with treatment activity. By setting $a_0 + b_0 = 1.0$, the prior for $\pi_{(j,p)}|M_j$ becomes a weekly informative prior with mean equal to π_A. It follows that the posterior distribution for $\pi_{(j,p)}$ is

$$\pi_{(j,p)}|\mathbf{D}, M_j \sim \text{Beta}(a_{(jp)}, b_{(jp)}),$$

where $a_{(jp)} = a_0 + \sum_{k \in \Omega_{j,p}} y_k$ and $b_{(jp)} = b_0 + \sum_{k \in \Omega_{j,p}} (n_k - y_k)$.

The marginal likelihood of the data, conditional on model M_j being the correct model, is

$$p(\mathbf{D}|M_j) = \prod_{k=1}^{K} \binom{n_k}{y_k} \times \prod_{p=1}^{P_j} \frac{\mathcal{B}(a_{(jp)}, b_{(jp)})}{\mathcal{B}(a_0, b_0)},$$

where $\mathcal{B}(\cdot,\cdot)$ is the complete beta function. We then calculate the PMPs by applying (2.2). For some arbitrary value x, we can calculate the posterior probabilities $P(\pi_k > x|M_j, \mathbf{D})$ for the kth basket, conditional on model M_j, $j = 1,\ldots,J$, $k = 1,\ldots,K$. By (2.3), we calculate the posterior probability of interest for the kth basket as

$$P(\pi_k > x|\mathbf{D}) = \sum_{j=1}^{J} P(\pi_k > x|M_j, \mathbf{D})p(M_j|\mathbf{D}).$$

In the simulation study, the proposed BMA method is compared to basket trial methodologies including Simon's optimal two-stage design in each basket and the calibrated Bayesian hierarchical model (CBHM). With $K = 5$, the authors consider all possible basket groupings into active and inactive baskets, and they also consider scenarios with uniform, fast active, and slow active accrual rates. Compared to the other models, they show that the BMA approach results in well-controlled false positive rates and reasonably high true positive rates with a good balance across the different scenarios. It also generally resulted in lower expected samples sizes and expected trial duration.

Their proposed BMA approach offers several advantages, including the flexibility to stop enrollment in any basket due to futility or efficacy and the ability to quantify the likelihood that two baskets have the same response. With the suggested prior model probabilities, this approach also allows for the control of information borrowing by specifying the choice α. Although BMA can be computationally demanding, the simulation scenarios detailed by the authors allow for closed-form computations, decreasing the computational burden compared to the CBHM.

2.3 Consistency of Treatment Effects Across Regions

Much of the literature on MRCTs focuses on the development of methods to evaluate the consistency of treatment effects across regions, however, consistency is often defined in different ways (Quan *et al.*, 2010a). The International Council for Harmonisation (ICH) defines consistency as "a lack of clinically relevant differences between treatment effects in different regions"

(ICH, 2017), whereas others focus on the comparison of any given region-specific treatment effect to the global treatment effect. We discuss several approaches for the assessment of consistency for MRCTs with a continuous endpoint, a TTE endpoint, or multiple continuous endpoints.

2.3.1 Consistency of Treatment Effects for Continuous Endpoints

When sponsors use an MRCT to show the effectiveness of an investigational treatment, regulatory authorities often require evidence that the global treatment effect can be bridged to the region-specific treatment effect for their corresponding region. The Japanese Medical Devices Agency in the Ministry of Health, Labour and Welfare (MHLW) was among the first to discuss the concept of consistency by establishing guidelines on the number of patients that should be enrolled in Japan. In 2007, the MHLW proposed two methods for sponsors to consider when designing an MRCT (MHLW, 2007). Method 1 requires that enough patients be enrolled in Japan such that the Japan-specific treatment effect is at least half of the overall treatment effect with probability greater than or equal to 0.8; i.e.,

$$P\left(\frac{\gamma_J}{\gamma_{all}} > \pi\right) \geqslant 1-\beta, \tag{2.4}$$

where γ_J denotes the Japan-specific treatment effect, γ_{all} denotes the global treatment effect, and $\pi \geqslant 0.5$ and $1-\beta \geqslant 0.8$. Alternatively, sponsors could consider Method 2 which states that consistency can be assumed if $\gamma_i > 0$ for all S regions, $i = 1, \ldots, S$.

Since the publication of the MHLW guidelines, many authors have proposed extensions of the two methods and developed calculations for determining the sample size of each region (Kawai *et al.*, 2008; Ko *et al.*, 2010; Liu *et al.*, 2016; Quan *et al.*, 2010a,b, 2013; Tsong *et al.*, 2012; Uesaka, 2009). Quan *et al.* (2010a) discuss the limitations of the MHLW guidelines, and they instead extend Method 1 to all regions by ensuring that the ratio of each region-specific treatment effect and the global treatment effect exceeds some prespecified proportion; i.e., $\gamma_1 > \pi\gamma_{all}, \gamma_2 > \pi\gamma_{all}, \ldots, \gamma_S > \pi\gamma_{all}$. For MRCTs with five or more regions, the MHLW's require-

ment of $\pi \geqslant 0.5$ for each region is not practical, and alternative values of π have been proposed (Quan *et al.*, 2014; Teng *et al.*, 2017).

2.3.2 Consistency of Treatment Effects for Time-to-Event Endpoints

Most methods for evaluating consistency of treatment effects in MRCTs are for trials with continuous endpoints with less research addressing TTE endpoints. Quan *et al.* (2010b) adapts the Japanese MHLW's Method 1 in (2.4) to survival data using the risk reduction (i.e., one minus the hazard ratio), rewriting this method as

$$P\left(\frac{1-\exp\left(\hat{\gamma}_i\right)}{1-\exp\left(\hat{\gamma}\right)} > \pi\right) \geqslant 1-\beta, \tag{2.5}$$

where $\hat{\gamma}_i$ is the estimated region-specific treatment effect for some region i and $\hat{\gamma}$ is the estimate for the overall treatment effect in (2.1). The authors show that the left-hand side of (2.5) can be approximated using a normal distribution, and Hayashi and Itoh (2017) provide a variation of their approximation in addition to showing that this probability can be computed using numerical integration. Alternatively, Chen *et al.* (2013) propose the evaluation of consistency through the use of normal probability plots, which, in comparison to other graphical tools such as funnel plots, decrease the false positive rate when identifying outlying countries.

For TTE endpoints, Teng *et al.* (2018) pose two possible requirements when evaluating if the overall hazard ratio (HR) results can be extended to the ith region:

(i) Hazard reduction: $(1-\exp(\hat{\gamma}_i)) > \pi_i\,(1-\exp(\hat{\gamma}_i)) \Leftrightarrow (1-HR_i) > \pi_i(1-HR)$, $0<\pi_i<1, i=1,\ldots,S$;

(ii) Log scale of hazard ratio: $\hat{\gamma}_i > \pi_i\hat{\gamma} \Leftrightarrow HR_i > HR^{\pi_i}$, $0<\pi_i<1, i=1,\ldots,S$.

When comparing the two requirements using the same value of π_i and when the overall results are positive (i.e., $HR<1$), the authors show that the requirement for the log scale of the hazard ratio is more stringent than the hazard reduction requirement. Thus, the authors suggest that the hazard reduction requirement may be preferable when assessing consistency.

2.3.3 Consistency of Treatment Effects for Multiple Co-Primary Endpoints

No methods have been proposed to evaluate consistency for joint models with a TTE outcome and a longitudinal marker, and more generally, little research has been done to evaluate consistency for MRCTs with multiple co-primary endpoints. Huang *et al.* (2017) consider the case of a trial that compares an active treatment to a control with respect to K continuous multiple co-primary endpoints, where the effect size is consistent across all S regions, with $K \geqslant 2$ and $S \geqslant 2$. For simplicity, they assume that the variances of and correlations between the outcomes are known, which is not the case in actual practice. For the kth co-primary endpoint, $k = 1, \ldots, K$, denote the following: D_{ik} is the observed mean difference between the treatment group and control group for the ith region, $i = 1, \ldots, S$, $k = 1, \ldots, K$, D_k^{SC} is the observed mean difference from all other regions without the ith region, and D_k is the observed mean difference from all regions. If the overall mean difference is significant at the α level, they propose evaluating the effectiveness of the treatment in the ith region using three criteria:

(i) $D_{i1} > \pi_1 D_1, \ldots, D_{iK} > \pi_K D_K$ for $0 < \pi_k < 1$, $k = 1, \ldots, K$;

(ii) $D_{i1} > \pi_1 D_1^{SC}, \ldots, D_{iK} > \pi_K D_K^{SC}$ for $0 < \pi_k < 1$, $k = 1, \ldots, K$;

(iii) $D_{i1} > h_{i1}, \ldots, D_{iK} > h_{iK}$ for $h_{ik} > 0$, $k = 1, \ldots, K$.

The values of π_k and h_{ik} are prespecified for $i = 1, \ldots, S$ and $k = 1, \ldots, K$. The authors recommend setting $h_{ik} = z_{1-\phi_{ik}} \sigma_k \sqrt{4/Np_i}$, where N is the total sample size, p_i is the proportion of all patients in the trial who are in the ith region, σ_k is the known standard deviation of the kth outcome, $z_{1-\phi_{ik}}$ is the $(1 - \phi_{ik})$th percentile of the standard normal distribution, and ϕ_{ik} can be thought of as a desired significance level if one were to test the effect of the treatment with respect to the kth endpoint using patients from only the ith region. Criteria (i) and (ii) compare the size of the ith region-specific treatment effect to the overall treatment effect, which is calculated with and without the ith region, respectively, and we note that the first criterion is similar to Method 1 proposed by the Japanese MHLW (MHLW, 2007). The authors suggest that the sample size for each region should be calculated such that at least one of the three criteria is met.

CHAPTER 3: BAYESIAN MODEL AVERAGING FOR MULTI-REGIONAL CLINICAL TRIALS WITH A CONTINUOUS ENDPOINT

3.1 Introduction

With the increasing globalization of medical drugs, pharmaceutical companies have turned to the frequent use of multi-regional clinical trials (MRCTs), or studies that include multiple geographic regions under the same study protocol. Although these studies introduce both logistical and statistical challenges, MRCTs provide the benefit of allowing drugs to more quickly enter the world market. Song *et al.* (2019) investigated all clinical trials registered on ClinicalTrials.gov by the top ten pharmaceutical companies between 1 January 2008 and 31 December 2017, and using the International Council for Harmonisation (ICH) E17 guideline to classify trials as either local or MRCT, they found that MRCTs made up 66.0% of phase II trials and 72.2% of phase III trials. While the total number of trials conducted by these ten companies has decreased over time, the proportion of MRCTs has greatly increased in all three phases. MRCTs gained further attention after the finalization of ICH E17 in November 2017, which provides nonbinding guidelines for the planning and design of MRCTs.

Multi-regional clinical trials generally have two objectives: (1) estimate the global effect of an active treatment, and (2) bridge trial results to individual regions having potentially different regulatory authorities. In this paper, we refer to the average treatment effect across regions as the global treatment effect, although it is sometimes referred to as the overall treatment effect in literature. Substantial evidence of a global effect indicates that the drug is effective in at least one region. However, region-specific effects can still differ due to differences in both intrinsic factors (e.g., genetic factors and pathological conditions) and extrinsic factors (e.g., environmental and cultural factors) between regions (ICH, 1998). Commonly used methods, such as fixed effects

models, often do not account for regional differences during the primary analysis and instead assume a homogeneous treatment effect across regions (Hung *et al.*, 2010).

We propose the use of Bayesian model averaging (BMA) which naturally provides estimates for the region-specific treatment effects, and we propose an approach for calculating the global treatment effect based on a weighted average of the region-specific treatment effects. Bayesian model averaging has previously been used to account for model uncertainty in which inferences from models with different sets of covariates are averaged (Hoeting *et al.*, 1999), and Psioda *et al.* (2021) extend the application of BMA to oncology basket trials by averaging models that categorize baskets into different homogeneous sets based on response rate. We extend the BMA approach to MRCTs, where each model considers a different possible grouping of how the regions are similar or different with regard to the effectiveness of an active treatment. For a given number of regions, the model space includes models pertaining to all possible classifications of regions into subsets that have common effects, with the two extremes being a model that assumes no regional differences by pooling all regions together (i.e., a fixed effects linear model that does not include a region-specific treatment effect) and a model that assumes all region-specific treatment effects differ (i.e., a fixed effects linear model with a separate region-specific treatment effect for each region). The BMA approach facilitates borrowing of information across regions while reasonably controlling the type I error rate, whereas other methods, such as Bayesian hierarchical models (BHMs), often incorporate information borrowing at the expense of greatly increasing type I error rates in regions with a small or null treatment effect.

Various models have been proposed for MRCTs. Fixed effects models are frequently used in practice, but Hung *et al.* (2010) suggest the use of random effects models where regional differences are treated as random variables. Wu *et al.* (2014) proposed a random effects model based on the assumption that the shift in variance across regions is random and that the mean response is fixed. Chiang and Hsiao (2019) point out that regional intrinsic and extrinsic factors are generally known and considered to be fixed, questioning the treatment of regional differences as random effects. Additionally, variability across regions is poorly estimated by random effects

models when the number of regions is small. Instead, Chiang and Hsiao assume that efficacy responses of regions are independently distributed with a common mean and heterogeneous variances, and they evaluate the efficacy response through interval estimation based on Howe, Cochran-Cox, and Satterthwaite approximations. Other methods have also been proposed that assume heterogeneous variances between the regions. For a continuous response, Lan and Pinheiro (2012) proposed a discrete random effects model (DREM) using discrete priors to account for between-region variability, and Lan *et al.* (2014) extended DREM to survival and binary responses. The DREM approach treats regional sample sizes as random by assuming that patients are randomly drawn from the full patient population of the selected regions, which may not be a realistic assumption in MRCTs due to minimum regulatory requirements.

In an MRCT, consistency of treatment effects is defined as a lack of clinically relevant differences between treatment effects in different regions (ICH, 2017). Evaluation of consistency provides an understanding of the extent to which global results of a trial can reasonably be bridged to individual regions. The Pharmaceuticals and Medical Devices Agency in the Ministry of Health, Labour and Welfare (MHLW) of Japan released a guideline for MRCTs (MHLW, 2007) that addressed the concept of consistency through two proposed methods, and many authors have extended these methods and developed regional sample size calculations (Kawai *et al.*, 2008; Ko *et al.*, 2010; Liu *et al.*, 2016; Quan *et al.*, 2010a,b, 2013; Tsong *et al.*, 2012; Uesaka, 2009). One common method to assess consistency is to test the interaction between treatment and region, however, interaction tests are known for having low power to detect heterogeneity. We propose three approaches that utilize posterior probabilities to quantify the strength of evidence in favor of consistency. Two approaches estimate the probability that region-specific treatment effects (pairwise and globally) differ by no more than a prespecified minimal clinically important regional difference, and the third approach compares region-specific effects to the global effect.

The rest of this paper is organized as follows. In Section 3.2, we discuss a motivating MRCT that used current statistical methods in the design and analysis stages, and we highlight limitations. In Section 3.3, we discuss the methodology of BMA and detail its application to MRCTs,

and we propose methods to assess treatment effect consistency. In Section 3.4, we define the global treatment effect and outline steps for hypothesis testing of both global and region-specific treatment effects. In Section 3.5, we present simulation studies to compare the proposed method to a fixed effects linear regression model and a BHM. We then close the paper with discussion in Section 3.6.

3.2 Motivating Example

To highlight the strengths of BMA in MRCTs, we discuss a phase III trial comparing two treatments for chronic obstructive pulmonary disease (COPD), and we design simulation studies to resemble the COPD trial in Section 3.5. The trial was designed to evaluate the efficacy and safety of fluticasone furoate/vilanterol inhalation powder (FF/VI) 100/25 mcg once daily compared to vilanterol inhalation powder (VI) 25 mcg once daily in subjects with COPD (Siler *et al.*, 2017). For patients in either treatment arm, investigators measured a patient's forced expiratory volume in one second (FEV1) twice on Treatment Day 1 (30 minutes pre-dose and immediately pre-dose) and averaged the two values to obtain a baseline measurement. On Treatment Day 84, investigators measured the patient's FEV1 twice more (both 23 and 24 hours after the previous morning's dosing) and averaged these values. The primary outcome of this study is the mean change from baseline in clinic visit trough FEV1 on Treatment Day 84.

The overall sample size was calculated using the hypothesized global treatment mean difference based on historical data from two phase III trials evaluating the efficacy and safety of FF/VI (GSK, 2014b). A two-sample t-test with an assumed common standard deviation of 230 mL and a two-sided 5% significance level was used. Each treatment arm would need 696 patients to detect a 40 mL difference between FF/VI 100/25 mcg and VI 25 mcg in trough FEV1 on Treatment Day 84 with 90% power. After the completion of the trial, the FF/VI arm and the VI arm had sample sizes of 759 and 749, respectively. Subjects were recruited and randomized from 211 centers in 11 countries (GSK, 2014a), and countries were grouped into five regions defined as US,

Asia Pacific, Eastern Europe, Western Europe, and Other. The primary analysis used region as a covariate.

The study results focus on the global treatment effect without the inclusion of a region-by-treatment interaction, and the model does not allow for the treatment effect to differ across regions. Without additional analyses, models such as that used are unable to provide estimates of region-specific treatment effects. To address the 2007 Japanese guideline of consistency, Japan-specific analyses were conducted using a new definition of region groupings that classified countries as either Japan or Not-Japan. Although these Japan-specific analyses were included in the clinical study report as required by the MHLW, investigators acknowledged that the sample size was not powered for a formal analysis in the subset of subjects recruited in Japan (GSK, 2014a). Such is generally true for all regional subgroup analyses when the sample size is powered based on the global treatment mean difference.

3.3 Model

3.3.1 BMA Applied to MRCTs With a Normally Distributed Outcome

Consider an MRCT comparing two treatments with S regions and a total sample size N. Let $\boldsymbol{Y} = (y_{ij})'$ be an $N \times 1$ vector where y_{ij} is the continuous response for the jth subject in the ith region, $i = 1, \ldots, S$, $j = 1, \ldots, n_i$, where n_i is the number of patients in the ith region and $\sum_{i=1}^{S} n_i = N$. Define $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_S)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S)'$ to be the $S \times 1$ vectors of region-specific intercepts and region-specific treatment effects, respectively, and let $\boldsymbol{\beta}$ be the $p \times 1$ vector of covariate effects. We let $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\gamma}', \boldsymbol{\beta}')'$, and we consider the case where $\boldsymbol{\theta}$ is unknown. The model is written as $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{W}$ is the $N \times (2S + p)$ design matrix with region indicators, region-by-treatment-group indicators, and optional covariates. We assume that $\boldsymbol{\epsilon} \sim N_N(\boldsymbol{0}, \tau^{-1}\boldsymbol{I}_N)$, where τ is the precision parameter and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix.

Let $\mathcal{M}_S$ denote the model space for S regions. Assume there are L models in $\mathcal{M}_S$ where the ℓth model is denoted by M_ℓ, $\ell = 1, \ldots, L$, and let D_ℓ be the number of distinct region-specific treatment effects for M_ℓ, $1 \leqslant D_\ell \leqslant S$. The number of models in $\mathcal{M}_S$ can be calculated as

$L = \sum_{i=1}^{S} \left\{ \frac{1}{i!} \sum_{\ell=0}^{i} (-1)^{\ell} \binom{i}{\ell} (i-\ell)^{S} \right\}$, where the bracketed term is the Stirling number of the second kind.

As an illustration, consider the case when $S = 3$. The possible treatment effect models are shown in Table 3.1, where $\gamma_{(\ell,d)}$ is the dth distinct treatment effect for model M_ℓ, $d = 1, \ldots, D_\ell$, $\ell = 1, \ldots, L$. Let $\Omega_{\ell,d}$ be the set of region labels corresponding to the regions that have the dth distinct treatment effect for model M_ℓ, and let $\boldsymbol{\gamma}_{(\ell)}$ be the $D_\ell \times 1$ vector of distinct treatment effects for model M_ℓ. For model M_1 in the example when $S = 3$, we assume that all three regions share a common treatment effect $\gamma_{(1,1)}$, and thus $D_1 = 1$, $\Omega_{1,1} = \{1, 2, 3\}$, and $\boldsymbol{\gamma}_{(1)} = \gamma_{(1,1)}$. For model M_5, we assume that each of three regions has a distinct treatment effect, and thus $D_5 = 3$, $\Omega_{5,1} = \{1\}$, $\Omega_{5,2} = \{2\}$, $\Omega_{5,3} = \{3\}$, and $\boldsymbol{\gamma}_{(5)} = (\gamma_{(5,1)}, \gamma_{(5,2)}, \gamma_{(5,3)})'$.

For the ℓth model, define the $(S + D_\ell + p) \times 1$ vector $\boldsymbol{\theta}_{(\ell)} = (\boldsymbol{\mu}', \boldsymbol{\gamma}_{(\ell)}', \boldsymbol{\beta}')'$, and let $\boldsymbol{W}_{(\ell)}$ be the corresponding design matrix. We write model M_ℓ as $\boldsymbol{Y} = \boldsymbol{W}_{(\ell)}\boldsymbol{\theta}_{(\ell)} + \boldsymbol{\epsilon}$. Conditional on model M_ℓ, we write the priors for $\boldsymbol{\theta}_{(\ell)}$ and τ as $p(\boldsymbol{\theta}_{(\ell)}, \tau | M_\ell) = p(\boldsymbol{\theta}_{(\ell)} | \tau, M_\ell) p(\tau | M_\ell)$, $\boldsymbol{\theta}_{(\ell)} | \tau, M_\ell \sim N_{S+D_\ell+p}(\boldsymbol{m}_{(\ell)}, \tau^{-1}\boldsymbol{\Sigma}_{(\ell)})$, $\tau | M_\ell \sim$ gamma $\left(\frac{\delta_0}{2}, \frac{\nu_0}{2}\right)$, where $\boldsymbol{m}_{(\ell)}$ is a $(D_\ell + S + p) \times 1$ vector and $\boldsymbol{\Sigma}_{(\ell)}$ is a $(D_\ell + S + p) \times (D_\ell + S + p)$ matrix. The hyperparameters $\left(\boldsymbol{m}_{(\ell)}, \boldsymbol{\Sigma}_{(\ell)}, \delta_0, \nu_0\right)$ are elicited. For the prior on $\tau | M_\ell$, we use the inverse-scale parameterization of the gamma distribution such that $E(\tau | M_\ell) = \frac{\delta_0}{\nu_0}$.

Let $S_p(b, \boldsymbol{m}, \boldsymbol{\Sigma})$ denote a p-dimensional t distribution with b degrees of freedom, location vector $\boldsymbol{m}$, and dispersion matrix $\boldsymbol{\Sigma}$, and let $\mathbf{D}$ be the observed data for all subjects across both treatment groups and all S regions. The marginal posterior distribution of $\boldsymbol{\theta}_{(\ell)}$, conditional on

Table 3.1: Possible treatment effect models for $S = 3$ regions.

—— Region ——			Distinct	
$i = 1$	$i = 2$	$i = 3$	Effects	Model
$\gamma_{(1,1)}$	$\gamma_{(1,1)}$	$\gamma_{(1,1)}$	1	M_1
$\gamma_{(2,1)}$	$\gamma_{(2,1)}$	$\gamma_{(2,2)}$	2	M_2
$\gamma_{(3,1)}$	$\gamma_{(3,2)}$	$\gamma_{(3,1)}$	2	M_3
$\gamma_{(4,1)}$	$\gamma_{(4,2)}$	$\gamma_{(4,2)}$	2	M_4
$\gamma_{(5,1)}$	$\gamma_{(5,2)}$	$\gamma_{(5,3)}$	3	M_5

$(M_\ell, \mathbf{D})$, can be expressed as

$$\boldsymbol{\theta}_{(\ell)}|M_\ell, \mathbf{D} \sim S_{D_\ell+S+p}\left(N+\delta_0, \tilde{\boldsymbol{\theta}}_{(\ell)}, \tilde{s}^2_{(\ell)}\left(\boldsymbol{W}'_{(\ell)}\boldsymbol{W}_{(\ell)} + \boldsymbol{\Sigma}^{-1}_{(\ell)}\right)^{-1}\right), \quad \ell = 1, \ldots, L, \tag{3.1}$$

where

$$\tilde{\boldsymbol{\theta}}_{(\ell)} = \boldsymbol{\Lambda}_{(\ell)}\boldsymbol{m}_{(\ell)} + \left(\boldsymbol{I}_{(D_\ell+S+p)} - \boldsymbol{\Lambda}_{(\ell)}\right)\hat{\boldsymbol{\theta}}_{(\ell)},$$

$$\boldsymbol{\Lambda}_{(\ell)} = \left(\boldsymbol{W}'_{(\ell)}\boldsymbol{W}_{(\ell)} + \boldsymbol{\Sigma}^{-1}_{(\ell)}\right)^{-1}\boldsymbol{\Sigma}^{-1}_{(\ell)},$$

$$\hat{\boldsymbol{\theta}}_{(\ell)} = \left(\boldsymbol{W}'_{(\ell)}\boldsymbol{W}_{(\ell)}\right)^{-1}\boldsymbol{W}'_{(\ell)}\boldsymbol{Y},$$

$$\tilde{s}^2_{(\ell)} = (N+\delta_0)^{-1}\left\{\boldsymbol{Y}'(\boldsymbol{I}_N - \boldsymbol{M}_W)\boldsymbol{Y} + (\hat{\boldsymbol{\theta}}_{(\ell)} - \boldsymbol{m}_{(\ell)})'(\boldsymbol{\Lambda}'_{(\ell)}\boldsymbol{W}'_{(\ell)}\boldsymbol{W}_{(\ell)})(\hat{\boldsymbol{\theta}}_{(\ell)} - \boldsymbol{m}_{(\ell)}) + \nu_0\right\},$$

$$\boldsymbol{M}_W = \boldsymbol{W}_{(\ell)}\left(\boldsymbol{W}'_{(\ell)}\boldsymbol{W}_{(\ell)}\right)^{-1}\boldsymbol{W}'_{(\ell)}.$$

For arbitrary x, we approximate $P(\gamma_i > x|M_\ell, \mathbf{D})$ via Monte Carlo integration.

3.3.2 Application of BMA

The marginal likelihood for the data conditional on model M_ℓ is written as

$$p(\mathbf{D}|M_\ell) = \left\{\Gamma\left(\frac{N+\delta_0}{2}\right)\Gamma\left(\frac{\delta_0}{2}\right)^{-1}(\pi\nu_0)^{-\frac{N}{2}}\left|\boldsymbol{I}_N + \boldsymbol{W}_{(\ell)}\boldsymbol{\Sigma}_{(\ell)}\boldsymbol{W}'_{(\ell)}\right|^{-\frac{1}{2}}\right\}$$

$$\times\left\{1 + \frac{1}{\nu_0}\left(\boldsymbol{Y} - \boldsymbol{W}_{(\ell)}\boldsymbol{m}_{(\ell)}\right)'\left(\boldsymbol{I}_N + \boldsymbol{W}_{(\ell)}\boldsymbol{\Sigma}_{(\ell)}\boldsymbol{W}'_{(\ell)}\right)^{-1}\left(\boldsymbol{Y} - \boldsymbol{W}_{(\ell)}\boldsymbol{m}_{(\ell)}\right)\right\}^{-\frac{N+\delta_0}{2}},$$

and we define $p(M_\ell)$ to be the prior model probability of M_ℓ, $\ell = 1, \ldots, L$. Using Bayes' theorem, the posterior model probability (PMP) for M_ℓ is

$$p(M_\ell|\mathbf{D}) = \frac{p(\mathbf{D}|M_\ell)p(M_\ell)}{\sum_{\ell'=1}^{L} p(\mathbf{D}|M_{\ell'})p(M_{\ell'})}. \tag{3.2}$$

Using the PMPs as weights, the model-averaged posterior $p(\boldsymbol{\theta}|\mathbf{D})$ is obtained as

$$p(\boldsymbol{\theta}|\mathbf{D}) = \sum_{\ell=1}^{L} p(\boldsymbol{\theta}_{(\ell)}|M_\ell, \mathbf{D}) p(M_\ell|\mathbf{D}). \tag{3.3}$$

3.3.3 Prior Elicitation

Based on simulation studies, we recommend setting the diagonals of $\boldsymbol{\Sigma}_{(\ell)}$ to Diag$\{(10 \times |\boldsymbol{\psi}_{(\ell)}|)^2\}$, where $\boldsymbol{\psi}_{(\ell)}$ is a vector of the best predictions for the corresponding parameters of interest. For diagonals corresponding to $\boldsymbol{\mu}$, $\boldsymbol{\psi}_{(\ell)}$ would be the best prediction for the control group means within each region. For diagonals corresponding to $\boldsymbol{\gamma}_{(\ell)}$, $\boldsymbol{\psi}_{(\ell)}$ would be the best prediction for the mean difference between the treatment group and control group.

For the prior model probability, we recommend $p(M_\ell) \propto D_\ell^{\alpha_0}$, where α_0 is a tuning parameter set by the investigator. The choice of $\alpha_0 = 0$ results in the uniform model prior in which all models in $\mathcal{M}_S$ are weighted equally *a priori*. Increasing values of α_0 place greater prior weight on models with a higher number of distinct treatment effects. We discuss sensitivity of analyses to the choices of $\boldsymbol{\psi}_{(\ell)}$ and α_0 in Section 3.5 with the use of simulation studies.

3.3.4 Assessing Consistency of Treatment Effects

The ICH E17 guideline (ICH, 2017) defines consistency of the treatment effect as a lack of clinically relevant differences between treatment effects in different regions of an MRCT. In this section, we introduce approaches for quantifying consistency of treatment effects overall (i.e., globally), between regions (i.e., pairwise), and between a region and all other regions (i.e., local).

3.3.4.1 Pairwise and Global Consistency

We refer to the ICH E17 definition as *pairwise consistency* for pairwise comparisons of regions and *global consistency* if comparing all regions. For each comparison, we propose an approach to assess the strength of evidence in favor of consistency at the ε level, where ε is the minimal clinically important regional difference (MCIRD). We assume that any two treatment

effects are consistent if their difference is less than ε. For any two regions i and j, we define the ε-level pairwise consistency probability as $P(|\gamma_i - \gamma_j| < \varepsilon|\mathbf{D}) = \sum_{\ell=1}^{L} P(|\gamma_i - \gamma_j| < \varepsilon|M_\ell, \mathbf{D})p(M_\ell|\mathbf{D})$. Similarly, we can define the ε-level pairwise inconsistency probability as $P(|\gamma_i - \gamma_j| > \varepsilon|\mathbf{D})$.

To assess global consistency at the ε level, we define an approach that utilizes PMPs and ε-level pairwise inconsistency probabilities. For some prespecified probability β^*, we evaluate whether $P(|\gamma_i - \gamma_j| > \varepsilon|\mathbf{D}) > \beta^*$ for all pairwise comparisons. If this holds true for some regions i and j, then we consider this to be sufficient evidence that the treatment effects for regions i and j differ by more than the MCIRD (i.e., evidence in support of pairwise inconsistency). We then consider all models that allow the ith and jth regions to differ. Let Θ be the set of labels for the models where at least one pair of distinct treatment effects meets the criterion for having an MCIRD. We note that evidence for pairwise inconsistency for one or more pairs of regions is also evidence for global inconsistency. Hence, all models not in Θ support the claim of global consistency, and thus, we calculate the probability for ε-level global consistency as

$$1 - \sum_{\ell \in \Theta} p(M_\ell|\mathbf{D}). \tag{3.4}$$

We note that this probability increases as the pairwise inconsistency threshold β^* increases, whereas small values of β^* allow for greater tolerance of inconsistency. A value of $\beta^* = 0.5$ provides balance between the tolerance of consistency and inconsistency. We also note that in the case of "exact" global consistency (i.e., $\varepsilon = 0$), the probability in (3.4) equals the PMP for the model that assumes all regions share one distinct treatment effect.

3.3.4.2 Local Consistency

Some regulatory authorities require sponsors to provide evidence that the global treatment effect can be bridged to a local region-specific treatment effect. In this paper, we refer to this definition of consistency as *local consistency*. The Japanese MHLW (MHLW, 2007) addressed the

assessment of local consistency by proposing two methods in a guideline for MRCTs. Method 1 requires that a sufficient number of patients from a given region should be enrolled such that the treatment effect for that region, denoted by γ_{reg}, is at least half of the global treatment effect, denoted by γ_{all}, with probability at least 0.8; i.e., $P(\gamma_{reg}/\gamma_{all} > \pi) \geqslant 1 - \beta$, where $\pi \geqslant 0.5$ and $1 - \beta \geqslant 0.8$. Method 2 asserts that consistency can be assumed if $\gamma_i > 0$ for all $i = 1, \ldots, S$. Quan *et al.* (2010a) describe the limitations of these guidelines and instead combine the two methods, proposing that each region-specific treatment effect should exceed some proportion of the global treatment effect; i.e., $\gamma_1 > \pi\gamma_{all}, \gamma_2 > \pi\gamma_{all}, \ldots, \gamma_S > \pi\gamma_{all}$. This guideline essentially extends Method 1 to all regions, and setting $\pi = 0$ results in Method 2 of the MHLW guideline. Quan *et al.* (2014) argue that the choice of $\pi \geqslant 0.5$ for each region is not practical in cases with five or more regions, and they recommend choosing the value of π based on the number of regions. One suggestion offered is to set π proportional to the reciprocal of the number of regions. Teng *et al.* (2017) propose guidelines for choosing the value of π and additional consistency assessment parameters where the values vary according to the number of planned regions. In the simulation studies in Section 3.5, we follow the guideline of Quan *et al.* (2014) by setting $\pi = 1/S$.

One criticism of the methods proposed by the MHLW and Quan *et al.* (2010a) is that the global treatment effect includes the region of interest to which it is being compared. A second drawback is that the ratio is difficult to interpret if either the region-specific treatment effect or global treatment effect is negative. We instead propose calculating $P(|\gamma_i - \gamma_{(-i)}| < \varepsilon|\mathbf{D})$, $i = 1, \ldots, S$, where ε is the MCIRD and $\gamma_{(-i)}$ is the global treatment effect, as defined in Section 3.4, calculated without the ith region. We refer to this as the *leave-one-out global treatment effect*. We note that the global treatment effect may not correspond to any region, particularly in scenarios with heterogeneous treatment effects. In these cases, comparing the region-specific treatment effects to the global treatment effect is not ideal, and the previously described metrics for global and pairwise consistency may provide more practical interpretations.

3.4 Hypothesis Testing

The proposed BMA method allows for simultaneous quantification of evidence of region-specific treatment benefit and global treatment benefit within the same analysis. First, we outline the hypotheses for each region-specific treatment effect, and then we define a statistic to estimate the global treatment effect.

For $i = 1, \ldots, S$, consider the hypotheses $H_0 : \gamma_i \leqslant \gamma_0$ versus $H_1 : \gamma_i > \gamma_0$ where γ_0 is a prespecified real value. For arbitrary x, we use (3.3) to calculate posterior probabilities as $P(\gamma_i > x|\mathbf{D}) = \sum_{\ell=1}^{L} P(\gamma_i > x|M_\ell, \mathbf{D})p(M_\ell|\mathbf{D})$, which can be used for inference for the ith region-specific treatment effect. For some prespecified threshold probability π_0, we have substantial evidence in favor of the alternative hypothesis if $P(\gamma_i > \gamma_0|\mathbf{D}) > \pi_0$. An example choice for π_0 could be $\pi_0 = 1 - \alpha$, where α is the one-sided significance level in a frequentist analysis.

Under model M_ℓ, we define the global treatment effect conditional on $(M_\ell, \mathbf{D})$ to be

$$\gamma_{G,\ell}|M_\ell, \mathbf{D} = \sum_{d=1}^{D_\ell} \omega_{(\ell,d)}\gamma_{(\ell,d)}, \tag{3.5}$$

where $\omega_{(\ell,d)}$ is a weight corresponding to $\gamma_{(\ell,d)}$, $d = 1, \ldots, D_\ell$, and $\sum_{d=1}^{D_\ell} \omega_{(\ell,d)} = 1$. We set $\omega_{(\ell,d)} = \frac{n_{(\ell,d)}}{N}$, where $n_{(\ell,d)} = \sum_{i\in\Omega_{\ell,d}} n_i$, $d = 1, \ldots, D_\ell$ (i.e., $n_{(\ell,d)}$ is the combined sample size of all regions that share the dth distinct treatment effect under M_ℓ). Using BMA, we then calculate the posterior global treatment effect as $p(\gamma_G|\mathbf{D}) = \sum_{\ell=1}^{L} p(\gamma_{G,\ell}|M_\ell, \mathbf{D})p(M_\ell|\mathbf{D})$.

Consider the hypotheses $H_0 : \gamma_G \leqslant \gamma_0$ versus $H_1 : \gamma_G > \gamma_0$. With the conditional posterior distribution of $\gamma_{(\ell,d)}|M_\ell, \mathbf{D}$ from (3.1) and the global treatment effect function defined in (3.5), we can obtain the posterior distribution of $\gamma_{G,\ell}|M_\ell, \mathbf{D}$ using Monte Carlo methods. For arbitrary x, we calculate posterior probabilities on the global treatment effect as $P(\gamma_G > x|\mathbf{D}) = \sum_{\ell=1}^{L} P(\gamma_{G,\ell} > x|M_\ell, \mathbf{D})p(M_\ell|\mathbf{D})$. If $P(\gamma_G > \gamma_0|\mathbf{D}) > \pi_0$, we conclude that we have substantial evidence in favor of the alternative hypothesis.

3.5 Simulation Studies

We compare the proposed BMA method to a fixed effects linear model (FELM) and a BHM via simulation studies. To calculate the global treatment effect for the FELM, we fit a model with a common treatment effect for all regions combined. For the BHM, we fit a model that allows for region-specific treatment effects with shrinkage towards an overall average effect. We compare the proposed global treatment effect from the BMA method to the global treatment effect from the FELM and to the overall fixed effect from the BHM. For the comparison of region-specific effects, we construct a second FELM with region-specific treatment effects. These treatment effects from the second FELM are compared to the region-specific effects from the BMA approach and the random region-specific treatment effects from the BHM. Full details on the model specification for the FELM and the BHM are discussed in Section A.1 of Appendix A.

When comparing both the global treatment effects and the region-specific treatment effects, we calculate the rejection rates of the null hypotheses discussed in Section 3.4 using simulation studies. For the region-specific treatment effects, we group the regions with a treatment effect (i.e., the alternative regions) and calculate the true positive rate (TPR), and we group the regions without a treatment effect (i.e., the null regions) and calculate the false positive rate (FPR).

The simulation studies are motivated by the COPD trial (NIH, 2014) detailed in Section 3.2. The FF/VI arm had a least squares mean change from baseline and standard deviation of 0.116 L and 0.204 L, respectively, with a sample size of 759, and the VI arm had a least squares mean and standard deviation of 0.082 L and 0.205 L, respectively, with a sample size of 749. In the simulation studies, we refer to the FF/VI arm as the treatment group and the VI arm as the control group. For each simulated dataset, we generated $N = 1508$ observations with 1:1 treatment allocation, and we set the treatment group mean to 0.116 L and the control group mean to 0.082 L (i.e., mean difference of 0.034 L). We assumed a common standard deviation of 0.205 L for the primary outcome of each group.

The BMA prior elicitation for the primary simulation studies reflects the realistic case when investigators rely on parameter predictions that differ from underlying "true" values. For elements of $\boldsymbol{m}_{(\ell)}$ and diagonals of $\boldsymbol{\Sigma}_{(\ell)}$ that correspond to region-specific treatment effects $\boldsymbol{\gamma}_{(\ell)}$, we specified values of 0 and $(10 \times 0.04)^2$ L, respectively, where 0.04 L is the assumed mean difference from the original study's sample size calculations (18% increase from the assumed true value of 0.034). We choose a value of 0.1 L for the region-specific control means, which is an increase from the true value by a similar factor (22% increase). Thus, we set values of 0.10 L and $(10 \times 0.10)^2$ L for the elements of $\boldsymbol{m}_{(\ell)}$ and the diagonals of $\boldsymbol{\Sigma}_{(\ell)}$ that correspond to region-specific intercepts $\boldsymbol{\mu}$. Assuming uniform prior model probabilities, we set $\alpha_0 = 0$, and we set $\gamma_0 = 0$ and $\pi_0 = 0.975$ for both the region-specific and global hypotheses in Section 3.4.

3.5.1 Simulation Results

3.5.1.1 Equal Regional Sample Sizes

In each simulation study, we set the number of regions to $S = 5$ and use the total sample size of $N = 1508$. We first consider the case of equal regional sample sizes, and we look at six scenarios that differ in the number of null regions. For one extreme, all five regions are specified as alternative regions, in which each has an underlying mean difference of 0.034 L. The opposite extreme considers five null regions, none of which have a true underlying mean difference. For each scenario, we generated $10{,}000$ datasets and calculated the global and regional null hypothesis rejection rates for each model. Both FELMs and the BHM were fit using the `rjags` package (Plummer *et al.*, 2022) in R. The rejection rates and relative MSE (FELM used as the reference method) are shown in Figure 3.1, and estimates of bias are reported in Section A.2 of Appendix A where positive bias indicates overestimation.

The BMA global test resulted in a similar global rejection rate as the FELM and the BHM in each scenario (see Panel A of Figure 3.1). Although the BHM had a higher TPR than the BMA approach and the FELM in most scenarios when testing region-specific treatment effects in alternative regions, it also had a much higher FPR when testing the effects in null regions for

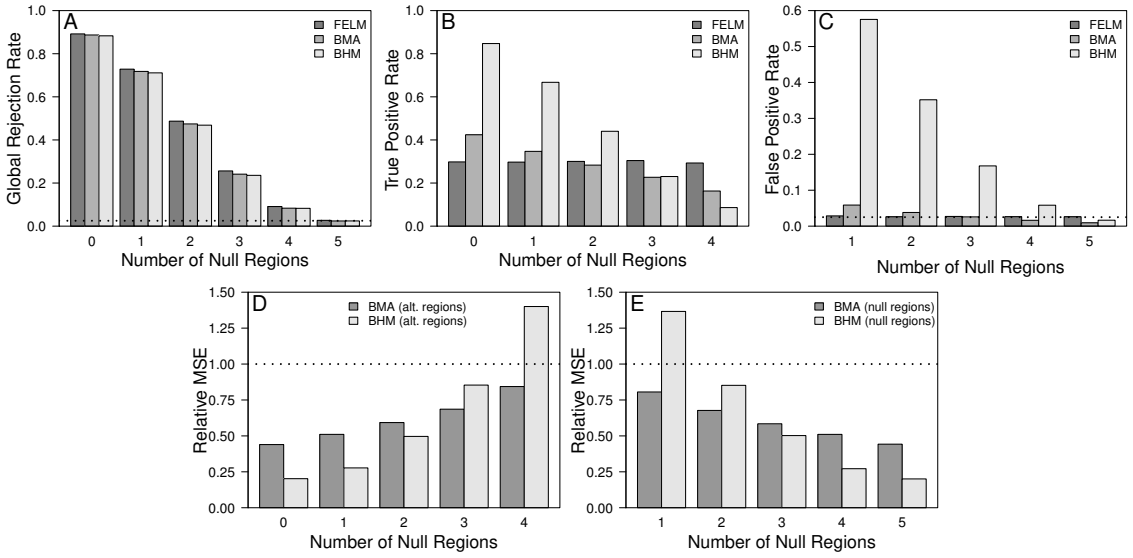

Figure 3.1: Global rejection rates (*Panel A*), true positive rates for alternative regions (*Panel B*), false positive rates for null regions (*Panel C*), relative MSE (FELM as reference) for alternative regions (*Panel D*), and relative MSE for null regions (*Panel E*) for simulations with equal regional sample sizes. Alternative regions have a treatment effect of 0.034 L.

all scenarios that included at least one alternative region (see Panels B and C of Figure 3.1). The BMA approach resulted in a greater TPR than the FELM for the region-specific tests for the 0- and 1-null-regions scenarios, with the tradeoff of having a slightly greater FPR in the 1- and 2-null-regions scenarios.

Compared to the FELM, the BMA approach provides appreciable reduction in MSE, resulting in a lower MSE than the FELM in each scenario. For both null and alternative regions, the MSE of the FELM estimates is the highest in four of the six scenarios (see Panels D and E of Figure 3.1). The BHM had the lowest MSE for alternative-region estimates in cases with more alternative regions than null regions and for null-region estimates in cases with a greater number of null regions. Additionally, the BMA approach resulted in the lowest MSE for alternative-region estimates in scenarios with a greater number of null regions than alternative and for null-region estimates in scenarios with a greater number of alternative regions than null. The FELM had minimal bias in all scenarios with a mix of null and alternative regions, and the BHM had the highest bias in these same scenarios. In the cases of only null regions and only alternative regions, the BMA approach resulted in the most-biased estimates. A common theme for both the BMA

approach and the BHM is an increase in bias for the alternative-region estimates as the number of alternative regions decreases. This increase in bias is more drastic with the BHM, where the increase in bias is often more than double the increase for the BMA approach between each scenario. These same trends are observed with the bias in null regions as the number of null regions decreases. The results of this simulation study highlight the strengths of the BMA approach over other models with its improved estimation quality of region-specific treatment effects while maintaining reasonable control over the FPR.

3.5.1.2 Behavior of Proposed Global Consistency Approach

We compute the ε-level global consistency probability defined in (3.4) by varying the value of ε between 0 and 0.07, and we consider two values of β^* to reflect the scenario with equal balance between consistency and inconsistency ($\beta^* = 0.5$) and a scenario with lower tolerance of inconsistency ($\beta^* = 0.8$). We also illustrate the proposed approach for total sample sizes of $N \in \{754, 1508, 7650\}$, where $N = 754$ is half the sample size of the COPD trial, and $N = 7650$ was chosen as a hypothetical case when each regional sample size is powered to achieve 90% power to detect a mean difference of 0.034 L with a standard deviation of 0.205 L and one-sided $\alpha = 0.025$. The results are shown in Figure 3.2. As the sample size increases, this approach has a greater ability to discern inconsistencies between region-specific treatment effects, which is seen in the increased separation of lines between the cases simulated under global consistency (0- and 5-null-regions scenarios) and all other cases. A higher value of β^* results in an increased probability of consistency across all scenarios, even in cases with a mix of alternative and null regions. While there is no correct choice for β^*, we recommend $\beta^* = 0.5$ if investigators believe that heterogeneous treatment effects across regions is plausible.

3.5.1.3 Measures of Consistency for First Simulation Study

In the context of COPD, one proposed minimal clinically important difference is a change of 100 mL in pre-dose FEV1 (Donohue, 2005), which corresponds to a treatment group mean

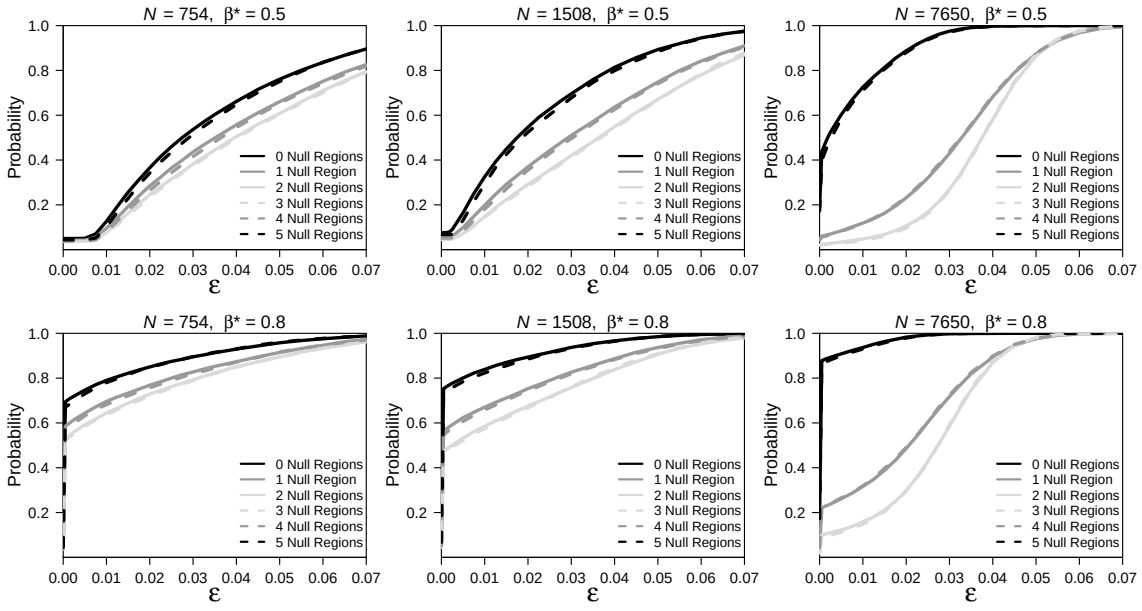

Figure 3.2: Average ε-level global consistency probabilities for varying values of ε, β^*, and N. We compare $\beta^* = 0.5$ (*top row*) vs. $\beta^* = 0.8$ (*bottom row*) across three different sample sizes: $N = 754$ (*left column*), $N = 1508$ (*middle column*), and $N = 7650$ (*right column*).

of 0.100 L and a mean difference of 0.018 L. We assume this mean difference to be the MCIRD in our simulation studies. With $N = 1508$ and $\beta^* = 0.5$, we estimate the average 0.018-level global consistency probability to be approximately 0.50 in the 0- and 5-null-regions scenarios, 0.33 in the 1- and 4-null-regions scenarios, and 0.26 for the 2- and 3-null-regions scenarios. For $\beta^* = 0.8$, the average probabilities for these three scenario groups increase to approximately 0.88, 0.73, and 0.65. We calculated all ε-level pairwise consistency probabilities for each dataset with $\varepsilon = 0.018$, and we report the average probabilities for each pairwise comparison of regions in Section A.2 of Appendix A.

To assess local consistency for the ith region under the six scenarios, $i = 1, \ldots, S$, we calculated both $P(\gamma_i/\gamma_G > \pi|\mathbf{D})$ and $P(|\gamma_i - \gamma_{(-i)}| < \varepsilon|\mathbf{D})$ for each dataset with $\pi = 0.2$ (i.e., the reciprocal of the number of regions) and $\varepsilon = 0.018$. The region-specific median probabilities of the $10{,}000$ datasets are presented in Table 3.2. For the 5-null-regions scenario, the ratio measure resulted in median probabilities that are approximately 0.23 lower than the results for the 0-null-regions scenario, whereas the medians for the leave-one-out absolute difference approach are

Table 3.2: Median probabilities of local consistency measures where $\pi = 0.20$ and $\varepsilon = 0.018$.

Number of Null Regions	$P(\gamma_i/\gamma_G > \pi \mid \mathbf{D})$ — Region — 1	2	3	4	5	$P(\lvert\gamma_i - \gamma_{(i)}\rvert < \varepsilon \mid \mathbf{D})$ — Region — 1	2	3	4	5
0	0.942	0.940	0.941	0.943	0.942	0.562	0.560	0.562	0.562	0.564
1	0.689	0.921	0.919	0.919	0.920	0.446	0.525	0.527	0.529	0.528
2	0.668	0.666	0.886	0.885	0.886	0.468	0.471	0.491	0.494	0.494
3	0.674	0.673	0.676	0.830	0.834	0.494	0.492	0.493	0.466	0.465
4	0.693	0.698	0.698	0.696	0.757	0.525	0.525	0.525	0.526	0.438
5	0.720	0.718	0.719	0.721	0.719	0.559	0.560	0.559	0.559	0.558

Null regions shaded; Treatment effect for alternative regions is 0.034

approximately the same for these two scenarios with underlying consistency across all regions. We also note that in the ideal case when every region has the same effect and $\gamma_G|D$ is perfectly estimated to be 0.034, there are values of $\gamma_i|D$ less than the proposed MCIRD that would satisfy the criterion $\gamma_i/\gamma_G > \pi$, even when using the MHLW's minimal requirement of $\pi = 0.5$. The MHLW method would still classify these region-specific treatment effects as consistent with the global effect, whereas the proposed method makes this classification only if the differences between the estimates is not clinically meaningful.

3.5.1.4 Unequal Regional Sample Sizes

For the second and third simulation studies, we varied the regional sample sizes based on the number of null regions in each scenario. In the second simulation study, we set the regional sample sizes of the null regions equal to half of the sample sizes of the alternative regions. We then set the alternative regional sample sizes to half the size of the null regions for the third simulation study. The rejection rates, relative MSE, and bias for the second and third simulation studies are presented in Section A.2 of Appendix A. Relative to the first simulation study, the BMA approach is more robust to changes in regional sample sizes than the BHM. We observe the same patterns in the rejection rates and relative MSE when using the BMA approach for the second and third simulation studies compared to the first study, whereas we observe greater changes in the relative MSE with the BHM across simulation studies. For the scenarios with a mix of null

and alternative regions, the average 0.018-level global consistency probabilities with $\beta^* = 0.5$ ranged as low as 0.22 in the second simulation study (3-null-regions case) and 0.25 in the third simulation study (2-null-regions case).

3.5.1.5 Heterogeneous Positive Treatment Effects

In the fourth simulation study, we compare the models in three scenarios where each region has a positive treatment effect with differing magnitudes and equal regional sample sizes. For the first scenario, we choose the underlying region-specific effects of the five regions by multiplying 0.034 L by 0.50, 0.75, 1.00, 1.25, and 1.50. For the second scenario, we choose two region-specific effects to be 0.034 while setting the effects of the other three regions as half of 0.034. For the third scenario, we choose one region-specific effect to be half of 0.034, two to be 0.034, and two to be 1.5 times 0.034. The results are shown in Figure 3.3. The BHM results in the highest TPR in each region with the FELM having the lowest TPR. The MSE for both the BMA approach and the BHM is much less than the MSE for the FELM, with the most substantial differences being observed for regions with treatment effects closest to the average effect. For $\beta^* = 0.5$, the average 0.018-level global consistency probabilities for the three scenarios are 0.37, 0.42, and 0.35, respectively.

3.5.2 Sensitivity Analysis

For the BMA prior elicitation in all simulation studies previously described, we chose values of hyperparameters that differ from the underlying "true" values. To further assess robustness of the proposed methods to the elicited priors, we repeated the first simulation study using several incorrect predictions for the assumed region-specific control mean and the mean difference used in the prior elicitation of $\boldsymbol{m}_{(\ell)}$ and $\boldsymbol{\Sigma}_{(\ell)}$. The results are included in Section A.3 of Appendix A. For both the region-specific control mean and the mean difference, we considered cases when the assumed values are both half and double the true values used to generate the data, and we also represent the ideal case when these values are perfectly predicted. The BMA rejection rate for the

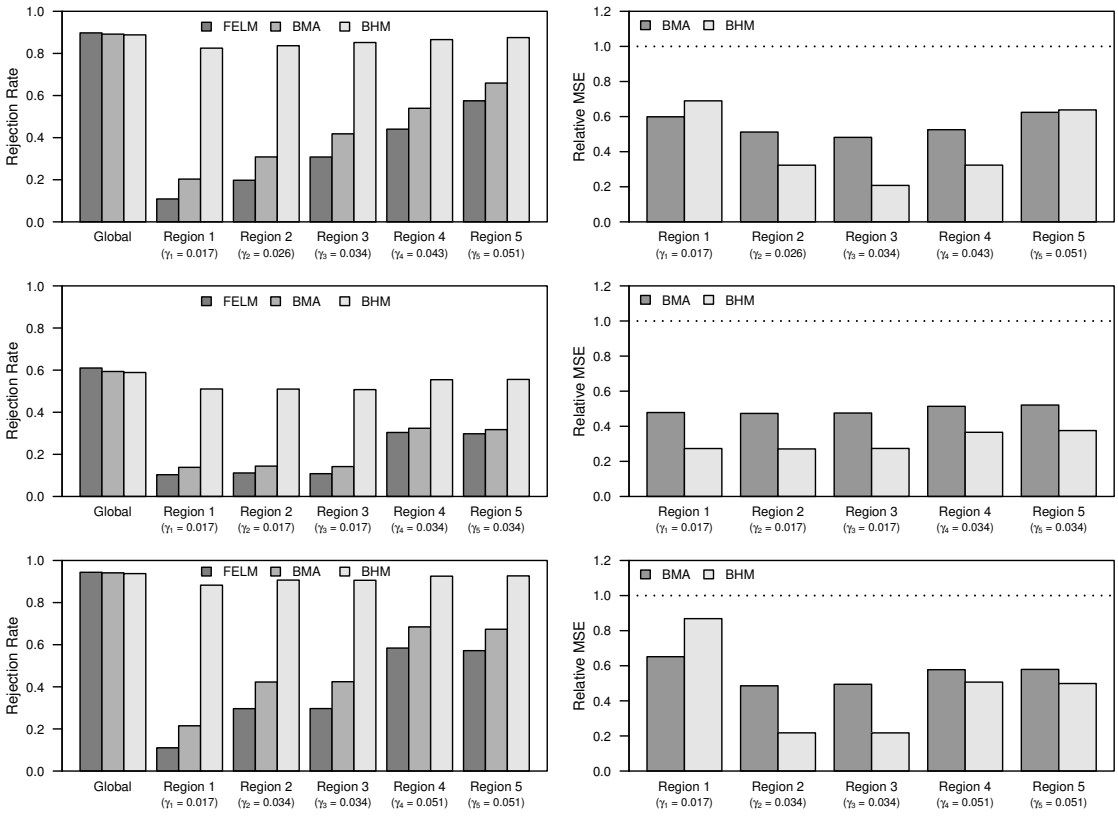

Figure 3.3: Global and regional rejection rates (*left column*) and relative MSE with the FELM as the reference method (*right column*) for simulations with equal regional sample size allocation and varying positive treatment effects across regions. We consider cases with five distinct effects ranging between 0.017 and 0.051 (*top row*), two distinct effects of 0.017 and 0.034 (*middle row*), and three distinct effects ranging between 0.017 and 0.051 (*bottom row*).

global test increased slightly when assuming the region-specific control mean to be half of the correct value, and the FPR under the 5-null-regions case increased by approximately 1 percentage point. When assuming the mean difference to be half of the correct value, the rejection rates decreased by 1-6 percentage points across the different scenarios. The rejection rates for all other incorrect predictions had only marginal changes from the primary simulation study, and the MSE remained substantially lower for the BMA approach compared to the FELM in each case. These results illustrate the robustness of using the BMA approach.

We performed a sensitivity analysis on the value of α_0 in the recommended prior model probability $p(M_\ell) \propto D_\ell^{\alpha_0}$, and the results are included in Section A.3 of Appendix A. To understand the influence of α_0 on the rejection rates, we repeated the first simulation study while considering

values of $\boldsymbol{\alpha_0} \in \{\pm 2, \pm 4, \pm 10\}$. Note that as α_0 increases, more weight is placed on models with a higher number of distinct region-specific treatment effects, thus favoring heterogeneous treatment effect models and reducing the amount of information borrowing. As expected, the global rejection rates across all six scenarios decrease as α_0 increases to account for less information borrowing, however, these decreases are marginal. Negative values of α_0, which favor models with fewer distinct treatment effects, allow for more information borrowing, resulting in increases in both TPR and FPR. On balance, we recommend $\alpha_0 \leqslant 0$, with $\alpha_0 = 0$ serving as a default choice.

3.6 Discussion

The proposed BMA design is motivated by the assumption that some regions in MRCTs may have similar treatment effects while recognizing that the treatment effects for other regions may differ. Compared to fixed effects models, the BMA approach results in substantially lower MSE when estimating region-specific effects while maintaining similar global rejection rates.

Both the BMA approach and the BHM incorporate information borrowing across regions, and an increase in TPR is associated with an increase in FPR. In most scenarios tested, this increase in FPR is minimal for the BMA approach compared to the BHM, which induces shrinkage towards the overall average mean and substantially inflates type I error rates for regions without a treatment effect (e.g., see Panel C of Figure 3.1). Hence, we recommend the BMA approach with $\alpha_0 = 0$ if it is plausible (in the opinion of stakeholders) that the investigational treatment may exhibit the desired level of efficacy for some regions while having no or minimal effect for others. In circumstances where such extreme levels of heterogeneity are not plausible, taking $\alpha_0 < 0$ will permit more information borrowing and greater efficiency gains in terms of MSE.

The BMA approach characterizes both a global treatment effect and region-specific effects, and its ability to probabilistically classify regions into sets with common effects via PMPs makes possible the use of innovative approaches for assessing global, local, and pairwise consistency of treatment effects. Unlike most existing methods for consistency assessment, these novel methods

ward against dichotomized thinking regarding whether consistency exists by focusing on quantifying the strength of evidence in favor of consistency rather than using an underpowered or ad hoc hypothesis testing approach.

We note that all methods produced the same degree of power for the overall treatment effect in each of the simulation scenarios. Hence, investigators can use traditional methods for sample size calculations and preserve the overall power with the use of BMA, and they can use simulations to further understand power for a given case. Nonetheless, after data are collected from an MRCT, the BMA approach should still be used to provide better estimation of region-specific treatment effects.

The marginal likelihood of the data and the marginal posterior distributions for the region-specific treatment effects have closed forms for the continuous-outcome situation described in this paper, allowing for quick and efficient computations. With $S = 5$ regions, $L = 52$ models in the model space, and $S + 1$ scenarios with a combined total of $60{,}000$ datasets per simulation study, all simulation studies were easily conducted using the Longleaf high-performance computing cluster at the University of North Carolina at Chapel Hill. For a single dataset, the BMA approach is easily executable on a computer using only a single core.

Conceptually, the proposed methodology for BMA can easily be extended to trials with non-Gaussian endpoints, however, these scenarios present a much harder computational problem. Without closed forms for the posterior distributions and marginal likelihoods, we must rely on approximation or Markov chain Monte Carlo methods when estimating PMPs and when sampling the region-specific treatment effects for each model. This is a topic for future research that would vastly increase the applicability of the BMA approach to MRCTs with different types of endpoints.

CHAPTER 4: BAYESIAN MODEL AVERAGING FOR MULTI-REGIONAL CLINICAL TRIALS WITH A TIME-TO-EVENT ENDPOINT

4.1 Introduction

Multi-regional clinical trials (MRCTs), or trials with multiple geographic regions included in the same study protocol, have become increasingly popular in the pharmaceutical industry due to their ability to allow sponsors to seek approval for investigational treatments from regulatory authorities for multiple geographic regions. Among all registered clinical trials conducted by the top ten pharmaceutical companies between the years 2008 and 2017 (the companies being ranked according to prescription sales in 2016), the proportion of trials that were MRCTs increased over time in each phase type, with the majority of phase II and phase III trials being MRCTs (Song *et al.*, 2019). The increased reliance on MRCTs prompted the publication of the International Council for Harmonisation (ICH) E17 guidelines for the planning and design of MRCTs in 2017 (ICH, 2017), which encourage the estimation of region-specific treatment effects in addition to the overall treatment effect.

Two major objectives of MRCTS are (1) to make inference on the overall treatment effect of an investigational treatment (referred to as the *global treatment effect* in the remainder of the paper), and (2) to estimate the region-specific treatment effects as part of subgroup analyses. With the increased popularity of MRCTs, greater research interest has been placed in assessing consistency of the treatment effect, or quantifying the degree of heterogeneity between region-specific treatment effects and the global treatment effect. The Japanese Ministry of Health, Labour and Welfare (MHLW) issued a guidance document describing two methods to determine the number of subjects needed in a trial to establish consistency between the Japan-specific treatment effect and the global treatment effect (MHLW, 2007), and adaptions of these guidelines have been ex-

tended for time-to-event (TTE) outcomes (Hayashi and Itoh, 2017; Huang *et al.*, 2012; Ko, 2020; Quan *et al.*, 2010b).

Data from MRCTs with TTE endpoints are commonly analyzed using the Cox proportional hazards model (Cox, 1972). The proportional hazards model can be formulated as a fixed effects model or a random effects model, in which the random effects are often referred to as "frailties" in survival analysis. When used for MRCTs, fixed effects models typically estimate only a global treatment effect in the primary analysis, ignoring regional differences that may result from intrinsic or extrinsic factors that vary across regions. If region-specific treatment effects are estimated, they are often done so as part of exploratory analyses and with the use of under-powered region-by-treatment interactions. Random effects models can estimate both a global treatment effect and region-specific effects if regions are considered to be a random sample from a larger population, however, this assumption may not be realistic depending on how geographical regions are defined in the study protocol. Additionally, if interest lies in assessing between-region heterogeneity, the region-level variation may be poorly estimated if the number of regions is small (Gelman and Hill, 2006).

The ICH E17 guidelines suggest the use of methods that allow for information borrowing across regions when regional sample sizes are small. For MRCTs with a TTE endpoint, we propose a method that naturally estimates region-specific treatment effects in addition to a global treatment effect, and the estimation quality of these effects is improved using information borrowing across regions. Specifically, we consider all possible partitions of the regions into sets where regions within a set share a common treatment effect. For each partition, we fit a model and estimate the distinct treatment effects for each set, and we then average posterior summaries from all of these models using Bayesian model averaging (BMA) (Hoeting *et al.*, 1999). Unlike traditional models that calculate only the global treatment effect, this approach accounts for regional heterogeneity in the treatment effects while also mitigating the possibility of obtaining poor estimates for the region-specific treatment effects.

While BMA has often been used for variable subset selection (Hoeting *et al.*, 1999), the idea of partitioning subgroups in clinical trials and then applying BMA to the corresponding models was first proposed by Psioda *et al.* (2021) as a method for estimating basket-specific response rates in oncology basket trials. Bean *et al.* (2021) extended the application to MRCTs with continuous endpoints while allowing for the inclusion of covariates. In both cases, estimates of region-specific treatment effects for each model can be directly sampled from closed-form posterior distributions. We further extend this methodology to TTE endpoints, and we develop a computationally efficient algorithm to obtain accurate approximations of posterior estimates using Laplace's method. Compared to standard statistical methods such as fixed effects models, this novel approach incorporates recommended guidelines from ICH E17—namely the estimation of both region-specific and global treatment effects in the primary analysis and the use of information borrowing methods—to greatly improve estimation quality of region-specific effects. Random effects models that use the hierarchical mean parameter as the global effect typically achieve low MSE when estimating region-specific effects at the cost of drastically decreasing the global rejection rate, whereas the BMA approach results in similar global rejection rates as fixed effects models.

The rest of this paper is organized as follows. Section 4.2 introduces a high-profile cardiovascular outcomes MRCT to highlight the shortcomings of commonly used statistical methods. In Section 4.3, we discuss the methodology of the proposed BMA approach, and we compare it to current methods using simulation studies in Section 4.4 to show how this approach results in similar global rejection rates while improving estimation quality of region-specific treatment effects. In Section 4.5, we use the proposed methodology to conduct a post hoc analysis of data from the MRCT discussed in Section 4.2, and we close with discussion in Section 4.6.

4.2 Motivating Example

The shortcomings of using standard statistical methods to analyze data from MRCTs are illustrated in the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome

Results (LEADER) trial (Marso *et al.*, 2016). In compliance with recently published guidelines from the U.S. Food and Drug Administration (FDA) for evaluating the cardiovascular safety of anti-diabetic medications (FDA, 2008), the study was designed with the primary objective of assessing the long-term effects of the treatment liraglutide versus placebo on the incidence of major adverse cardiovascular events (MACE), where the primary composite outcome in the TTE analysis is defined as time from randomization to first occurrence of a MACE. A hierarchical testing strategy was implemented to first test the primary hypothesis that liraglutide is non-inferior to placebo with regard to the primary outcome, and then superiority was subsequently tested. According to the FDA guidelines (FDA, 2008), non-inferiority is established if the upper boundary of the 95% confidence interval for the hazard ratio comparing to placebo is less than 1.30, and superiority is established if this upper boundary is below 1.00.

A total of 9340 patients were enrolled in the trial across 32 countries, and patients were assigned to either liraglutide or placebo. Significantly fewer patients in the liraglutide arm experienced the primary outcome compared to patients in the placebo arm (hazard ratio of 0.87; 95% confidence interval of 0.78–0.97), supporting the hypothesis that liraglutide results in a lower MACE incidence rate than placebo; however, the findings were not consistent when data were analyzed by geographical region as part of exploratory subgroup analyses. Of the four geographical regions analyzed, the North America region had an estimated hazard ratio favoring placebo (estimate = 1.01) while the three other regions (Asia, Europe, Rest of the World) had point estimates favoring liraglutide (i.e., less than 1.00).

All members of the Endocrinologic and Metabolic Drugs Advisory Committee agreed that the trial results supported the conclusion that liraglutide does not increase cardiovascular risk to patients with type 2 diabetes, but the regional subgroup analysis for North America raised concerns regarding liraglutide's ability to decrease cardiovascular risk in patients with type 2 diabetes for that group (Nielsen *et al.*, 2021). In June 2017, the advisory committee voted 17-2 in support of the claim that liraglutide reduces the risk for myocardial infarction, stroke, and cardiovascular death in adults with type 2 diabetes mellitus, and the FDA followed these recommen-

dations to approve a new indication for liraglutide. Additional post hoc analyses later provided support for this decision by suggesting that discrepancies in the estimated region-specific hazard ratios may be due to decreased drug exposure or chance rather than differences in intrinsic and extrinsic factors (Nielsen *et al.*, 2021).

We note that the LEADER trial was completed with published results in 2016 prior to the publication of ICH E17, and both the trial design and statistical analyses methods are standard for cardiovascular outcome MRCTs. Yusuf and Wittes (2016) detail additional examples of MRCTs, including four cardiovascular outcomes trials, where results vary across regions. Considering trial-specific explanations for these differences, they conclude that most variations in regional results are likely due to chance. The challenge posed by random extremes in region-specific treatment effect estimates can largely be overcome in MRCTs using information borrowing methods, which in turn can provide evidence in support of bridging a beneficial global effect of an investigational treatment to each region.

4.3 Model

4.3.1 Piecewise Constant Hazard Model for Time-to-Event Outcomes

Consider an MRCT comparing two treatments with S regions and a total sample size N, and let n_i be the number of patients in the ith region where $\sum_{i=1}^{S} n_i = N$. Let T_{ij} be the true survival time and C_{ij} the potential censoring time for the jth subject in the ith region, and we observe $y_{ij} = \min(T_{ij}, C_{ij})$ and $\nu_{ij} = 1(T_{ij} < C_{ij})$, where $1(\cdot)$ is the indicator function. Define the $N \times 1$ vectors $\boldsymbol{Y} = (y_{ij})'$ and $\boldsymbol{\nu} = (\nu_{ij})'$, and let $\boldsymbol{W}$ be the $N \times (S + p)$ design matrix where the first S columns correspond to region-specific treatment indicators and the last p columns are optional baseline covariates. We denote the observed data by $\mathbf{D} = \{\boldsymbol{Y}, \boldsymbol{\nu}, \boldsymbol{W}\}$.

We construct a finite partition of the time axis, $m_0 < m_1 < m_2 < \ldots < m_K$, with $m_0 \equiv 0$ and $m_K > \max(y_{ij})$, $i = 1, \ldots, S$, $j = 1, \ldots, n_i$, creating K intervals $(0, m_1], (m_1, m_2], \ldots, (m_{K-1}, m_K]$. In the kth interval, we assume a separate constant baseline hazard $h_0(y_{ij}) = \lambda_{ik}$ for each region where $y_{ij} \in (m_{k-1}, m_k]$, and we let $\boldsymbol{\lambda} = (\lambda_{ik})'$.

For the parameters of interest, we define an $S \times 1$ vector of region-specific treatment effects as $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S)'$ where γ_i is the treatment effect (i.e., log of the hazard ratio) for the ith region. We let $\boldsymbol{\beta}$ be the $p \times 1$ vector of covariate effects, $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$, and $\boldsymbol{\xi} = \{\boldsymbol{\lambda}, \boldsymbol{\theta}\}$. Using the formulation of Ibrahim *et al.* (2001), we write the likelihood for $\boldsymbol{\xi}$ as

$$\mathcal{L}(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{S}\prod_{j=1}^{n_i}\prod_{k=1}^{K}\Bigg[\left\{\lambda_{ik}\exp\left(\boldsymbol{w}_{ij}'\boldsymbol{\theta}\right)\right\}^{\delta_{ijk}\nu_{ij}} \times \exp\left(-\delta_{ijk}\left\{\lambda_{ik}(y_{ij} - m_{k-1}) + \sum_{g=1}^{k-1}\lambda_{ig}(m_g - m_{g-1})\right\}\exp\left(\boldsymbol{w}_{ij}'\boldsymbol{\theta}\right)\right)\Bigg],$$

where $\boldsymbol{w}_{ij}$ is the row of $\boldsymbol{W}$ corresponding to the jth subject in the ith region, and $\delta_{ijk} = 1$ if the subject had an event or was censored in the kth interval and 0 otherwise.

4.3.2 Definition of the Model Space and Classification of Region-Specific Treatment Effects

We define a model space that considers all partitions of regions, and we note the similarities of this approach to other applications of product partition models (PPMs) (Hartigan, 1990; Barry and Hartigan, 1992) in clinical trials. While PPMs are commonly used to identify subpopulations by clustering patients into covariate-dependent partitions (Muller *et al.*, 2011; Xu *et al.*, 2019), they can also be used to combine levels of categorical covariates in subgroup analyses. Sivaganesan *et al.* (2011) define a class of models for each covariate by considering partitions of subgroups based on treatment effects, and they propose a model selection approach to identify the presence of treatment-by-subgroup interactions. Similarly, Psioda *et al.* (2021) define a model space based on all possible partitions of baskets in a basket trial, however, they propose the use of BMA instead of model selection. Bean *et al.* (2021) extend this approach to linear models for MRCTs.

Following the model formulation of Bean *et al.* (2021), we assume that the model space $\mathcal{M}_S$ contains L models where each model corresponds to a unique partition of regions into sets such that all regions within the same set share a common treatment effect. The number of models is a function of the number of regions, calculated as $L = \sum_{i=1}^{S}\left\{\frac{1}{i!}\sum_{\ell=0}^{i}(-1)^{\ell}\binom{i}{\ell}(i-\ell)^S\right\}$. We

denote the ℓth model by M_ℓ and the prior model probability for M_ℓ by $p(M_\ell)$, $\ell = 1, \ldots, L$. Additionally, let D_ℓ denote the number of distinct region-specific treatment effects for the ℓth model, where $1 \leqslant D_\ell \leqslant S$. We let $\gamma_{(\ell,d)}$ denote the dth distinct treatment effect for model M_ℓ, $d = 1, \ldots, D_\ell$, $\ell = 1, \ldots, L$, and we define $\Omega_{\ell,d}$ to be the set of region labels corresponding to the regions that have the distinct effect $\gamma_{(\ell,d)}$.

Let $\boldsymbol{W}_\ell$ be the $N \times (D_\ell + p)$ matrix where the first D_ℓ columns correspond to region-specific treatment indicators under M_ℓ, and let $\boldsymbol{w}_{\ell ij}$ be the row of $\boldsymbol{W}_\ell$ corresponding to the jth patient in the ith region. We define $\boldsymbol{\gamma}_\ell$ to be the $D_\ell \times 1$ vector of distinct treatment effects for model M_ℓ, and let $\boldsymbol{\theta}_\ell = (\boldsymbol{\gamma}_\ell', \boldsymbol{\beta}')'$ and $\boldsymbol{\xi}_\ell = \{\boldsymbol{\lambda}, \boldsymbol{\theta}_\ell\}$. We may now write the likelihood for $\boldsymbol{\xi}_\ell$ conditional on M_ℓ being the true model as

$$\mathcal{L}(\boldsymbol{\xi}_\ell|\mathbf{D}, M_\ell) = \prod_{d=1}^{D_\ell} \prod_{i \in \Omega_{\ell,d}} \prod_{j=1}^{n_i} \prod_{k=1}^{K} \Bigg[\left\{ \lambda_{ik} \exp\left(\boldsymbol{w}_{\ell ij}' \boldsymbol{\theta}_\ell\right) \right\}^{\delta_{ijk}\nu_{ij}} \times \exp\left(-\delta_{ijk} \left\{ \lambda_{ik}(y_{ij} - m_{k-1}) + \sum_{g=1}^{k-1} \lambda_{ig}(m_g - m_{g-1}) \right\} \exp\left(\boldsymbol{w}_{\ell ij}' \boldsymbol{\theta}_\ell\right) \right) \Bigg].$$

4.3.3 Prior Formulation and Posterior Distributions

For model M_ℓ, we write the joint prior distribution for $\boldsymbol{\xi}_\ell$ as

$$p(\boldsymbol{\xi}_\ell|M_\ell) = p(\boldsymbol{\theta}_\ell|M_\ell) \times \left\{ \prod_{i=1}^{S} \prod_{k=1}^{K} p(\lambda_{ik}|M_\ell) \right\},$$

where $\boldsymbol{\theta}_\ell|M_\ell \sim N_{(D_\ell+p)}(\boldsymbol{\mu}_{0\ell}, \boldsymbol{\Sigma}_{0\ell})$ and $\lambda_{ik}|M_\ell \sim \text{Gamma}(\eta_{ik}, \phi_{ik})$, with hyperparameters $(\boldsymbol{\mu}_{0\ell}, \boldsymbol{\Sigma}_{0\ell}, \eta_{ik}, \phi_{ik})$, $i = 1, \ldots, S$, $k = 1, \ldots, K$. The full conditional distribution for λ_{ik} is $\lambda_{ik}|\boldsymbol{\theta}_\ell, \mathbf{D}, M_\ell \sim \text{Gamma}(\tilde{\eta}_{ik}, \tilde{\phi}_{ik})$, where

$$\tilde{\eta}_{ik} = \eta_{ik} + \sum_{j=1}^{n_i} \left(\delta_{ijk}\nu_{ij}\right),$$

$$\tilde{\phi}_{ik} = \phi_{ik} + \sum_{j=1}^{n_i} \left[\left\{ \delta_{ijk}(y_{ij} - m_{k-1}) + \sum_{g=k+1}^{K} \delta_{ijg}(m_k - m_{k-1}) \right\} \times \exp(\boldsymbol{w}_{\ell ij}' \boldsymbol{\theta}_\ell) \right].$$

After integrating out the baseline hazards, the marginal posterior distribution for $\boldsymbol{\theta}_\ell$ is

$$p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell) \propto \left\{ \prod_{d=1}^{D_\ell} \prod_{i \in \Omega_{\ell,d}} \prod_{k=1}^{K} \frac{\Gamma\left(\tilde{\eta}_{ik}\right)}{\tilde{\phi}_{ik}^{\tilde{\eta}_{ik}}} \exp\left(\sum_{j=1}^{n_i} \delta_{ijk} \nu_{ij} \boldsymbol{w}'_{\ell ij} \boldsymbol{\theta}_\ell \right) \right\} \times p(\boldsymbol{\theta}_\ell|M_\ell).$$

We recommend eliciting $\boldsymbol{\mu}_{0\ell}$ to be the vector of best predictions for the corresponding treatment effects or covariate effects under model M_ℓ and setting $\boldsymbol{\Sigma}_{0\ell}$ to be the diagonal matrix Diag$\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$. In the context of a non-inferiority trial, the elements of $\boldsymbol{\mu}_{0\ell}$ that correspond to region-specific treatment effects could be set as the non-inferiority bound on the log hazard ratio scale.

4.3.4 Efficient Sampling from Posterior Distributions Using a Laplace Approximation

In the case of normal endpoints, the marginal posterior distributions of all regression parameters have closed forms with recognizable distributions, allowing for direct sampling (Bean *et al.*, 2021). Additionally, the marginal likelihood of the data has a closed form, and the posterior model probabilities (PMPs) can be easily computed. In the case with TTE endpoints, however, the marginal distribution of the regression parameters is not a recognizable distribution, and the marginal likelihood of the data does not have a closed form. We implement Laplace's method (Tierney and Kadane, 1986) to derive accurate approximations of these posterior distributions and the marginal likelihood.

Laplace approximations have been applied to a wide class of survival models. Martino *et al.* (2011) apply integrated nested Laplace approximations (INLA) (Rue *et al.*, 2009) to parametric and semi-parametric Cox models formulated as latent Gaussian models, and Niekerk *et al.* (2021) extend INLA to joint models for TTE and longitudinal data. As an alternative to INLA, Gressani and Lambert (2018) approximate posterior distributions for latent variables in semi-parametric promotion time cure models by combining Laplace's method with penalized B-splines. Laplace approximations have also been used to integrate over random effects in mixed effects Cox models (Pankratz *et al.*, 2005) and joint models (Rizopoulos *et al.*, 2009).

Let $h_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell) = \log\{p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell)\}$. We approximate $p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell)$ by the $N_{(D_\ell+p)}(\hat{\boldsymbol{\theta}}_\ell, \hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_\ell})$ distribution, where $\hat{\boldsymbol{\theta}}_\ell$ is the posterior mode of $p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell)$ and $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_\ell}$ is the negative inverse Hessian of $h_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell)$ evaluated at $\hat{\boldsymbol{\theta}}_\ell$ (see Section B.1 of Appendix B for details of $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_\ell}$). The marginal likelihood of the data conditional on M_ℓ is approximated as

$$p(\mathbf{D}|M_\ell) \approx (2\pi)^{\frac{D_\ell+p}{2}} \mid \hat{\boldsymbol{\Psi}}_\ell \mid^{\frac{1}{2}} p(\mathbf{D}|\hat{\boldsymbol{\theta}}_\ell, M_\ell) p(\hat{\boldsymbol{\theta}}_\ell|M_\ell),$$

where $p(\mathbf{D}|\boldsymbol{\theta}_\ell, M_\ell) = \int \mathcal{L}(\boldsymbol{\xi}_\ell|\mathbf{D}, M_\ell) p(\boldsymbol{\lambda}|M_\ell) d\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\Psi}}_\ell$ is the negative inverse Hessian of $\log\{p(\mathbf{D}|\boldsymbol{\theta}_\ell, M_\ell) p(\boldsymbol{\theta}_\ell|M_\ell)\}$ evaluated at $\hat{\boldsymbol{\theta}}_\ell$ (Kass and Raftery, 1995). We note that $p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell)$ is proportional to $p(\mathbf{D}|\boldsymbol{\theta}_\ell, M_\ell) p(\boldsymbol{\theta}_\ell|M_\ell)$ and hence $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_\ell} = \hat{\boldsymbol{\Psi}}_\ell$. By first integrating out the baseline hazards from the joint posterior distribution, we can make posterior inference on the region-specific treatment effects without the need to sample posterior baseline hazards. Thus, we improve computational efficiency and obtain more accurate approximations of $p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell)$ and $p(\mathbf{D}|M_\ell)$ when working with the marginal distribution of $\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell$ rather than the full conditional distribution of $\boldsymbol{\theta}_\ell|\boldsymbol{\lambda}, \mathbf{D}, M_\ell$, $\ell = 1, \ldots, L$.

4.3.5 Inference via Bayesian Model Averaging

The PMP for M_ℓ is calculated as

$$p(M_\ell|\mathbf{D}) = \frac{p(\mathbf{D}|M_\ell) p(M_\ell)}{\sum_{\ell'=1}^{L} p(\mathbf{D}|M_{\ell'}) p(M_{\ell'})}, \tag{4.1}$$

where $p(M_\ell)$ denotes the prior probability for M_ℓ. We recommend setting $p(M_\ell) \propto e^{D_\ell \times \alpha_0}$, where positive values of the tuning parameter α_0 place greater prior probability on models with more distinct region-specific treatment effects. This recommendation is the default prior used in the `bmabasket` package in R (Psioda and Alt, 2022), but applied to MRCTs instead of basket trials.

Using BMA, we obtain the averaged posterior distributions of the region-specific treatment effects as

$$p(\gamma_i|\mathbf{D}) = \sum_{\ell=1}^{L} \left\{ \sum_{d=1}^{D_\ell} 1(i \in \Omega_{\ell,d}) p(\gamma_{(\ell,d)}|\mathbf{D}, M_\ell) \right\} p(M_\ell|\mathbf{D}), \quad i = 1, \ldots, S.$$

We define the global treatment effect under model M_ℓ to be $\gamma_G|M_\ell = \sum_{d=1}^{D_\ell} \frac{n_{(\ell,d)}}{N} \gamma_{(\ell,d)}$, where $n_{(\ell,d)}$ is the combined sample size of all regions that share the dth distinct treatment effect. We then compute the posterior global treatment effect as $p(\gamma_G|\mathbf{D}) = \sum_{\ell=1}^{L} p(\gamma_G|\mathbf{D}, M_\ell) p(M_\ell|\mathbf{D})$. If we allow γ to denote either the global treatment effect or a region-specific treatment effect, we can use BMA to test the hypothesis $H_0 : \gamma \geqslant \gamma_0$ versus $H_1 : \gamma < \gamma_0$ for some prespecified value γ_0 by calculating $P(\gamma < \gamma_0|\mathbf{D}) = \sum_{\ell=1}^{L} P(\gamma < \gamma_0|\mathbf{D}, M_\ell) p(M_\ell|\mathbf{D})$.

4.3.6 Measures of Consistency of the Treatment Effect

Consistency of the treatment effect across regions is defined differently throughout the literature with a common definition focusing on the comparison of region-specific effects to the global effect. For the continuous endpoint case, Bean *et al.* (2021) address the differences between definitions by proposing three measures to assess consistency: (1) the comparison of any two region-specific treatment effects (pairwise consistency), (2) the comparison of region-specific treatment effects across all regions (global consistency), and (3) the comparison of the treatment effect for a given region to the global treatment effect calculated without that region (local consistency). Here, we extend these three measures to TTE endpoints.

Both the pairwise and global consistency measures were designed to assess consistency as it is defined by ICH E17, which is a lack of clinically relevant differences between treatment effects in different regions of an MRCT (ICH, 2017). We note that two region-specific hazard ratios (HRs) with perfect consistency have a ratio of one, and we define ε as the smallest acceptable ratio between two region-specific HRs while still considering them to be consistent with one another; i.e., two region-specific HRs are consistent if $\varepsilon < e^{\gamma_i - \gamma_j} < \varepsilon^{-1}$, $i \neq j$, $\varepsilon \in (0, 1)$. The

value of ε, which we refer to as the minimal clinically important regional difference (MCIRD), should be defined prior to assessing any of the proposed measures of consistency. For example, the MCIRD can be set as the inverse of the non-inferiority margin.

We define the ε-level pairwise consistency probability as $P(\varepsilon < e^{\gamma_i - \gamma_j} < \varepsilon^{-1}|\mathbf{D}) = \sum_{\ell=1}^{L} P(\varepsilon < e^{\gamma_i - \gamma_j} < \varepsilon^{-1}|\mathbf{D}, M_\ell)p(M_\ell|\mathbf{D})$, and the ε-level pairwise inconsistency probability as $P(|\gamma_i - \gamma_j| > -\log(\varepsilon)|\mathbf{D}) = 1 - P(\varepsilon < e^{\gamma_i - \gamma_j} < \varepsilon^{-1}|\mathbf{D})$. The ε-level global consistency probability quantifies the strength of evidence that no clinically meaningful difference exists between any of the region-specific HRs. Hence, the stronger the evidence in favor of pairwise inconsistency for any two regions, the weaker the evidence in favor of global consistency.

To calculate the global consistency probability, we first calculate all ε-level pairwise inconsistency probabilities and compare them to some prespecified probability threshold β^*. If the pairwise inconsistency probability is greater than this threshold, we assume that the HRs from the two regions are inconsistent (i.e., the ratio of the smaller to larger HR is likely less than the MCIRD), and we consider all models in $\mathcal{M}_S$ that allow these two regions to differ in the calculation of the global inconsistency probability as described below. Let Θ be the set of labels for the models where at least one pair of distinct HRs meets the criterion for being inconsistent. These models provide evidence in favor of ε-level pairwise inconsistency for at least one pair of regions, and the sum of their PMPs measures the strength of evidence in favor of global inconsistency. Thus, the ε-level global consistency probability is calculated as $1 - \sum_{\ell \in \Theta} p(M_\ell|\mathbf{D})$. High values of β^* are more conservative with respect to classifying two regions as inconsistent, which ultimately results in an increase in the ε-level global consistency probability; conversely, low values of β^* correspond to a decrease in the global consistency probability. We recommend the choice of $\beta^* = 0.5$ as to not overly favor nor discriminate against the classification of regions as inconsistent. We note that the ε-level global consistency probability is an exploratory metric that can inform investigators of the degree of heterogeneity between region-specific HRs. By measuring the strength of evidence in support of consistency as defined by ICH E17, this metric discourages the notion of attempting to prove consistency.

Consistency is commonly assessed by comparing region-specific treatment effects to the overall treatment effect. For the ith region, we define the ε-level local consistency probability as $P(\varepsilon < e^{\gamma_i - \gamma_{(-i)}} < \varepsilon^{-1} | \mathbf{D})$, where $\gamma_{(-i)}$ is the global treatment effect defined in Section 4.3.5 calculated without the ith region. If the underlying treatment effects are heterogeneous, the global treatment effect may not be representative of any region; we recommend the use of pairwise or global consistency measures in these cases due to practical interpretability.

4.4 Simulation Studies

We compare the BMA approach to two Cox proportional hazards models (CPHMs) and a Bayesian hierarchical model (BHM) using simulated data designed to imitate aspects of the LEADER trial. The first CPHM estimates only a global treatment effect, whereas the second CPHM estimates only region-specific treatment effects. In many MRCTs with a TTE endpoint, these CPHMs correspond to models typically used for the primary analysis and as part of exploratory subgroup analyses. We designed the BHM using the same model specification and prior elicitation used by the FDA in a post hoc analysis of the LEADER trial data (Rothmann, 2021), where the response modeled is the crude log hazard ratio for each region. Additional details about the model setup and prior elicitation for the CPHMs and the BHM can be found in Section B.2 of Appendix B.

4.4.1 Setup of Simulation Studies

For each simulation study, we randomly generated datasets with four regions and a total sample size of $N = 9340$, and we considered scenarios in which the sample size allocation and underlying region-specific treatment effects were varied. We simulated $10{,}000$ datasets for each scenario using a uniform accrual rate for the first 1.5 years and a maximum follow-up time of 5 years, and we chose the constant baseline hazard and dropout rate so that the datasets mirror the LEADER trial data with respect to the average number of events, the average percentage of

subjects who either completed follow-up or experienced an event, and the median follow-up time (see Section B.3 of Appendix B for details).

We focused our simulation studies on testing superiority of an investigational treatment over placebo. Setting $\gamma_0 = 0$ for both the global and region-specific hypotheses, we calculated the global rejection rate, the true positive rate (TPR) for regions with a beneficial treatment effect (which we refer to as *alternative regions*), and the false positive rate (FPR) for regions with no treatment effect (which we refer to as *null regions*). Additionally, we assessed estimation quality by calculating the mean squared error (MSE) of the region-specific treatment effect estimates from the BMA approach and the BHM relative to the estimates obtained from the second CPHM. We then compared the three modeling approaches with respect to the rejection rates and relative MSE.

We used $K = 8$ intervals for the piecewise constant baseline hazard under the BMA approach, where the interval boundaries were chosen such that each interval contained approximately the same number of events. To reflect *a priori* prior elicitation, we chose each element of $\boldsymbol{\mu}_{0\ell}$ to equal $\log(1.3)$ (i.e., the log of the non-inferiority bound established by the FDA), $\ell = 1, \ldots, L$, and we set $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$, $i = 1, \ldots, S$, $k = 1, \ldots, K$. We chose $\boldsymbol{\Sigma}_{0\ell}$ and $p(M_\ell)$ according to the recommendations in Sections 4.3.3 and 4.3.5, respectively, with $\alpha_0 = 0$ (i.e., uniform prior model probabilities). For the MCIRD, we set $\varepsilon = 0.77$ (i.e., 1.3^{-1}) when calculating the consistency probability measures. Lastly, we set $\beta^* = 0.5$ for the inconsistency probability threshold used to calculate ε-level global consistency.

4.4.2 First Simulation Study: Equal Regional Sample Sizes

For the first set of simulation studies, we defined the underlying treatment hazard ratio for alternative regions to equal 0.868 (i.e., treatment effect of -0.142). We considered five cases in which we incremented the number of null regions between zero and four, and we set the regional sample sizes to be equal for each case. In MRCTs with time-to-event outcomes, the expected number of events for each region depends on both the sample size and the underlying treatment

effect. Specifically, the number of events is expected to be greater in null regions than in alternative regions despite having the same sample size. Thus, this first set of simulation studies with equal sample size should not be mistaken for the case in which each region contributes the same expected number of events.

The results for the first set of simulation studies are shown in Figure 4.1. Compared to the CPHMs, the BMA approach results in similar global rejection rates and substantially lower MSE in each scenario. While the BHM results in lower MSE than the BMA approach for two of the four scenarios when estimating the treatment effect in alternative and null regions, the BHM's global rejection rate is drastically lower across all scenarios (e.g., 0.27 compared to the BMA approach's rejection rate of 0.72 in the 0-null-regions case). We note that the BMA approach results in higher TPRs for the 0- and 1-null-regions cases while also having inflated FPRs; however, this should be of comparatively less concern considering that the primary objectives of MRCTs

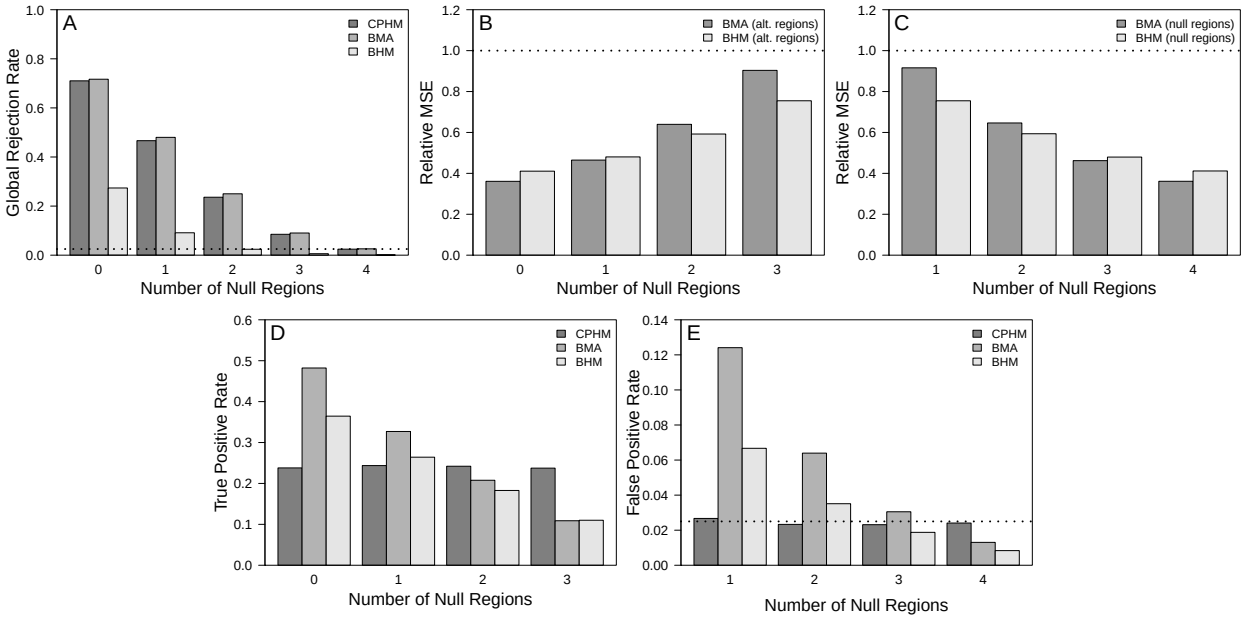

Figure 4.1: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for the simulation study with equal regional sample sizes. Alternative regions have a treatment-to-placebo hazard ratio of 0.868.

include testing the global treatment effect and estimating region-specific treatment effects (as opposed to formal hypothesis testing).

When $\varepsilon = 0.77$, the average ε-level global consistency probabilities for the 0- and 2-null-regions scenarios are estimated to be 0.97 and 0.90, respectively. We further investigated the nature of the global consistency metric by plotting the average probabilities over a range of ε for values of $\beta^* \in \{0.2, 0.5, 0.8\}$ (see Figure B.1 in Appendix B). As expected, increasing β^* for a fixed ε led to an increase in the ε-level global consistency probability across all five scenarios in which the number of null regions vary.

4.4.3 Second Simulation Study: Unequal Regional Sample Sizes

In the second set of simulation studies, we changed the regional sample sizes to equal the number of patients enrolled in Asia (AS), Europe (EU), North America (NA), and Rest of the World (RW) in the LEADER trial ($n_{AS} = 711$, $n_{EU} = 3296$, $n_{NA} = 2847$, $n_{RW} = 2486$). We defined the underlying region-specific treatment HRs for two different scenarios: (1) set HR = 0.868 for all regions, and (2) set the HRs equal to the estimates obtained in the original analysis of the LEADER data ($\text{HR}_{AS} = 0.62$, $\text{HR}_{EU} = 0.82$, $\text{HR}_{NA} = 1.01$, $\text{HR}_{RW} = 0.83$). We note that this second scenario is an extreme case to reflect the possibility that the HR point estimates from the original analysis were accurately estimated (i.e., the true underlying treatment effects are heterogeneous across regions with one region having no beneficial effect).

The results for the second set of simulation studies are shown in Figure 4.2. For the first scenario, the BMA approach and the CPHM have comparable global rejection rates, and the BMA approach results in the lowest MSE for three of the four regions. The BHM has a slightly lower MSE for Asia, but the global rejection rate is only a third of the rejection rate from the BMA approach (0.24 versus 0.72). In the second scenario, the BMA approach has a similar global rejection rate as the CPHM, and both the BMA approach and BHM have lower MSE than the CPHM for Europe and Rest of the World (the two regions with both a larger sample size and

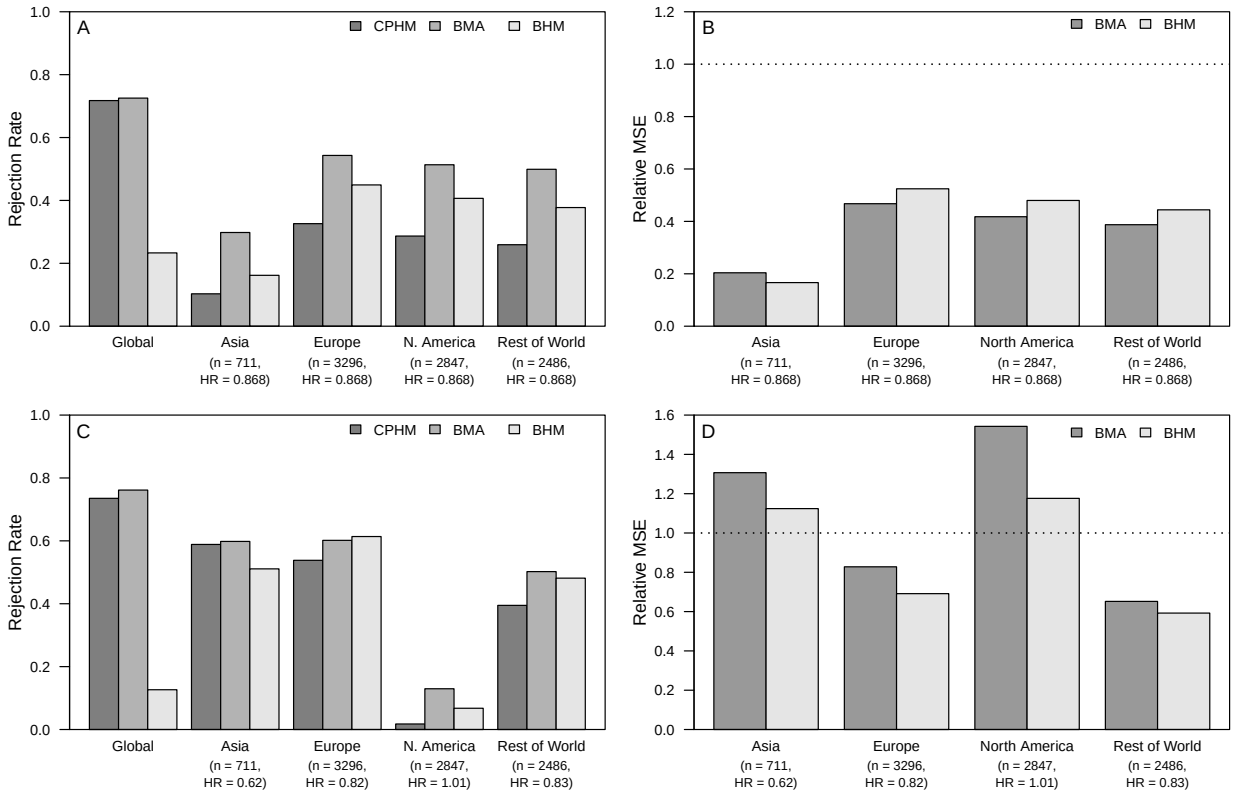

Figure 4.2: Rejection rates (*Panel A*) and MSE relative to CPHM (*Panel B*) for the scenario with all treatment-to-placebo hazard ratios equal to 0.868, and rejection rates (*Panel C*) and relative MSE (*Panel D*) for the scenario with differing treatment-to-placebo hazard ratios.

a favorable treatment effect). The BHM has lower MSE than the BMA approach for each region and a drastically lower global rejection rate.

We plotted the average ε-level global consistency probabilities over a range of ε values for both scenarios (see Figure B.2). For the first scenario with equal underlying region-specific HRs, the average global consistency probability is approximately 0.97 for $\varepsilon = 0.77$. In the second scenario with unequal region-specific HRs, this average probability decreases to 0.63, illustrating the ability of this metric to discern differences between the two extreme scenarios when quantifying the strength of evidence in favor of global consistency.

The average ε-local consistency probabilities for the first scenario when $\varepsilon = 0.77$ are 0.96 (AS) and 0.99 (EU, NA, and RW). For the second scenario, these average ε-local consistency probabilities are approximately 0.82 (AS), 0.87 (NA), and 0.96 (EU and RW). While the results

for local consistency clearly differ between the two scenarios, we again note that the local consistency probabilities alone can be difficult to interpret in the likely occurrence that some degree of heterogeneity exists between the regions. Thus, we recommend using both the global and local consistency probabilities to obtain a holistic understanding of the extent to which region-specific treatment effects are consistent with one another and the global effect.

4.4.4 Additional Simulation Studies

To better understand the performance of the BMA approach under different scenarios, we conducted additional simulation studies similar to the first study. We first considered two cases when the sample sizes of null regions are half/double the sizes of alternative regions, and we then considered the case with equal samples sizes and a non-constant piecewise baseline hazard. The results for these simulation studies are included in Section B.5 of Appendix B. The patterns observed in the global rejection rates and MSE of region-specific treatment effects for these three simulation cases are similar to the patterns from the first simulation study, demonstrating the robustness of the BMA approach in scenarios when regional sample sizes vary or when the underlying baseline hazard is non-constant.

We also compared the BHM with a gamma prior on the hierarchical precision parameter τ to two BHMs with priors (uniform and half-Cauchy) on the standard deviation $\tau^{-\frac{1}{2}}$ according to the recommendations of Gelman (2006). The BHMs with a gamma prior on τ and the half-Cauchy prior resulted in similar global rejection rates and MSE while the BHM with the uniform prior resulted in lower global rejection rates. Further details on the results, sampling methods, and convergence diagnostics of the BHMs are included in Section B.5 of Appendix B.

4.4.5 Sensitivity Analyses

We also performed sensitivity analyses on the choices of K, $\boldsymbol{\mu}_{0\ell}$, and α_0. The results for each analysis, which are included in Section B.6 of Appendix B, show the robustness of the BMA

approach when choosing K or eliciting $\boldsymbol{\mu}_{0\ell}$ and α_0 with respect to the global rejection rate and MSE of the region-specific treatment effects.

For the first analysis, we considered values of $K \in \{4, 8, 12, 16\}$ when defining the number of time intervals in the piecewise constant baseline hazard. The global rejection rates and MSE of the region-specific treatment effects are approximately equal across scenarios for all values of K considered, as is to be expected with an underlying constant baseline hazard. Thus, overfitting with respect to the constant baseline hazard increases the computational complexity without providing additional benefit when testing the global treatment effect or estimating the region-specific effects. If the underlying baseline hazard is believed to be non-constant, K should be sufficiently large and the intervals defined such that each interval contains approximately the same number of events.

The next sensitivity analysis compares values of $\boldsymbol{\mu}_{0\ell}$ where all elements of $\boldsymbol{\mu}_{0\ell}$ are equal across models M_ℓ, $\ell = 1, \ldots, L$. Specifically, we considered values of $e^{\boldsymbol{\mu}_{0\ell}} \in \{0.7, 1.05, 1.3, 1.5\}$. While the TPRs and FPRs of the region-specific treatment effects increased as the magnitude $\boldsymbol{\mu}_{0\ell}$ increases, the choice of $\boldsymbol{\mu}_{0\ell}$ (and in turn $\boldsymbol{\Sigma}_{0\ell}$ using the recommended elicitation) made negligible difference on the global rejection rates. All choices of $\boldsymbol{\mu}_{0\ell}$ resulted in lower MSE for the BMA approach than the CPHMs; however, the BMA approach had slightly greater MSE than the BHM in both null and alternative regions for all scenarios when $e^{\boldsymbol{\mu}_{0\ell}} = 1.05$ (i.e., the case with the smallest prior variances along the diagonals of $\boldsymbol{\Sigma}_{0\ell}$). For the other three values of $e^{\boldsymbol{\mu}_{0\ell}}$, the BMA approach had the lowest MSE in alternative regions for the 0- and 1-null-regions scenarios and in null regions for the 3- and 4-null-regions scenarios.

For the third sensitivity analysis, we elicited different prior model probabilities with $\alpha_0 \in \{0, \pm 0.5, \pm 1, \pm 2, \pm 5\}$, which in turn varied the amount of information borrowing. The choice of α_0 made no discernable difference on the global rejection rate, and the MSE of region-specific treatment effects remained lower for the BMA approach compared to the CPHM across all scenarios for $\alpha_0 \geqslant -1$. Negative values of α_0 with large magnitude allow for greater information borrowing, resulting in major increases in TPRs and FPRs of the region-specific effects. As α_0 in-

creases and greater prior model probability is placed on models with more distinct region-specific treatment effects (i.e., less information borrowing), the BMA approach more closely resembles the CPHMs with respect to the MSE, TPRs, and FPRs.

4.5 Data Analysis: LEADER Trial

We analyzed the full analysis set of 9340 subjects from the LEADER trial using the BMA approach, and we compared these results to the treatment effects estimated from the original planned analysis and the post hoc analysis conducted by statisticians at the Center for Drug Evaluation and Research of the FDA. The original analysis estimated the global treatment effect using a CPHM with treatment arm as the only covariate, and region-specific treatment effects were estimated as part of an exploratory subgroup analysis using a CPHM with main effects for treatment arm, region, and their interaction (Esbjerg and Ogenstad, 2012). The FDA's post hoc analysis incorporated shrinkage estimation by modeling the crude log hazard ratio for each region using a BHM (Rothmann, 2021), the details of which are described further in Section 4.4 of this paper and Section B.2 of Appendix B.

For the BMA approach, we used a piecewise constant baseline hazard with $K = 8$ intervals, and we used the same priors as in the first simulation study; i.e., $\boldsymbol{\mu}_{0\ell} = \log(1.3)$ and $\boldsymbol{\Sigma}_{0\ell} = \text{Diag}\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$ ($\ell = 1, \ldots, L$), $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$ ($i = 1, \ldots, S$; $k = 1, \ldots, K$), and $\alpha_0 = 0$ (i.e., uniform prior model probabilities). We calculated 95% credible intervals for the global and region-specific treatment effects, and we compared these interval estimates to both the 95% confidence intervals from the original analysis and the 95% credible intervals from the FDA analysis. The results from all modeling approaches are shown in Figure 4.3.

The three approaches result in similar point estimates of the global treatment effect, and the BMA approach and CPHM from the original analysis provide enough evidence to support the superiority of liraglutide over placebo with respect to the global MACE incidence rate. As seen in the simulation studies in Section 4.4, the BHM suffers from low power when testing the global effect based on the hierarchical mean parameter, and the BHM in this data analysis does

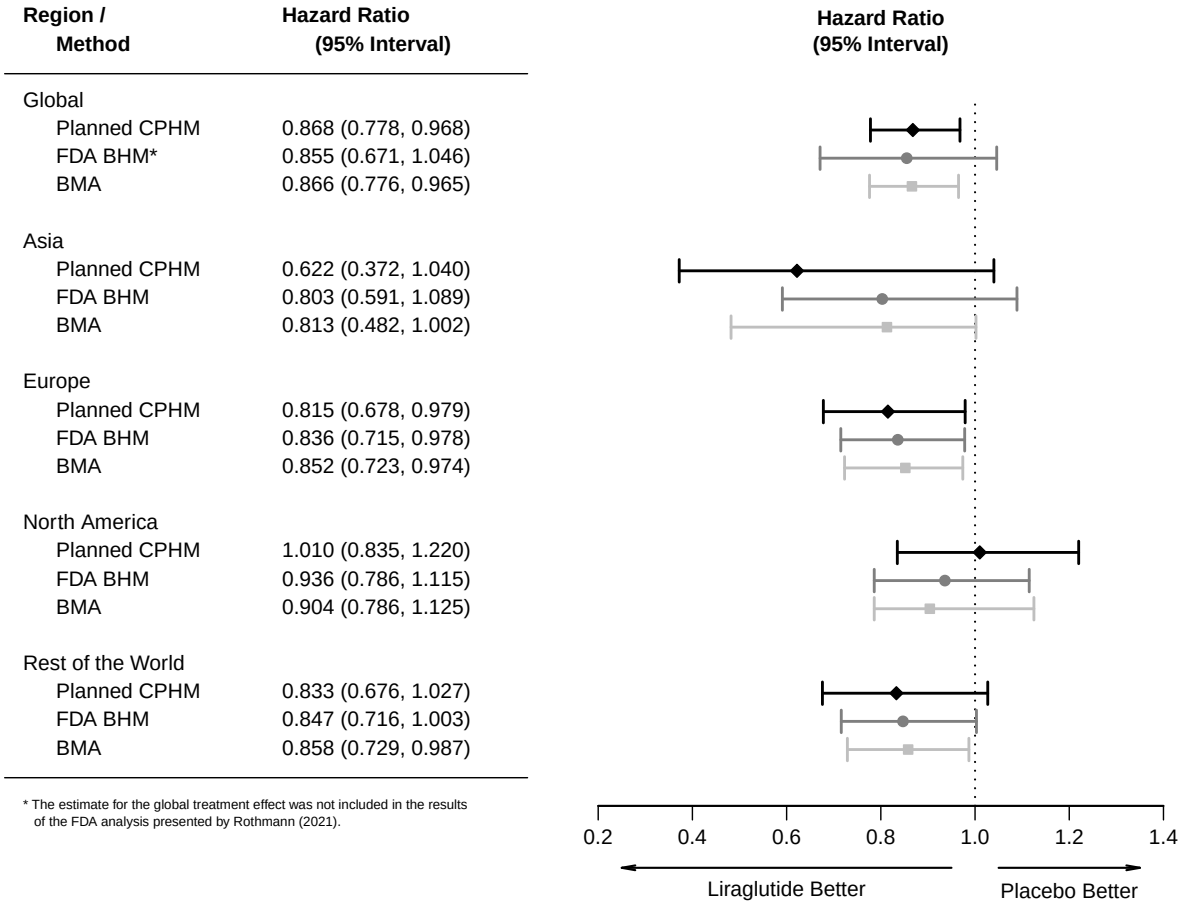

Figure 4.3: Comparison of global and region-specific hazard ratio estimates and 95% intervals for each analysis of the LEADER trial data.

not provide sufficient evidence to support the superiority of liraglutide based on the 95% credible interval. When estimating region-specific treatment effects, all point estimates obtained by both the BMA approach and the BHM are less than 1.00 and align with the global treatment effect estimate. Both approaches, which use information borrowing, result in a "regression to the global treatment effect" for region-specific effects.

It is worth noting that the purpose of the FDA's post hoc analysis of the LEADER trial data was to investigate the differences between the region-specific effect estimates. Hence, the FDA provided estimates only for the region-specific treatment effects and not the global effect (Rothmann, 2021). While the BHM may be appropriate for an exploratory analysis to better understand the extent to which treatment effects differ across regions, it is less suitable for evaluating

the global effect, at least based on the hierarchical mean parameter which is the most logical estimator. Additional information on the BHM's hierarchical parameters can be found in Section B.7 of Appendix B.

We further explored how the region-specific treatment effects relate to one another and to the global treatment effect by calculating the ε-level global and local consistency probabilities for values of $\varepsilon \in [0.7, 1.0)$ (see Figure 4.4). When $\beta^* = 0.5$, the global consistency probability is equal to 1.0 for values of $\varepsilon < 0.91$, indicating very strong evidence in favor of global consistency for these values of the MCIRD. If $\beta^* < 0.5$, then any pair of regions is more likely to be considered inconsistent, which in turn decreases the global consistency probability for a given value of ε. Thus, lower values of ε are required for the global consistency probability to reach 1.0 when β^* is small, and the opposite behavior is observed for values of $\beta^* > 0.5$. We note that the global consistency probability is the sum of PMPs for models that do not provide sufficient evidence of inconsistency, resulting in a step function with respect to ε for a set value of β^*. As ε increases toward one (with equivalence of effects corresponding to $\varepsilon = 1.0$), the global consistency probability for any value of β^* equals the PMP for the model that constrains all regions to share a common treatment effect.

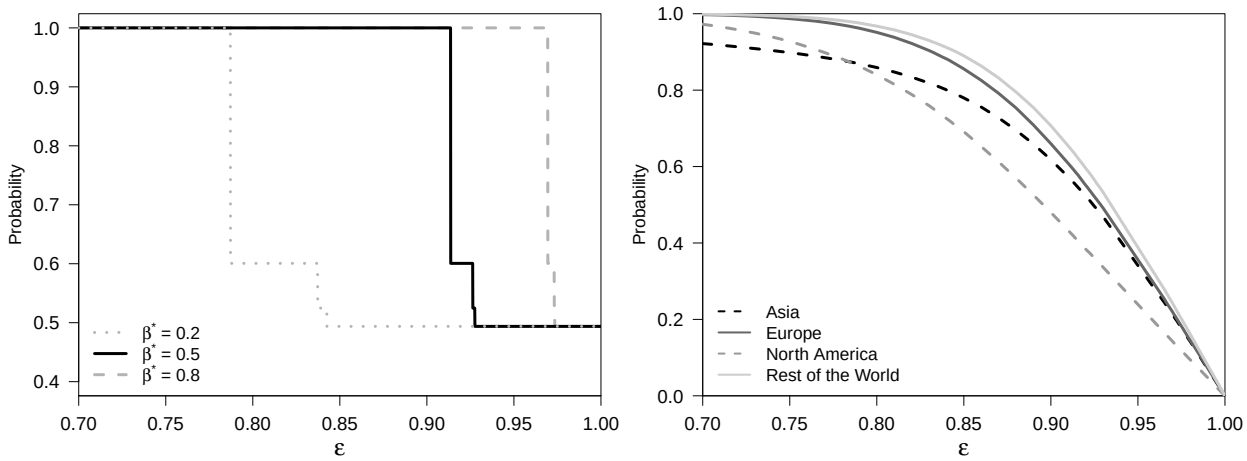

Figure 4.4: Comparison of ε-level global consistency probabilities for values of $\beta^* \in \{0.2, 0.5, 0.8\}$ (*left*) and ε-level local consistency probabilities for all four regions (*right*) for values of $\varepsilon \in [0.7, 1.0)$.

The right panel of Figure 4.4 shows how the ε-level local consistency probabilities differ across regions for $\varepsilon \in [0.7, 1.0)$. Unsurprisingly, the probabilities for North America and Asia are the lowest for all ε values, with probabilities of approximately 0.90 and 0.89, respectively, when $\varepsilon = 0.77$. When interpreted in conjunction with the global consistency probabilities using the same ε value, these probabilities provide reasonably strong evidence in favor of consistency between all region-specific hazard ratios and the global hazard ratio.

4.6 Discussion

The BMA approach provides better estimation quality of the region-specific treatment effects than CPHMs while maintaining a high global rejection rate within a single, comprehensive analysis. While it is plausible that region-specific treatment effects differ to some degree in MRCTs, major observed differences between the estimated treatment effects may be due in part to chance. The BMA approach mitigates this problem through information borrowing while still providing the flexibility of assuming that some level of heterogeneity between regions is possible. Like other approaches that utilize information borrowing, the BMA approach results in lower MSE when estimating the region-specific treatment effects at the price of increased false positive rates when testing these effects; however, it is worth noting that hypothesis testing on region-specific effects is uncommon in practice considering that the two main objectives of MRCTs relate to inference on the global effect and *estimation* of region-specific effects. Thus, emphasis of model performance should be placed on operating characteristics relating to these two objectives, both with which the BMA approach excels in comparison to competitor methods.

We also note that the BHM used by the FDA is appropriate for a post hoc analysis of the LEADER trial data to better understand the effects of liraglutide in each region; however, this model is not a reasonable choice for the primary analysis of MRCT data if treating the hierarchical mean parameter as the global treatment effect due to its low power and the challenge of *a priori* prior elicitation. Additionally, the BHM approach used by the FDA provides information only on the hazard ratio, whereas the piecewise constant baseline hazard regression model of

the BMA approach allows for flexible estimation of the survival curve and can be formulated as a BHM. Further, the BMA approach estimates PMPs that can inform investigators about similarities between the region-specific treatment effects to help with the assessment of consistency across regions.

A key assumption of BHMs is that the treatment effects of regions are a random sample from an underlying population, which may not be valid depending on how regions are defined; e.g., regions based on geography without accounting for intrinsic and extrinsic factors (Tanaka *et al.*, 2011). If investigators believe location effects are a random sample from a continuous distribution, they should give special care when defining regions in the planning stage of an MRCT to ensure that this assumption is reasonable. Careful consideration should also be given to the prior choice on the hierarchical variance parameter (Gelman, 2006).

While the BMA approach requires posterior samples of the region-specific treatment effects for all L models, the application of Laplace's method to approximate the full conditional distributions results in tremendous computational efficiency with minimal decrease in estimation accuracy. The BMA approach is easily and quickly executable for one dataset using only a single core on a computer, and a substantial increase in the number of subjects in a dataset makes little difference on the computation time due to the Laplace approximation.

CHAPTER 5: BAYESIAN MODEL AVERAGING FOR MULTI-REGIONAL CLINICAL TRIALS WITH A JOINT TIME-TO-EVENT ENDPOINT AND LONGITUDINAL MARKER

5.1 Introduction

As greater emphasis is placed on the globalization of drug development, multi-regional clinical trials (MRCTs), or trials that include multiple geographic regions under the same study protocol, have become a viable method for global data collection. Drug sponsors often rely on MRCTs to simultaneously seek approval for an investigational treatment from multiple regulatory authorities, resulting in their increased popularity among pharmaceutical companies in recent years (Song *et al.*, 2019). While MRCTs allow sponsors to more quickly introduce new treatments into the global market, these trials are also associated with several logistical and statistical challenges, prompting the International Council for Harmonisation (ICH) to publish the E17 document with guidelines for the planning and design of MRCTs in 2017. Among the guidelines include the recommendation to estimate region-specific treatment effects in addition to the overall treatment effect and the consideration of statistical methods that allow for information borrowing across regions if regional sample sizes are small (ICH, 2017).

Generally, the first major objective of MRCTs is to test and estimate the global treatment effect, and the second major objective is to assess the degree to which these results can be bridged to the individual regions. Some degree of heterogeneity in the treatment effects likely exists between regions due to intrinsic and extrinsic factors, however, these regional differences are often addressed only in exploratory subgroup analyses. Typically, the overall treatment effect (hereafter referred to as the *global treatment effect*) is estimated in the primary analysis using a fixed effects model, and region-specific treatment effects are later estimated using a separate fixed ef-

fects model as part of a regional subgroup analysis. Alternative models that account for regional heterogeneity include a continuous random effects model (CREM) with region-specific random treatment effects and the discrete random effects model (DREM) (Lan and Pinheiro, 2012), both of which assume different random mechanisms. The CREM assumes that region-specific treatment effects are samples from an underlying normal distribution, whereas the DREM assumes that regional sample sizes jointly follow a multinomial distribution. Depending on how regions are defined or if regional samples sizes are predetermined (e.g., based on regional regulatory requirements), these assumptions about the random effects may not be valid for an MRCT, and the model performances can be sensitive to the accuracy of these assumptions and the magnitude of the between-region variability (Li *et al.*, 2021).

To address the ICH E17 recommendations, we propose a Bayesian approach that incorporates information borrowing when estimating the region-specific treatment effects in addition to the global treatment effect. We consider all possible partitions of regions into sets, and we fit a unique model for each partition in which regions within a set are constrained to share the same treatment effect. We then average the posterior results from each model using Bayesian model averaging (BMA) (Hoeting *et al.*, 1999). This approach has previously been proposed for MRCTs with either a continuous endpoint (Bean *et al.*, 2021) or a time-to-event (TTE) endpoint (Bean *et al.*, 2022), and we now extend this method to models that jointly analyze a TTE endpoint and a continuous longitudinal marker. To the best of the authors' knowledge, this is the first proposed application of joint models to MRCTs.

The rest of this paper is organized as follows. In Section 5.2, we provide a brief description of joint models. We then discuss our proposed methodology using BMA in Section 5.3, and we compare this approach to survival models using simulation studies in Section 5.4. In Section 5.5, we apply the proposed approach to data from a high-profile MRCT that investigated the cardiovascular safety of an anti-diabetic treatment. Lastly, we close with discussion in Section 5.6.

5.2 Joint Models

Joint models typically include both a survival submodel and a longitudinal submodel. Different approaches for linking the two submodels include (i) using the observed longitudinal values as covariates in the survival submodel, (ii) fitting the longitudinal submodel first and then including the fitted values of the longitudinal trajectory for each subject in the survival submodel as covariates, and (iii) incorporating shared subject-specific random effects in both the longitudinal and survival likelihood. The majority of recent and ongoing research in joint modeling consider the third method, which generally results in less biased estimates than the first two methods (Sweeting and Thompson, 2011). In this paper, we discuss the joint modeling framework in the context of linking longitudinal and survival submodels via shared random effects.

Compared to classical models, such as the linear mixed model for longitudinal data and the Cox proportional hazards model for survival data, joint models can lead to higher power and lower sample sizes when testing the treatment effects on both the TTE outcome and the longitudinal marker by accounting for possible associations or dependencies between the two (Ibrahim *et al.*, 2010). When the longitudinal and survival outcomes are correlated, significant treatment effects on either outcome can be detected with greater sensitivity, which is especially beneficial when longitudinal outcomes are subjectively measured and susceptible to high variability (e.g., patient-reported outcomes) (Gould *et al.*, 2015).

The longitudinal submodel is commonly formulated as a linear mixed model of the form

$$X_i(t) = X_i^*(t) + \epsilon_i(t), \tag{5.1}$$

where $X_i(t)$ is the observed longitudinal outcome for the ith subject at time t, $X_i^*(t)$ is a trajectory function that depends on subject-specific random effects $\boldsymbol{b}_i$ (normally distributed), and $\epsilon_i(t) \sim N(0, \sigma^2)$. The errors $\epsilon_i(t)$ are independent and can be thought of as deviations due to

measurement error. The proportional hazards survival submodel at time t is generally written as

$$h\left(t|\, X_i^*, \boldsymbol{w}_{Y,i}\right) = h_0(t) \exp\left(g\left(\boldsymbol{\alpha}, X_i^*\right) + \boldsymbol{w}_{Y,i}^{\intercal} \boldsymbol{\theta}_Y\right), \tag{5.2}$$

where $h_0(t)$ is the baseline hazard function, $g(\cdot)$ is a function that defines the association structure, $\boldsymbol{\alpha}$ measures the association between the longitudinal marker and the TTE endpoint, and $\boldsymbol{w}_{Y,i}$ is a vector of covariates for the ith subject with corresponding effects $\boldsymbol{\theta}_Y$. In this paper, we consider an association structure that connects the longitudinal and survival submodels via shared random effects; i.e., $g\left(\boldsymbol{\alpha}, X_i^*\right) = \boldsymbol{\alpha}^{\intercal} \boldsymbol{b}_i$.

Faucett and Thomas (1996) propose a Bayesian joint model as defined in (5.1) and (5.2), and they define a piecewise constant baseline hazard for the survival submodel. Several variations of this model have been proposed, including joint models for multivariate longitudinal and survival data (Ibrahim *et al.*, 2004; Chi and Ibrahim, 2006) and models that allow for greater flexibility in the structure of the longitudinal submodel (Wang and Taylor, 2001; Brown and Ibrahim, 2003). Despite the computational challenges associated with Bayesian joint models, their implementation has become more accessible with R packages such as `JMbayes` (Rizopoulos, 2020) and `rstanarm` (Gabry *et al.*, 2022). We use the model formulation of Faucett and Thomas (1996) in our extension of joint models to MRCTs.

5.3 Methodology

5.3.1 Joint Model for a Time-to-Event Outcome and a Continuous Longitudinal Marker

Consider an MRCT with two treatment groups, S regions, and N subjects, and let n_i denote the sample size for the ith region where $\sum_{i=1}^{S} n_i = N$. We define the longitudinal and survival submodels in the form of (5.1) and (5.2), respectively, and we change all i indices to ij to indicate the jth patient from the ith region, $i = 1, \ldots, S$, $j = 1, \ldots, n_i$.

For the longitudinal submodel, let $\boldsymbol{w}_{X,ij}(t)$ be a vector of region indicators, functions of time (e.g., the identity function, polynomials, splines), region-specific treatment indicators, region-

specific interactions between treatment and all functions of time, and optional covariates (potentially time varying). Additionally, let $\boldsymbol{\theta}_X$ be the vector of regression effects on the longitudinal marker corresponding to $\boldsymbol{w}_{X,ij}(t)$. The trajectory function from (5.1) at time t can be written as

$$X^*_{ij}(t) = \boldsymbol{z}(t)^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}_{X,ij}(t)^\intercal \boldsymbol{\theta}_X, \tag{5.3}$$

where $\boldsymbol{z}(t)$ is an $r \times 1$ vector of functions at time t and $\boldsymbol{b}_{ij} \sim N_r(\boldsymbol{0}, \boldsymbol{G})$ is an $r \times 1$ vector of subject-specific random effects with positive-definite covariance matrix $\boldsymbol{G}$. We assume $\boldsymbol{b}_{ij}$ and $\epsilon_{ij}(t)$ are independent, and we define $\boldsymbol{\xi}_X = \{\boldsymbol{\theta}_X, \boldsymbol{G}, \tau\}$ where $\tau = \sigma^{-2}$. Suppose we observe K_{ij} measurements of the longitudinal marker for the jth patient from the ith region at times $t_{ij1}, \ldots, t_{ijK_{ij}}$, and we denote $X_{ijk} = X_{ij}(t_{ijk})$, $X^*_{ijk} = X^*_{ij}(t_{ijk})$, and $\boldsymbol{w}_{X,ijk} = \boldsymbol{w}_{X,ij}(t_{ijk})$ for notational convenience.

We model the TTE endpoint using the proportional hazards model from (5.2) with a piecewise constant baseline hazard function. Let T_{ij} be the true survival time and C_{ij} a potential right-censoring time for the jth subject from the ith region, and suppose we observe the TTE outcome $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\nu_{ij} = 1(T_{ij} < C_{ij})$, where $1(\cdot)$ is the indicator function. Define $\boldsymbol{w}_{Y,ij}$ to be an $(S + p_Y) \times 1$ vector of region-by-treatment indicators and p_Y optional baseline covariates (possibly with overlapping covariates in $\boldsymbol{w}_{X,ijk}$), and we define the $(S + p_Y) \times 1$ vector $\boldsymbol{\theta}_Y = (\boldsymbol{\gamma}_Y^\intercal, \boldsymbol{\beta}_Y^\intercal)^\intercal$ where $\boldsymbol{\gamma}_Y = (\gamma_{Y,1}, \ldots, \gamma_{Y,S})^\intercal$ is the vector of region-specific treatment effects on the TTE outcome (log hazard ratio scale) and $\boldsymbol{\beta}_Y$ is the vector of covariate effects. For the piecewise constant baseline hazards, we construct a finite partition of the time axis into Q intervals $(m_0, m_1]$, $(m_1, m_2]$, …, $(m_{Q-1}, m_Q]$, where $m_0 \equiv 0$ and $m_Q > \max(Y_{ij})$, $i = 1, \ldots, S$, $j = 1, \ldots, n_i$. We assume each interval has a separate region-specific constant baseline hazard $h_0(y_{ij}) = \lambda_{iq}$ for $y_{ij} \in (m_{q-1}, m_q]$, $q = 1, \ldots, Q$, and we define $\boldsymbol{\lambda} = (\lambda_{iq})^\intercal$. Using the formulation for the piecewise constant baseline hazard model as presented by Ibrahim *et al.* (2001), we

write the likelihood of the survival submodel given the random effects and $\boldsymbol{\xi}_Y = \{\boldsymbol{\alpha}, \boldsymbol{\theta}_Y, \boldsymbol{\lambda}\}$ as

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu} | \boldsymbol{\xi}_Y, \boldsymbol{b}) = \prod_{i=1}^{S} \prod_{j=1}^{n_i} \prod_{q=1}^{Q} \left[\left\{ \lambda_{iq} \exp\left(\boldsymbol{\alpha}^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}_{Y,ij}^\intercal \boldsymbol{\theta}_Y \right) \right\}^{\delta_{ijq}\nu_{ij}} \times \exp\left(-\delta_{ijq} \left\{ \lambda_{iq}(y_{ij} - m_{q-1}) + \sum_{g=1}^{q-1} \lambda_{ig}(m_g - m_{g-1}) \right\} \exp\left(\boldsymbol{\alpha}^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}_{Y,ij}^\intercal \boldsymbol{\theta}_Y \right) \right) \right],$$

where $\boldsymbol{Y} = (Y_{ij})^\intercal$ and $\boldsymbol{\nu} = (\nu_{ij})^\intercal$ are $N \times 1$ vectors and $\delta_{ijq} = 1(y_{ij} \in (m_{q-1}, m_q])$.

We denote the observed data for all subjects by $\mathbf{D} = \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\nu}, \boldsymbol{W}_X, \boldsymbol{W}_Y\}$ where $\boldsymbol{X}$ is the set of observed longitudinal markers, and $\boldsymbol{W}_X$ and $\boldsymbol{W}_Y$ are the design matrices corresponding to $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Y$, respectively. Setting $\boldsymbol{b} = (\boldsymbol{b}_{ij})^\intercal$, we write the joint density as $\mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu}, \boldsymbol{X} | \boldsymbol{\xi}_Y, \boldsymbol{\xi}_X, \boldsymbol{b}) = \mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu} | \boldsymbol{\xi}_Y, \boldsymbol{b}) \mathcal{L}(\boldsymbol{X} | \boldsymbol{\xi}_X, \boldsymbol{b})$, where $\mathcal{L}(\boldsymbol{X} | \boldsymbol{\xi}_X, \boldsymbol{b})$ is the likelihood of the longitudinal model.

5.3.2 Definition of the Model Space

We define the model space such that each model corresponds to a unique classification of regions into sets where regions within a set are constrained to share the same treatment effect (i.e., data from all regions within a set are pooled together to estimate a common treatment effect). Thus, the number of distinct treatment effects for a given model equals the number of sets into which regions are classified. This definition of a model space was first proposed by Psioda *et al.* (2021) in the context of basket trials, and Bean *et al.* (2021, 2022) extend the application to MRCTs with either a continuous or TTE endpoint. Similarly, we define both the longitudinal and survival submodels in this manner. For a given joint model, we allow the classification of regions into sets for one submodel to differ from the sets for the other submodel.

Let D_0 denote the maximum number of distinct treatment effects allowed in either submodel, where $1 \leqslant D_0 \leqslant S$. Considering the longitudinal or survival component individually, the number of submodels based on unique classifications of sets is a function of D_0 and S, calculated as $L_0 = \sum_{i=1}^{D_0} \left\{ \frac{1}{i!} \sum_{\ell=0}^{i} (-1)^\ell \binom{i}{\ell} (i-\ell)^S \right\}$. We let $M_{\ell,\ell'}$ denote the joint model that consists of the ℓth longitudinal submodel and the ℓ'th survival submodel, $1 \leqslant \ell \leqslant L_0$, $1 \leqslant \ell' \leqslant L_0$, and we denote the model space containing all $L = L_0^2$ possible joint models as $\mathcal{M}_{S,D_0}$. Additionally, we

let $D_{X,\ell}$ and $D_{Y,\ell'}$ be the number of distinct treatment effects on the longitudinal outcome and TTE outcome, respectively, for model $M_{\ell,\ell'}$ where $1 \leqslant D_{X,\ell} \leqslant D_0$ and $1 \leqslant D_{Y,\ell'} \leqslant D_0$.

For model $M_{\ell,\ell'}$, we define $\boldsymbol{W}_{X,\ell}$ and $\boldsymbol{W}_{Y,\ell'}$ to be variations of the design matrices $\boldsymbol{W}_X$ and $\boldsymbol{W}_Y$ in which we collapse together columns of region-specific treatment indicators that correspond to regions within the same set. We then let $\boldsymbol{w}_{X,\ell ijk}$ be the row of $\boldsymbol{W}_{X,\ell}$ for the jth patient in the ith region at time point k, and we similarly define $\boldsymbol{w}_{Y,\ell' ij}$ to be a row of $\boldsymbol{W}_{Y,\ell'}$ for the same patient. We also define $\boldsymbol{\theta}_{X,\ell}$ and $\boldsymbol{\theta}_{Y,\ell'}$ to be the updated vectors of regression coefficients corresponding to $\boldsymbol{W}_{X,\ell}$ and $\boldsymbol{W}_{Y,\ell'}$, respectively, and we let $\boldsymbol{\xi}_{X,\ell} = \{\boldsymbol{\theta}_{X,\ell}, \boldsymbol{G}, \tau\}$ and $\boldsymbol{\xi}_{Y,\ell'} = \{\boldsymbol{\alpha}, \boldsymbol{\theta}_{Y,\ell'}, \boldsymbol{\lambda}\}$. Given that $M_{\ell,\ell'}$ is the true model, we rewrite the trajectory function in (5.3) as

$$X^*_{ijk} = \boldsymbol{z}(t)^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}^\intercal_{X,\ell ijk} \boldsymbol{\theta}_{X,\ell},$$

and we write the survival submodel as

$$h\left(t \middle| X^*_{ijk}, \boldsymbol{w}_{Y,\ell' ij}\right) = h_0(t) \exp\left(\boldsymbol{\alpha}^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}^\intercal_{Y,\ell' ij} \boldsymbol{\theta}_{Y,\ell'}\right).$$

The likelihoods $\mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu} | \boldsymbol{\xi}_{Y,\ell'}, \boldsymbol{b}, M_{\ell,\ell'})$, $\mathcal{L}(\boldsymbol{X} | \boldsymbol{\xi}_{X,\ell}, \boldsymbol{b}, M_{\ell,\ell'})$, and $\mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu}, \boldsymbol{X} | \boldsymbol{\xi}_Y, \boldsymbol{\xi}_X, \boldsymbol{b}, M_{\ell,\ell'})$ are then updated accordingly.

We define the prior distributions for each element of $\{\boldsymbol{\xi}_{X,\ell}, \boldsymbol{\xi}_{Y,\ell'}, \boldsymbol{b}\}$ to be independent, where $\boldsymbol{\theta}_{X,\ell} | \tau, M_{\ell,\ell'} \sim N(\boldsymbol{\mu}_{X,\ell}, \tau^{-1} \boldsymbol{\Sigma}_{X,\ell})$, $\boldsymbol{G}^{-1} | M_{\ell,\ell'} \sim$ Wishart$_r(\nu_0, \boldsymbol{C}_0)$, $\tau | M_{\ell,\ell'} \sim$ Gamma $\left(\frac{\eta_\tau}{2}, \frac{\phi_\tau}{2}\right)$ such that $E(\tau | M_{\ell,\ell'}) = \frac{\eta_\tau}{\phi_\tau}$, $\boldsymbol{\alpha} | M_{\ell,\ell'} \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$, $\boldsymbol{\theta}_{Y,\ell'} | M_{\ell,\ell'} \sim N(\boldsymbol{\mu}_{Y,\ell'}, \boldsymbol{\Sigma}_{Y,\ell'})$, and $\lambda_{iq} | M_{\ell,\ell'} \sim$ Gamma(η_{iq}, ϕ_{iq}), $i = 1, \ldots, S$, $q = 1, \ldots, Q$, with hyperparameters $\{\boldsymbol{\mu}_{X,\ell}, \boldsymbol{\Sigma}_{X,\ell}, \nu_0, \boldsymbol{C}_0, \eta_\tau, \phi_\tau, \boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha, \boldsymbol{\mu}_{Y,\ell'}, \boldsymbol{\Sigma}_{Y,\ell'}, \eta_{iq}, \phi_{iq}\}$. For the prior on the survival regression effects, we follow the recommendation of Bean *et al.* (2022) by eliciting $\boldsymbol{\mu}_{Y,\ell'}$ to be the vector of best predictions for the corresponding treatment effects or covariate effects under model $M_{\ell,\ell'}$ and setting $\boldsymbol{\Sigma}_{Y,\ell'}$ to be the diagonal matrix Diag$\{(10 \times |\boldsymbol{\mu}_{Y,\ell'}|)^2\}$. Similarly, we recommend setting $\boldsymbol{\Sigma}_{X,\ell} =$ Diag$\{(10 \times |\boldsymbol{\mu}_{X,\ell}|)^2\}$ where the first S elements of $\boldsymbol{\mu}_{X,\ell}$ are the best predictions for the region-specific intercepts in the longitudinal submodel and the remaining elements are set as the best

prediction of the rate of change in the longitudinal marker (e.g., possibly obtained from previous clinical trials).

5.3.3 Posterior Distributions for Model $M_{\ell,\ell'}$

Let $\boldsymbol{\xi}^*_{\ell,\ell'} = \{\boldsymbol{\theta}_{X,\ell}, \boldsymbol{G}, \tau, \boldsymbol{\alpha}, \boldsymbol{\theta}_{Y,\ell'}\}$ be the set of parameters without the piecewise constant baseline hazards $\boldsymbol{\lambda}$ for model $M_{\ell,\ell'}$. We derive the joint posterior distribution of $\{\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}\}|M_{\ell,\ell'}$ as

$$p(\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}|\mathbf{D}, M_{\ell,\ell'}) = \int \mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu}, \boldsymbol{X}|\boldsymbol{\xi}_Y, \boldsymbol{\xi}_X, \boldsymbol{b}, M_{\ell,\ell'}) p(\boldsymbol{\xi}_{X,\ell}, \boldsymbol{\xi}_{Y,\ell'}, \boldsymbol{b}|M_{\ell,\ell'}) d\boldsymbol{\lambda},$$

where $p(\boldsymbol{\xi}_{X,\ell}, \boldsymbol{\xi}_{Y,\ell'}, \boldsymbol{b}|M_{\ell,\ell'})$ is the joint prior distribution. By integrating over $\boldsymbol{\lambda}$, we reduce the complexity of the joint model and eliminate the need to sample the posterior baseline hazards.

To make inference on $\boldsymbol{\xi}^*_{\ell,\ell'}$, we first integrate over the subject-specific random effects; i.e.,

$$p(\boldsymbol{\xi}^*_{\ell,\ell'}|\mathbf{D}, M_{\ell,\ell'}) = \int p(\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}|\mathbf{D}, M_{\ell,\ell'}) d\boldsymbol{b}. \tag{5.4}$$

We note that this integral does not have a closed-form solution, and thus we must rely on approximation methods. One common approach for solving high-dimensional integrals over random effects is to apply a Laplace approximation in which inference for model parameters can be made by conditioning on maximized values of the random effects. In the context of non-linear mixed models, Wolfinger and Lin (1997) approximate the integral by making a quadratic Taylor-series expansion of the integrand about the empirical best linear unbiased predictors of the random effects, and Lindstrom and Bates (1990) and Vonesh (1996) obtain the modes of the random effects using maximization algorithms that iterate between the fixed effects and random effects until convergence. Similarly, Ripatti and Palmgren (2000) and Pankratz *et al.* (2005) use Laplace approximations for survival frailty models by maximizing the fixed effects and random effects with a two-step iterative algorithm. For joint models, Rizopoulos *et al.* (2009) estimate the model parameters using an EM algorithm in which they apply a Laplace approximation to high-dimensional integrals over the random effects in the E step. Specifically, they obtain the

modes of the random effects for each subject using the Newton-Raphson algorithm, and then they take a second-order Taylor series expansion of the logarithm of the integrand around these modes. Additionally, they note that many repeated measurements per subject are necessary to achieve a satisfactory approximation when using Laplace's method.

Similar to the previously discussed methods, we apply a Laplace approximation to (5.4) using an iterative algorithm to obtain the maximum a posteriori (MAP) estimates of the parameters in $\boldsymbol{\xi}^*_{\ell,\ell'}$ and the random effects $\boldsymbol{b}$ with respect to the full conditional distributions obtained from $p(\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}|\mathbf{D}, M_{\ell,\ell'})$. We denote these maximized values by $\hat{\boldsymbol{\xi}}^*_{\ell,\ell'}$ and $\hat{\boldsymbol{b}}$. For each parameter or vector of subject-specific random effects $\boldsymbol{b}_{ij}$, we condition on the most current MAP estimates of all other parameters and random effects, and we iterate through each parameter and vector of subject-specific random effects until convergence. Additional details for the algorithm and the full conditional distributions are included in Section C.1 in Appendix C.

Conditional on τ and $\boldsymbol{b}$, the longitudinal regression effects $\boldsymbol{\theta}_{X,\ell}$ under model $M_{\ell,\ell'}$ follow a multivariate normal distribution; however, the full conditional distribution of the survival regression effects $\boldsymbol{\theta}_{Y,\ell'}$ does not have a closed form. Thus, we use Laplace's method to approximate $p(\boldsymbol{\theta}_{Y,\ell'}|\boldsymbol{\alpha}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'})$ with the $N(\hat{\boldsymbol{\theta}}_{Y,\ell'}, \hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_{Y,\ell'}})$ distribution, where $\hat{\boldsymbol{\theta}}_{Y,\ell'}$ is the posterior mode of $p(\boldsymbol{\theta}_{Y,\ell'}|\boldsymbol{\alpha}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'})$ conditional on the MAP estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{b}$, and $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_{Y,\ell'}}$ is the negative inverse Hessian of $\log\{p(\boldsymbol{\theta}_{Y,\ell'}|\boldsymbol{\alpha}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'})\}$ evaluated at the MAP estimates. A similar approximation is made for the full conditional distribution of $\boldsymbol{\alpha}$.

We approximate the marginal likelihood of the data conditional on $M_{\ell,\ell'}$ as

$$p(\mathbf{D}|M_{\ell,\ell'}) \approx (2\pi)^{\frac{D(\boldsymbol{\xi}^*_{\ell,\ell'})}{2}} \mid \hat{\boldsymbol{\Psi}}_{\boldsymbol{\xi}^*_{\ell,\ell'}} \mid^{\frac{1}{2}} p(\mathbf{D}|\hat{\boldsymbol{\xi}}^*_{\ell,\ell'}, \hat{\boldsymbol{b}}, M_{\ell,\ell'}) p(\hat{\boldsymbol{\xi}}^*_{\ell,\ell'}|M_{\ell,\ell'}),$$

where $D(\boldsymbol{\xi}^*_{\ell,\ell'})$ is the sum of the lengths of all parameters in $\boldsymbol{\xi}^*_{\ell,\ell'}$, $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\xi}^*_{\ell,\ell'}}$ is the negative inverse Hessian of $\log\left\{p(\mathbf{D}|\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}, M_{\ell,\ell'}) p(\boldsymbol{\xi}^*_{\ell,\ell'}|M_{\ell,\ell'})\right\}$ evaluated at $\hat{\boldsymbol{\xi}}^*_{\ell,\ell'}$ and $\hat{\boldsymbol{b}}$, and $p(\mathbf{D}|\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}, M_{\ell,\ell'}) = \int \mathcal{L}(\boldsymbol{Y}, \boldsymbol{\nu}, \boldsymbol{X}|\boldsymbol{\xi}_Y, \boldsymbol{\xi}_X, \boldsymbol{b}, M_{\ell,\ell'}) p(\boldsymbol{\lambda}|M_{\ell,\ell'}) d\boldsymbol{\lambda}$ (Kass and Raftery, 1995). We refer readers to Section C.2 in Appendix C for details of the negative inverse Hessian matrices.

5.3.4 Inference via Bayesian Model Averaging

Given the marginal likelihoods for all models and the prior model probabilities, which we denote as $p(M_{\ell,\ell'})$ for model $M_{\ell,\ell'}$, we calculate the posterior model probability (PMP) for $M_{\ell,\ell'}$ as

$$p(M_{\ell,\ell'}|\mathbf{D}) = \frac{p(\mathbf{D}|M_{\ell,\ell'})p(M_{\ell,\ell'})}{\sum_{k=1}^{L_0}\sum_{k'=1}^{L_0} p(\mathbf{D}|M_{k,k'})p(M_{k,k'})}.$$

As an extension of the prior model probability elicitation suggested by Bean *et al.* (2022), we recommend setting $p(M_{\ell,\ell'}) \propto \exp\left(a_X D_{X,\ell} + a_Y D_{Y,\ell'}\right)$, where a_X and a_Y are scalars that influence the amount of information borrowing across regions in the longitudinal and survival submodels, respectively. Negative values of these scalars result in more information borrowing by placing greater prior weight on models with fewer distinct treatment effects, whereas positive values lead to less borrowing.

With the PMPs as weights, we obtain model-averaged posterior probabilities for $\boldsymbol{\theta}_Y$ as

$$p(\boldsymbol{\theta}_Y|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}, \mathbf{D}) = \sum_{\ell=1}^{L_0}\sum_{\ell'=1}^{L_0} p(\boldsymbol{\theta}_{Y,\ell'}|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}, \mathbf{D}, M_{\ell,\ell'})p(M_{\ell,\ell'}|\mathbf{D}). \tag{5.5}$$

Similarly for other parameters, we can apply BMA to get model-averaged posterior quantities.

Under model $M_{\ell,\ell'}$, we define the global treatment effect on the TTE outcome to be

$$\gamma_{Y,G}|M_{\ell,\ell'} = \sum_{d=1}^{D_{Y,\ell'}} \frac{n_{(\ell',d)}}{N}\gamma_{(Y,\ell',d)},$$

where $\gamma_{(Y,\ell',d)}$ is the dth distinct region-specific treatment effect for survival submodel ℓ' and $n_{(\ell',d)}$ is the combined sample size of all regions that share the dth distinct treatment effect. For the longitudinal submodel, we let $\boldsymbol{\gamma}_{X,\ell} = (\gamma_{(X,\ell,1)}, \ldots, \gamma_{(X,\ell,D_{X,\ell})})^\intercal$ denote the $D_{X,\ell} \times 1$ vector of region-specific treatment effects of interest (e.g., main effects, treatment-by-time interaction

effects). We calculate the corresponding global treatment effect as

$$\gamma_{X,G}|M_{\ell,\ell'} = \sum_{p=1}^{D_{X,\ell}} \frac{K_{(\ell,p)}}{K_N}\gamma_{(X,\ell,p)},$$

where $K_{(\ell,p)}$ is the number of longitudinal observations for subjects in regions that share the pth distinct treatment effect and K_N is the total number of longitudinal observations for all subjects; i.e., $K_N = \sum_{i=1}^{S}\sum_{j=1}^{n_i} K_{ij}$. If we allow γ_Y to denote either the global treatment effect or a region-specific treatment effect on the TTE outcome, we can test the hypotheses $H_0 : \gamma_Y \geqslant \gamma_0$ versus $H_1 : \gamma_Y < \gamma_0$ for some prespecified value γ_0 by calculating $P(\gamma_Y < \gamma_0|\mathbf{D}) = \sum_{\ell=1}^{L_0}\sum_{\ell'=1}^{L_0} P(\gamma_Y < \gamma_0|\mathbf{D}, M_{\ell,\ell'})p(M_{\ell,\ell'}|\mathbf{D})$, and we consider $P(\gamma_Y < \gamma_0|\mathbf{D}) > \pi_0$ to be strong enough evidence in favor of H_1 for a prespecified π_0. Likewise, the global and region-specific treatment effects on the longitudinal marker can be tested for any one- or two-sided hypotheses using BMA.

5.4 Simulation Studies

5.4.1 Motivating Example

As motivation for the simulation studies, we consider the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results (LEADER) trial, a phase 3 MRCT designed to test the cardiovascular safety of an anti-diabetic treatment liraglutide in adults with type 2 diabetes (Marso *et al.*, 2016). In 2008, the U.S. Food and Drug Administration (FDA) issued a guidance document requiring pharmaceutical companies with anti-diabetic treatments to establish non-inferiority of the treatment to a placebo with respect to the primary outcome of time from enrollment to first occurrence of a major adverse cardiovascular event (MACE) (FDA, 2008). Per FDA guidelines, a hierarchical testing strategy was implemented for the LEADER trial in which non-inferiority of liraglutide to placebo would be established if the upper bound of the 95% confidence interval (CI) was below the non-inferiority margin of 1.3, and superiority of liraglutide would subsequently be established if this upper bound was below 1.0.

The global treatment hazard ratio (HR) was estimated to be 0.868 (95% CI: 0.778–0.968), providing evidence of the superiority of liraglutide. Three of the four regions had treatment effect point estimates that supported this conclusion, however, the HR point estimate for North America marginally favored placebo (HR: 1.010; 95% CI: 0.835–1.220). Post hoc analyses suggest that the heterogeneity originally observed between the regions may be due to chance more than differences in intrinsic and extrinsic factors (Rothmann, 2021, Nielsen *et al.*, 2021), highlighting challenges that can arise in MRCTs when estimating region-specific treatment effects without borrowing information across regions.

A secondary objective of the LEADER trial was to assess the efficacy and safety of liraglutide with regard to clinically important events or other surrogate parameters of treatment (Esbjerg and Ogenstad, 2012). As one of several clinical assessments, glycated hemoglobin (HbA1c) was measured at 11 post-screening visits from the time of randomization through the end of the treatment period, and change from baseline in HbA1c to the last assessment during the treatment period was used as a supportive endpoint to the secondary objective.

5.4.2 Simulation Setup

To assess the performance of the proposed joint modeling approach, we simulate survival and longitudinal data that mirror the LEADER trial with respect to the MACE incidence rate and the HbA1c trajectories for both treatment groups. Each dataset has a total of $N = 9340$ subjects from $S = 4$ regions, and the number of longitudinal measurements per subject range between 1 and 11. We connect the survival and longitudinal data with shared subject-specific random intercepts, and we consider different cases in which we vary the regional sample sizes, the magnitude of the association parameter, and the underlying region-specific treatment effects on the TTE outcome across scenarios, where regions may have a beneficial effect (*alternative regions*) or no effect (*null regions*). Additional details of the data generation (e.g., the dropout rate, a steadily increasing baseline hazard, and the mean and standard deviation of the longitudinal data for each visit) are provided in Section C.3 in Appendix C.

Considering that the TTE outcome was the primary endpoint in the LEADER trial and HbA1c measurements were used only in part as a supportive endpoint to the secondary objective, we report simulation results with respect to only the global and region-specific treatment effects on the TTE outcome. Any subsequent mention of treatment effects refers to these survival effects. In the absence of other proposed joint models for MRCTs, we compare the BMA joint modeling approach (BMA-JM) to survival models that have previously been used to analyze data from the LEADER trial, each of which ignore potential correlations with the longitudinal data. Specifically, we consider (i) Cox proportional hazards models (CPHMs) similar to the models used in the primary and subgroup analyses of the LEADER trial (i.e., one CPHM to estimate the global treatment effect without accounting for region, and a second CPHM to estimate region-specific treatment effects with region-by-treatment interactions) and (ii) the BMA approach for survival data only (BMA-S) (Bean *et al.*, 2022). Additional model details for the CPHMs and the BMA-S approach are included in Section C.4 in Appendix C. We simulate $10{,}000$ datasets for each simulation scenario and set $\gamma_0 = 0$ and $\pi_0 = 0.975$ when testing the hypotheses in Section 5.3.4, and we compare the models with respect to the global rejection rate, the mean square error (MSE) when estimating region-specific treatment effects in both alternative and null regions, and the true positive rate (TPR) and false positive rate (FPR) when testing treatment effects for alternative and null regions, respectively.

For the prior elicitation of regression effects in both submodels of the BMA-JM approach, we set each element of $\boldsymbol{\mu}_{Y,\ell'}$ equal to $\log(1.3)$ (i.e., the non-inferiority bound established by the FDA on the log scale), and we rely on HbA1c data from previous trials to predict the intercept values and rate of change when eliciting $\boldsymbol{\mu}_{X,\ell}$. In a set of phase 3 trials comparing liraglutide to placebo in the presence of other oral anti-diabetic agents, one study estimated the change in HbA1c from a baseline value of 8.6% to be as much as -1.5% over a 26-week period (-0.25% per month, assuming a linear rate of change) (Raskin and Mora, 2010). Thus, we set the first S elements of $\boldsymbol{\mu}_{X,\ell}$ corresponding to region-specific intercepts equal to 8.6, and we set all other elements equal to -0.25. We let $\boldsymbol{\Sigma}_{Y,\ell'} = \text{Diag}\{(10 \times |\boldsymbol{\mu}_{Y,\ell'}|)^2\}$ and $\boldsymbol{\Sigma}_{X,\ell} = \text{Diag}\{(10 \times |\boldsymbol{\mu}_{X,\ell}|)^2\}$,

and we set $\nu_0 = 1$, $\boldsymbol{C}_0 = 1$, $\eta_\tau = 0.001$, $\phi_\tau = 0.001$, $\boldsymbol{\mu}_\alpha = 0$, $\boldsymbol{\Sigma}_\alpha = 1000$, and $\eta_{iq} = 0.01$ and $\phi_{iq} = 0.01$ for $i = 1, \ldots, S$, $q = 1, \ldots, Q$. For each model $M_{\ell,\ell'}$, we elicit the prior model probabilities as recommended in Section 5.3.4 with $a_X = 0$ and $a_Y = 0$ (i.e., uniform prior model probabilities), and we discuss the sensitivity of this elicitation in Section 5.4.5.

We define the piecewise constant baseline hazards in the survival submodel of the BMA-JM approach by dividing the time axis into $Q = 8$ intervals where each interval contains approximately the same number of observed events, and we estimate a single association parameter corresponding to shared random intercepts. For the longitudinal submodel, we construct linear splines with knots at $t \in \{3, 18\}$, and we account for potential time-by-treatment interactions for each spline. Considering the small regional sample sizes typical of MRCTs, we assume that only models with at most two distinct treatment effects in either submodel can practically be identified with meaningful PMPs, and thus we set $D_0 = 2$ (i.e., $L_0 = 8$ longitudinal/survival submodels, or 64 possible joint models in $\mathcal{M}_{S,D_0}$).

5.4.3 First Simulation Study: Equal Sample Sizes

In the first set of simulation studies, we consider the case of equal regional sample sizes and two values of $\alpha \in \{0.5, 1.0\}$. We set the underlying treatment HR of alternative regions equal to 0.868, and we define five unique scenarios by differing the ratio of alternative to null regions.

The results for the simulation study when $\alpha = 0.5$ are shown in Figure 5.1. Compared to the survival models, the BMA-JM approach resulted in the highest global rejection rates for the scenarios with 0-3 null regions. Both BMA approaches resulted in lower MSE than the CPHM in all scenarios with the BMA-JM approach having the lowest MSE when estimating the treatment effect in alternative regions in the 2- and 3-null-regions scenarios. In exchange for better estimation quality of region-specific treatment effects, the BMA approaches had higher FPRs than the CPHM in scenarios with 1-3 null regions; however, we note that a major objective of MRCTs is to *estimate* the treatment effects within regions (as opposed to formal testing), and thus the inflated FPRs are of relatively low practical concern.

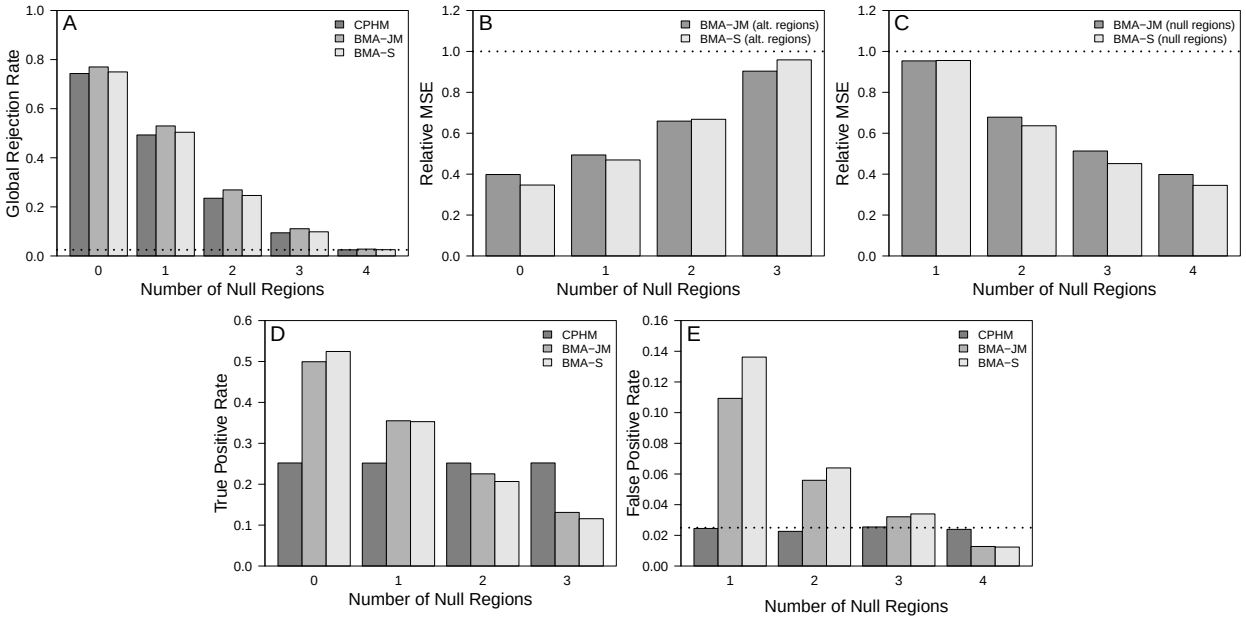

Figure 5.1: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for the simulation study with equal regional sample sizes and $\alpha = 0.5$. Alternative regions have a treatment-to-placebo hazard ratio of 0.868.

We also consider the case when $\alpha = 1.0$ (see Figure 5.2). In contrast to the BMA-S approach and the CPHM, the global rejection rates of the BMA-JM approach substantially increased as the magnitude of α increased. Additionally, we also observe an increase in the MSE with the BMA-JM approach relative to the CPHM when estimating the treatment effects of null regions in all scenarios. These results highlight the tradeoff of using a joint model versus a survival model when the TTE outcome and longitudinal marker are strongly associated: joint models may result in higher rejection rates of the global treatment effect, whereas survival models may estimate region-specific treatment effects with lower MSE as a result of ignoring the additional source of variation from the random effects. By not accounting for the correlated longitudinal data, survival models are prone to underestimate the variance of the region-specific treatment effects.

To further illustrate the behavior of reduced MSE with survival models, we conducted an additional simulation study in which we fit the CPHM and the BMA-S approach to joint data with a strong association ($\alpha = 1.0$) and to data with no association ($\alpha = 0$), and we compare the

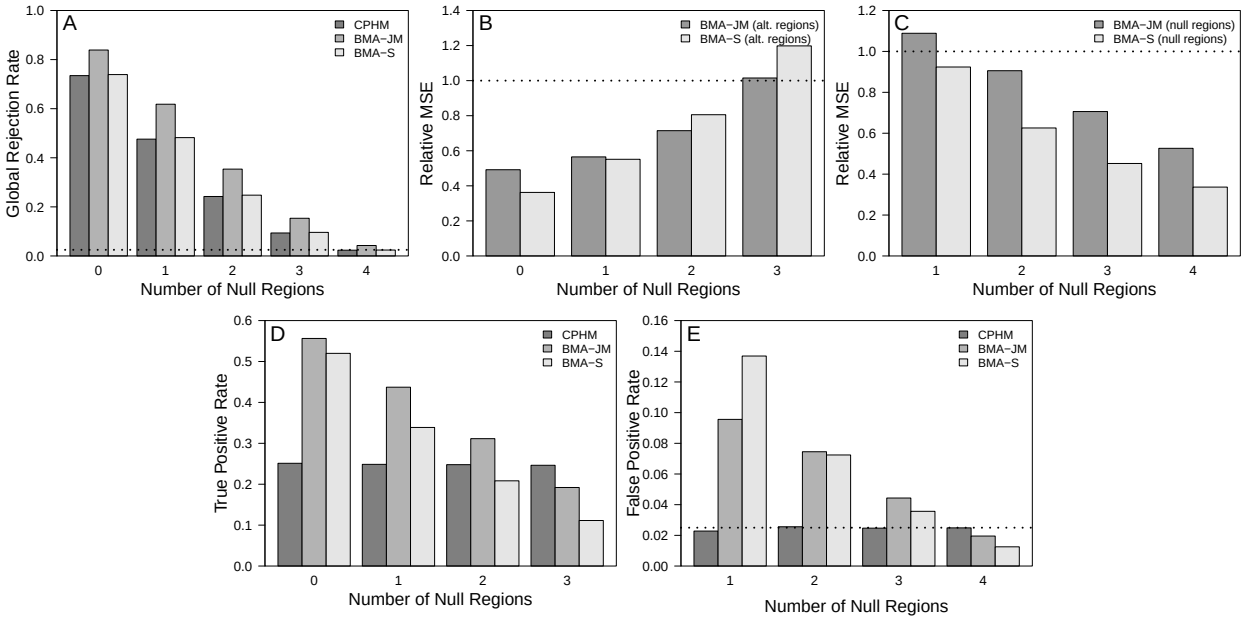

Figure 5.2: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for the simulation study with equal regional sample sizes and $\alpha = 1$. Alternative regions have a treatment-to-placebo hazard ratio of 0.868.

MSE and variance between the two types of data. As seen in Figure 5.3, we observe a consistent decrease in variance when estimating region-specific treatment effects as α increases, which in turn leads to lower MSE across scenarios. This reduction in MSE is particularly notable for the BMA-S approach when estimating the treatment effects for null regions.

5.4.4 Second Simulation Study: Unequal Sample Sizes

The second simulation study considers the case in which the regional sample sizes mirror the LEADER trial. Specifically, we set the sample sizes for Asia (AS), Europe (EU), North America (NA), and Rest of the World (RW) equal to $n_{AS} = 711$, $n_{EU} = 3296$, $n_{NA} = 2847$, and $n_{RW} = 2486$. The first scenario reflects the case where each region shares an underlying HR of 0.868, and the second scenario represents the extreme case in which the underlying HR of each region is equal to the point estimates from the LEADER trial subgroup analysis; i.e., $\text{HR}_{AS} = 0.622$, $\text{HR}_{EU} = 0.815$, $\text{HR}_{NA} = 1.010$, and $\text{HR}_{RW} = 0.833$. We once again choose $\alpha \in \{0.5, 1.0\}$.

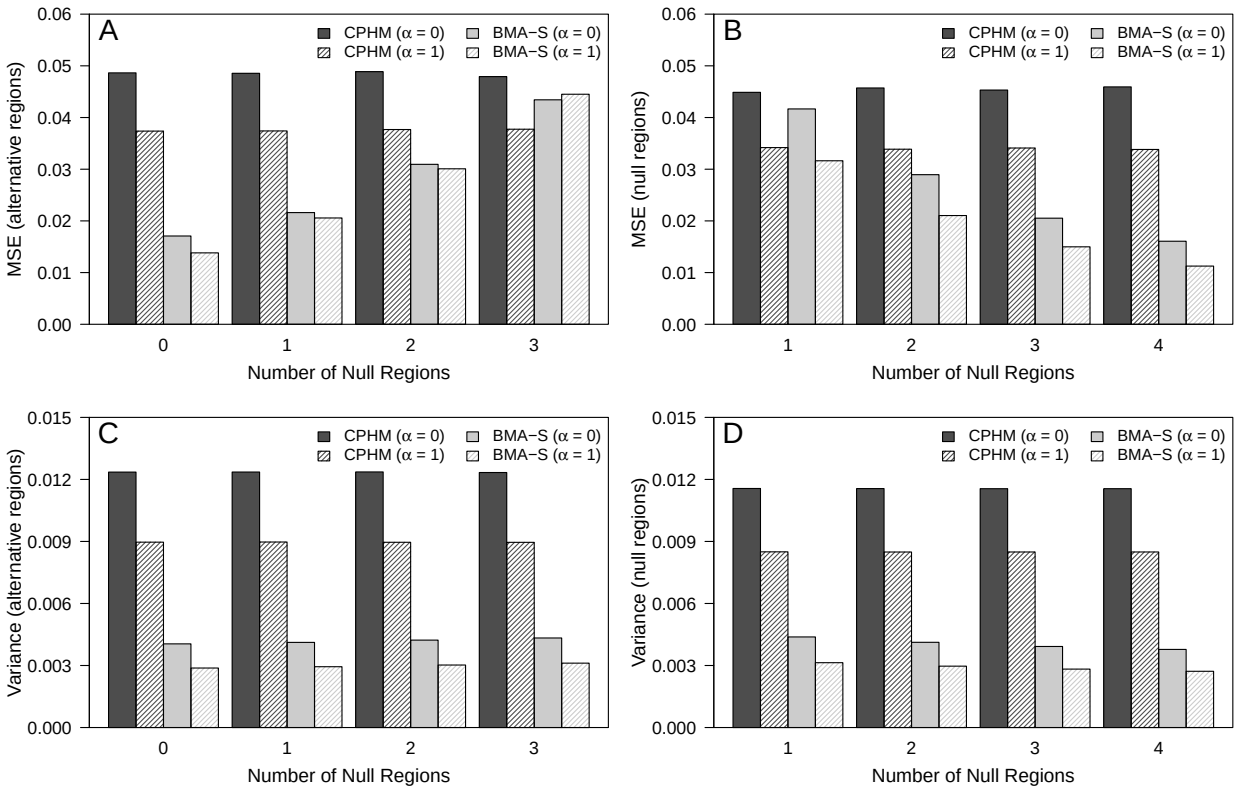

Figure 5.3: MSE for alternative regions (*Panel A*) and null regions (*Panel B*), and variance of region-specific treatment effects in alternative regions (*Panel C*) and null regions (*Panel D*), estimated from survival models fit to data with either a strong association ($\alpha = 1$) or no association ($\alpha = 0$) between the TTE outcome and longitudinal data. Alternative regions have a treatment-to-placebo hazard ratio of 0.868.

Figure 5.4 shows the results for the case when $\alpha = 0.5$. In both scenarios, the BMA-JM approach resulted in a higher global rejection rate compared to the survival models. For the scenario with equal underlying treatment effects, both BMA approaches resulted in substantially lower MSE than the CPHM with the BMA-S approach having marginally lower MSE than the BMA-JM approach for each region. Compared to the CPHM in the second scenario, both BMA approaches had lower MSE for Europe and Rest of the World (BMA-S slightly lower) and higher MSE for Asia and North America, the two regions with more extreme underlying HRs (BMA-JM lower).

The results when $\alpha = 1.0$ are included in Section C.5 in Appendix C. Similar to the first set of simulation studies, the BMA-JM approach resulted in a large increase in the global rejection

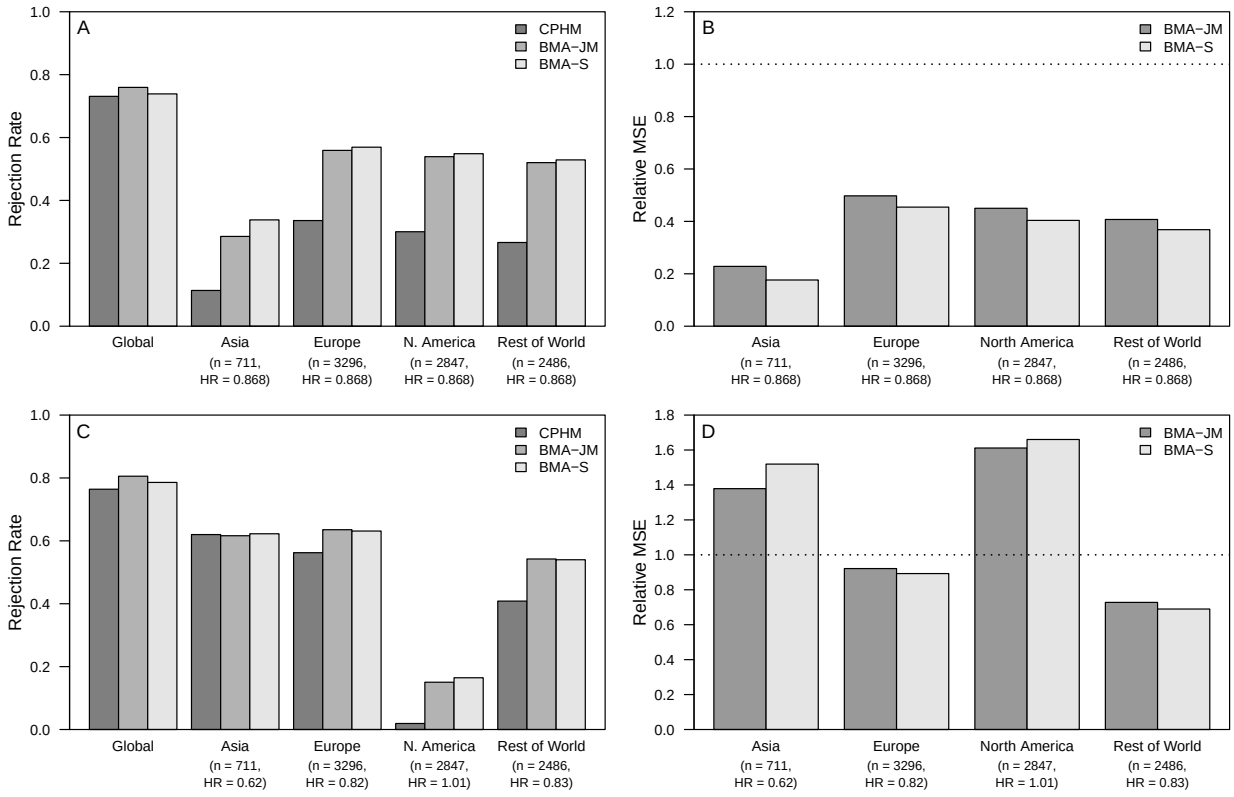

Figure 5.4: Rejection rates (*Panel A*) and MSE relative to CPHM (*Panel B*) for the scenario with all treatment-to-placebo hazard ratios equal to 0.868, and rejection rates (*Panel C*) and relative MSE (*Panel D*) for the scenario with differing treatment-to-placebo hazard ratios. Both scenarios consider unequal regional sample sizes and $\alpha = 0.5$.

rate and moderate increases in the MSE for several regions in both scenarios compared to the case when $\alpha = 0.5$. The increase in MSE for the first scenario is moderate for each region while remaining well below the MSE of the CPHM. In the second scenario, the largest increase in MSE that resulted from the increase in α occured for North America (i.e., the region with essentially no underlying treatment effect) while the MSE for Asia (i.e., the region with the greatest effect) remained approximately the same.

5.4.5 Additional Simulation Studies

We further investigate the behavior of the BMA-JM approach by varying the association parameter, the sample sizes of null regions compared to alternative regions, the prior model

probability elicitation, and the number of time intervals when defining the constant baseline hazards. The results from these simulation studies are included in Section C.5 in Appendix C.

We repeat the first two simulation studies using values of the association parameter in $\alpha \in \{0, 0.15\}$. As the magnitude of α decreased, the global rejection rates of the BMA-JM approach became more similar to the rates from the survival models in each scenario. For both values of α in the case with equal sample sizes, the BMA-JM had lower MSE compared to the BMA-S approach when estimating treatment effects in alternative regions for the 3-null-region scenario and in null regions for the 1-null-region scenario. The patterns in MSE between the BMA-JM and BMA-S approaches remained similar for all other scenarios with equal and unequal sample sizes in comparison to the corresponding scenarios when $\alpha = 0.5$.

Next, we considered the two cases when the sample sizes of null regions are half/double the sizes of alternative regions with $\alpha = 0.5$. Comparing these cases to the first simulation study with equal sample sizes, we observe similar relationships between the two BMA approaches with respect to the global rejection rates and MSE across all scenarios. For the case with smaller null regions, the global rejection rate substantially increased for all three modeling approaches in the scenarios with 1-3 null regions, and both BMA approaches experienced an increase in the MSE of alternative regions for the 3-null-regions scenario and a decrease in the MSE of null regions for the 1-null-region scenario; we observe opposite behaviors in the global rejection rates and MSE for these same scenarios when null regions are larger than alternative regions.

To better understand the sensitivity of the prior model probabilities, we consider all pairwise combinations of $a_X, a_Y \in \{-1, 0, 1\}$. The tested values of a_X and a_Y made no discernible difference with respect to the global rejection rate, and the choice of a_Y resulted in no change in the MSE or the TPR/FPR of the region-specific treatment effects for a given value of a_X. For a set a_Y, negative values of a_X resulted in decreases in the MSE when estimating the effect of null regions for the 0- and 1-null-region scenarios while increasing the MSE in the 4-null-regions scenario, and the MSE for alternative regions increased in the 1-null-region scenario while de-

creasing in the 3- and 4-null-regions scenarios; the opposite behavior is observed for positive values of a_X.

Lastly, we vary the number of time intervals when defining the constant baseline hazards for each region by considering values of $Q \in \{5, 8, 12\}$. For the case with a steadily increasing baseline hazard, the number of time intervals made no discernible difference with respect to any of the calculated operating characteristics. If the baseline hazard is believed to greatly increase over time, a sufficiently large value of Q should be chosen such that each interval contains approximately the same number of observed events.

5.5 Data Analysis: LEADER Trial

We apply the BMA-JM approach to data from the LEADER trial discussed in Section 5.4.1, and we compare the estimates of the global and region-specific HRs to estimates from the original primary and subgroup analyses using CPHMs and from the BMA-S approach. For both BMA approaches, we use the same model formulation and prior elicitation detailed in Section 5.4.2.

As seen in Figure 5.5, all three approaches resulted in similar point and interval estimates for the global treatment HR. Compared to the CPHM, both BMA approaches estimated region-specific HRs that are closer to the global estimate, and the corresponding 95% credible intervals are more narrow than the 95% confidence intervals from the CPHM. The point estimates and credible intervals from the BMA-JM approach and the BMA-S approach are similar for Europe, North America, and Rest of the World with the BMA-JM interval for North America being marginally smaller. Of the four regions, Asia has the largest observed differences in the point and interval estimates between the BMA-JM and BMA-S approaches.

The association parameter corresponding to the shared random intercepts is estimated to be 0.183 (95% credible interval: 0.135–0.232), indicating that HbA1c may have some effect on the time to first occurrence of a MACE. While this association is non-zero, both the BMA-JM and BMA-S approaches produced similar estimates for the treatment effect on the TTE outcome for this particular trial, likely due to the small magnitude of the association parameter. Despite

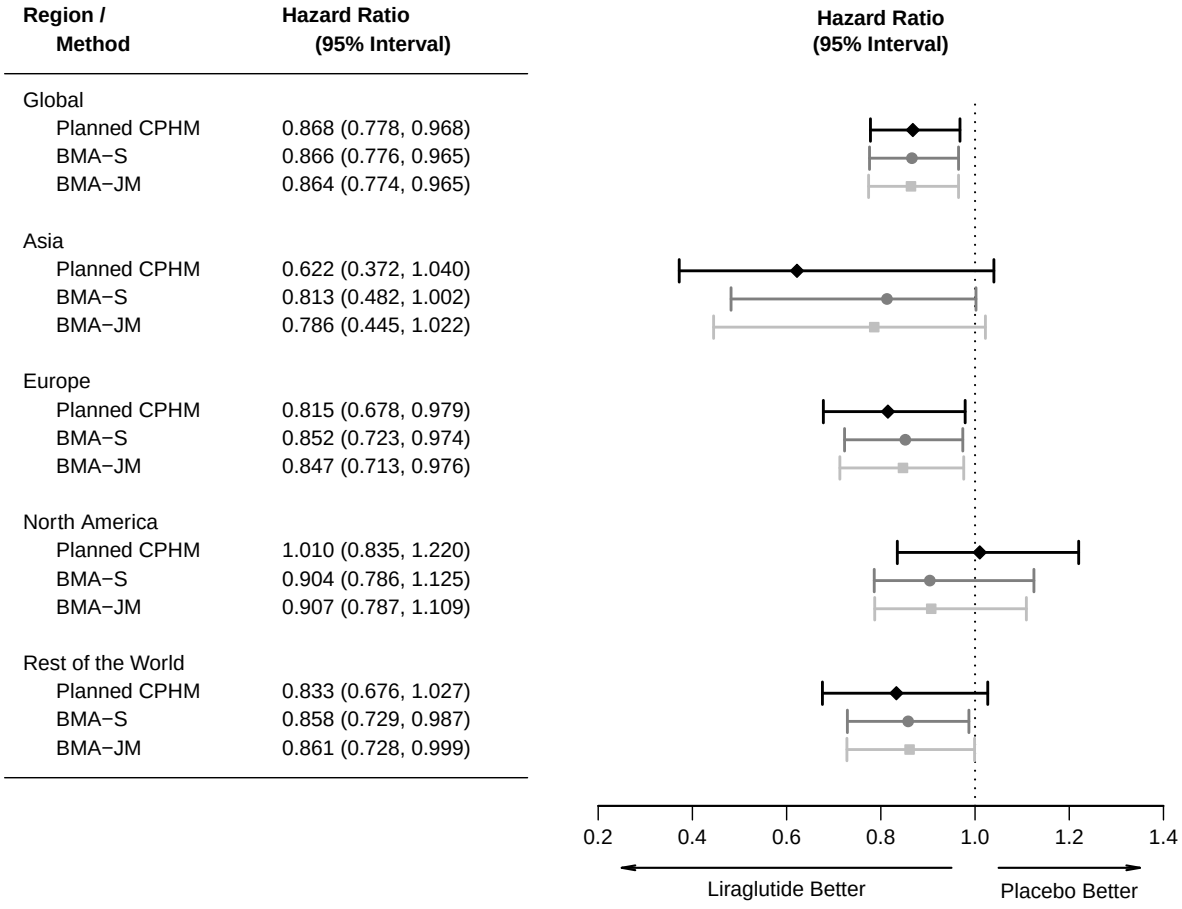

Figure 5.5: Comparison of global and region-specific hazard ratio estimates and 95% intervals for each analysis of the LEADER trial data.

the strength of the association, the BMA-JM approach provides the benefit of simultaneously estimating treatment effects on a longitudinal marker which may be of clinical interest in an MRCT (results for these treatment effects and PMPs are reported in Section C.6 in Appendix C). Further, as was illustrated with simulation studies in Section 5.4, we would expect to see additional benefits of the joint modeling approach when testing the global treatment effect on the TTE outcome for trials in which the association parameter is estimated to be of a higher magnitude (e.g., 0.5 or greater).

5.6 Discussion

The proposed BMA approach was motivated in part by the absence of joint modeling applications to MRCTs and by the need for statistical methodology that accounts for possible heterogeneity of the treatment effect across regions. While some degree of heterogeneity is expected to exist, only a few distinct treatment effects are likely to be identified in practice due to small regional sample sizes. Thus, we recommend reducing the number of possible models in the model space by choosing a small maximum number of distinct treatment effects allowed in either the longitudinal or survival submodel (e.g., set $D_0 = 2$ for a trial with $S = 4$ regions).

In the presence of a strong association between the TTE outcome and a longitudinal marker, the BMA-JM approach results in a substantially greater global rejection rate compared to survival models, whereas survival models may potentially estimate region-specific treatment effects on the TTE outcome with lower MSE by failing to account for the uncertainty from the random effects. Investigators should carefully consider this tradeoff if they suspect the TTE outcome to be associated with a longitudinal marker, and the choice of model may depend on the objectives of the study. In the case of most MRCTs, the first major objective is to make inference on the global treatment effect, and thus the joint modeling approach may be more appropriate if the survival and longitudinal data are believed to be correlated.

Using a computer with a single core, the data analysis from Section 5.5 took approximately 1.1 hours to run when $D_0 = 2$ (64 models in $\mathcal{M}_{S,D_0}$). The BMA-JM approach relies on a maximization algorithm that implements Laplace's method to integrate over subject-specific random effects and to approximate the posterior distributions of the treatment effects on the TTE endpoint. While existing software such as `JMbayes` may be appropriate for fitting each joint model defined in the model space with a single dataset, the computational demands and memory requirements make the use of this package and other pre-existing software infeasible in the necessary simulation studies when investigating the operating characteristics of the proposed approach.

An active research topic relating to MRCTs involves the evaluation of the consistency of treatment effects across regions, where consistency is defined by the ICH E17 document as a lack of clinically relevant differences (ICH, 2017). Possibilities for future research include developing methods to assess the consistency of the longitudinal treatment effect over time or exploring different association structures in the survival submodel.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

In this dissertation, we proposed a novel methodology to analyze data from a multi-regional clinical trial (MRCT). By utilizing Bayesian model averaging (BMA), this approach accounts for possible heterogeneity between the region-specific treatment effects while utilizing information borrowing across regions. We developed this approach for trials with a continuous endpoint (Chapter 3), a time-to-event (TTE) endpoint (Chapter 4), and a joint TTE endpoint and longitudinal marker (Chapter 5).

In Chapters 3 and 4, we detailed the application of the BMA approach to MRCTs with either a continuous or TTE endpoint, and we developed computationally efficient algorithms to obtain posterior samples by relying on either closed-form solutions (if the endpoint is continuous) or Laplace approximations (if the endpoint is TTE). We illustrated through simulation studies several advantages of this approach over commonly used models, including the ability to estimate region-specific treatment effects with lower mean squared error compared to fixed effects models (e.g., fixed effects linear models, Cox proportional hazards models). Unlike Bayesian hierarchical models, we showed that the BMA approach results in improved estimation quality of region-specific effects without compromising the rejection rate when testing the global treatment effect. Additionally, we developed three novel measures for testing the consistency of treatment effects across all regions, between any two regions, and between any given region and all other regions averaged together.

In Chapter 5, we further extended the BMA approach to joint models, and we used simulation studies to show the increased global rejection rate that can be obtained with the joint modeling approach compared to models that ignore underlying associations between the TTE outcome and longitudinal data. To overcome the computational complexities of fitting joint models, we

developed an algorithm that applies a Laplace approximation to integrate over the subject-specific random effects.

In addition to simulation studies, we conducted post hoc analyses of data from the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results (LEADER) trial using the proposed BMA approaches. In Chapter 4, we demonstrated the advantages of the BMA approach in a practical setting by highlighting the similar point and interval estimations of the global treatment effect compared to a Cox proportional hazards model and the similar estimates of the region-specific treatment effects compared to a Bayesian hierarchical model. In Chapter 5, we jointly analyzed the TTE outcome with repeated HbA1c measurements to simultaneously obtain estimates of the treatment effect on both outcomes.

Areas of future research include the extension of the BMA approach to MRCTs with other types of endpoints (e.g., binomial, count). The proposed approaches for evaluating the consistency of the treatment effects across regions can also be developed for longitudinal data, in which case we recommend methods that compare the trajectories for different regions. In the case of joint models, additional work can be done to incorporate different association structures in the survival submodel (e.g., inclusion of the entire trajectory function) and to further increase the computational efficiency of the approach when fitting all models in the model space.

APPENDIX A: ADDITIONAL RESULTS FOR CHAPTER 3

A.1 Comparison Models

A.1.1 Fixed Effects Linear Regression Models

Let $\boldsymbol{z}$ be the $N \times 1$ vector of treatment indicators ($z_i = 1$ if the ith patient received treatment and $z_i = 0$ otherwise), and $\boldsymbol{J}$ the $N \times 1$ vector of ones. Let $\boldsymbol{V}$ be the $N \times (S-1)$ matrix with region indicators corresponding to regions 2–S (region 1 is the reference region), and denote the observed data as $\mathbf{D} = (\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{V})$. For the first fixed effects linear model (FELM), we write the model as

$$\boldsymbol{Y} = \beta_0 \boldsymbol{J} + \boldsymbol{V}\boldsymbol{\beta}_R + \gamma \boldsymbol{z} + \boldsymbol{\epsilon},$$

where β_0 is the intercept, $\boldsymbol{\beta}_R$ is the $(S-1) \times 1$ vector of region effects corresponding to regions 2–S, γ is the global treatment effect, and $\boldsymbol{\epsilon}$ is distributed as $N_N(0, \tau^{-1}\boldsymbol{I}_N)$. We specify the following non-informative priors: $\beta_0 \sim N(0, 10{,}000)$, $\beta_R \sim N_{S-1}(0, 10{,}000\boldsymbol{I}_{S-1})$, $\gamma \sim N(0, 10{,}000)$, and $\tau \sim \text{gamma}(0.001, 0.001)$.

To test the global treatment effect, we calculate a 95% credible interval for $\gamma|\mathbf{D}$. We reject the global null hypothesis from Section 3.4 if γ_0 is less than the lower limit of the credible interval, where $\gamma_0 = 0$.

For the second FELM, let $\boldsymbol{V}_z$ be the $N \times (S-1)$ matrix with region-by-treatment indicators corresponding to subjects who received treatment in regions 2–S, and let $\mathbf{D} = (\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{V}, \boldsymbol{V}_z)$. When testing the region-specific treatment effects, we write the FELM as

$$\boldsymbol{Y} = \beta_0 \boldsymbol{J} + \boldsymbol{V}\boldsymbol{\beta}_R + \gamma \boldsymbol{z} + \boldsymbol{V}_z\boldsymbol{\gamma}_R + \boldsymbol{\epsilon},$$

where γ is the treatment effect for region 1 and $\boldsymbol{\gamma}_R$ is the $(S-1) \times 1$ vector of region-specific treatment effects corresponding to regions 2–S. In addition to the non-informative priors specified for the first FELM, we set the prior for γ_R as $N_{S-1}(0, 10{,}000\boldsymbol{I}_{S-1})$.

The region-specific treatment effects for regions 2–S are calculated as $(\gamma + \gamma_{Ri})|\mathbf{D}$, where γ_{Ri} is the ith element of γ_R, $i = 1, \ldots, S-1$. We calculate 95% credible intervals for each region-specific treatment effect and conclude that the data favor the alternative hypothesis from Section 3.4 if γ_0 is below the lower limit.

A.1.2 Bayesian Hierarchical Model

Now let $\boldsymbol{V}$ be the $N \times S$ matrix with region indicators for all S regions, and let $\boldsymbol{V}_z$ be the $N \times S$ matrix with region-by-treatment indicators corresponding to subjects who received treatment in regions 1–S. For the likelihood portion of the Bayesian hierarchical model (BHM), we assume

$$\boldsymbol{Y} = \beta_0 \boldsymbol{J} + \boldsymbol{V}\boldsymbol{\beta}_R + \gamma \boldsymbol{z} + \boldsymbol{V}_z \boldsymbol{\gamma}_R + \boldsymbol{\epsilon},$$

where β_0 is the fixed intercept, γ is the fixed treatment effect, $\boldsymbol{\beta}_R$ is the $S \times 1$ vector of random region-specific intercepts, $\boldsymbol{\gamma}_R$ is the $S \times 1$ vector of random region-specific treatment effects, and $\boldsymbol{\epsilon}$ is distributed as $N_N(0, \sigma_e^2 \boldsymbol{I}_N)$. We assume that $\boldsymbol{\beta}_R$ and $\boldsymbol{\gamma}_R$ are distributed $N_S(0, \sigma_\beta^2 \boldsymbol{I}_S)$ and $N_S(0, \sigma_\gamma^2 \boldsymbol{I}_S)$, respectively. For the fixed effects, we specify the prior distributions as $\beta_0 \sim N(0, 10{,}000)$ and $\gamma \sim N(0, 10{,}000)$. Due to the high sensitivity of gamma priors for the precision parameters when modeling BHMs (Cunanan *et al.*, 2019), we use bounded uniform priors for the standard deviation parameters. Specifically, we choose $\sigma_e \sim \text{uniform}(0, 100)$, $\sigma_\beta \sim \text{uniform}(0, 100)$, and $\sigma_\gamma \sim \text{uniform}(0, 0.01)$. The prior for σ_γ was chosen to be appropriately informative so that the global rejection rate is comparable to that of both the fixed effects model and the BMA approach. We note that the behavior of this model when testing and estimating region-specific treatment effects more closely resembles the second fixed effects model as the prior for σ_γ becomes more noninformative, and the global rejection rate dramatically decreases.

To test the null hypothesis for the global treatment effect, we calculate a 95% credible interval for $\gamma|\mathbf{D}$. For the region-specific treatment effects, we calculate a 95% credible interval for $(\gamma + \gamma_{Ri})|\mathbf{D}$, $i = 1, \ldots, S$.

A.2 Additional Results from Simulation Studies

A.2.1 Bias and Average Pairwise Consistency Probabilities for Simulation Study 1

- Simulation Study 1:
  - Equal regional sample sizes
  - Alternative regions have treatment effects equal to 0.034

Table A.1: Bias of region-specific treatment effects for simulations with equal regional sample sizes.

Number of Null Regions	Method	Bias (Alternative† Regions)	Bias (Null Regions)
0	FELM	-9.95×10^{-5}	—
	BMA	-1.50×10^{-3}	—
	BHM	-8.57×10^{-5}	—
1	FELM	-1.94×10^{-4}	5.95×10^{-4}
	BMA	-3.88×10^{-3}	9.55×10^{-3}
	BHM	-6.42×10^{-3}	2.56×10^{-2}
2	FELM	-1.54×10^{-5}	1.05×10^{-4}
	BMA	-5.87×10^{-3}	6.59×10^{-3}
	BHM	-1.27×10^{-2}	1.92×10^{-2}
3	FELM	1.60×10^{-4}	2.27×10^{-4}
	BMA	-8.06×10^{-3}	4.57×10^{-3}
	BHM	-1.89×10^{-2}	1.30×10^{-2}
4	FELM	-2.21×10^{-4}	-7.33×10^{-5}
	BMA	-1.12×10^{-2}	2.16×10^{-3}
	BHM	-2.57×10^{-2}	6.28×10^{-3}
5	FELM	—	-7.71×10^{-5}
	BMA	—	-1.90×10^{-4}
	BHM	—	-7.89×10^{-5}

† Treatment effect for alternative regions is 0.034

The average pairwise consistency probabilities (averaged across datasets) are reported in Table A.2. These average probabilities are approximately $0.08 - 0.11$ higher when assessing

consistency between two regions with equal effects (null or alternative) compared to assessments between one region with a null effect and another with an alternative effect. As the regional sample sizes increase, the average pairwise consistency probabilities for two regions with the same effect will increase for a set value of ε, and the average probabilities for two regions with different effects will decrease.

Table A.2: Average pairwise consistency probabilities $P(|\gamma_i - \gamma_j| < \varepsilon|\mathbf{D})$ between regions i and j where $\varepsilon = 0.018$. Pairwise comparisons are between two alternative† regions (darkly shaded), one alternative† and one null region (lightly shaded), and two null regions (no shading).

# Null Regions	i \ j	Region 1	Region 2	Region 3	Region 4	Region 5
0	1	1.000	0.605	0.606	0.607	0.608
	2	0.605	1.000	0.605	0.606	0.606
	3	0.606	0.605	1.000	0.608	0.608
	4	0.607	0.606	0.608	1.000	0.608
	5	0.608	0.606	0.608	0.608	1.000
1	1	1.000	0.503	0.505	0.503	0.505
	2	0.503	1.000	0.595	0.593	0.592
	3	0.505	0.595	1.000	0.595	0.594
	4	0.503	0.593	0.595	1.000	0.597
	5	0.505	0.592	0.594	0.597	1.000
2	1	1.000	0.586	0.486	0.488	0.488
	2	0.586	1.000	0.486	0.487	0.487
	3	0.486	0.486	1.000	0.593	0.590
	4	0.488	0.487	0.593	1.000	0.594
	5	0.488	0.487	0.590	0.594	1.000
3	1	1.000	0.580	0.578	0.488	0.487
	2	0.580	1.000	0.578	0.487	0.487
	3	0.578	0.578	1.000	0.491	0.488
	4	0.488	0.487	0.491	1.000	0.602
	5	0.487	0.487	0.488	0.602	1.000
4	1	1.000	0.580	0.582	0.584	0.503
	2	0.580	1.000	0.582	0.582	0.504
	3	0.582	0.582	1.000	0.582	0.507
	4	0.584	0.582	0.582	1.000	0.505
	5	0.503	0.504	0.507	0.505	1.000
5	1	1.000	0.597	0.597	0.598	0.597
	2	0.597	1.000	0.601	0.598	0.600
	3	0.597	0.601	1.000	0.599	0.597
	4	0.598	0.598	0.599	1.000	0.599
	5	0.597	0.600	0.597	0.599	1.000

† Treatment effect for alternative regions is 0.034

A.2.2 Rejection Rates, MSE, and Bias for Simulation Study 2

- Simulation Study 2:
  - Sample sizes of null regions are half the size of alternative regions
  - Alternative regions have treatment effects equal to 0.034

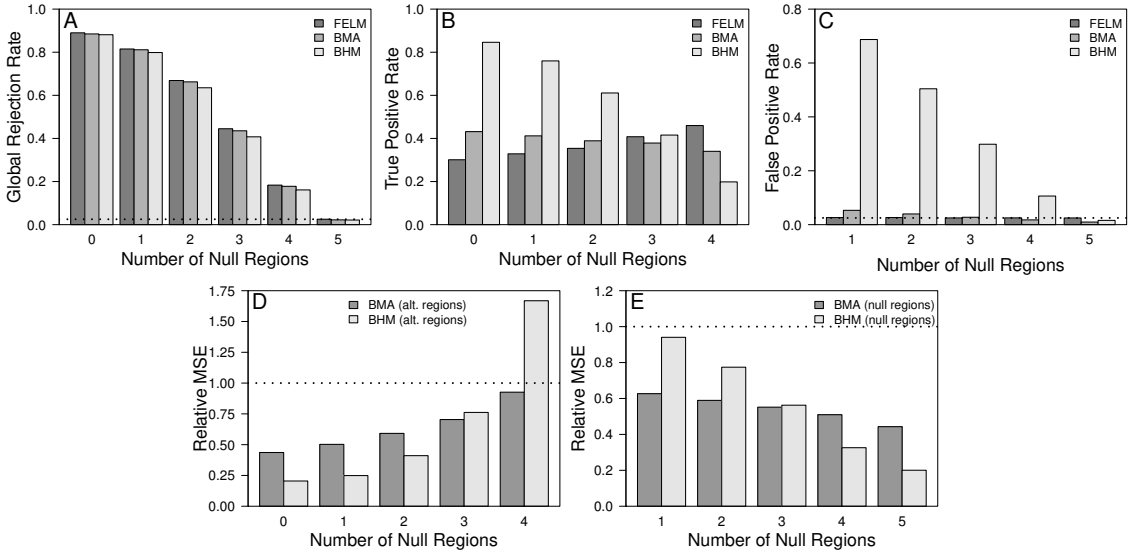

Figure A.1: Global rejection rates (*Panel A*), true positive rates for alternative regions (*Panel B*), false positive rates for null regions (*Panel C*), relative MSE (FELM as reference) for alternative regions (*Panel D*), and relative MSE for null regions (*Panel E*) for simulations with regional sample size allocation such that null regions are half the size of alternative regions. Alternative regions have a treatment effect of 0.034 L.

Table A.3: Bias of region-specific treatment effects for simulations where sample sizes of null regions are half the size of alternative regions.

Number of Null Regions	Method	Bias (Alternative[†] Regions)	Bias (Null Regions)
0	FELM	-2.10×10^{-5}	—
	BMA	-1.42×10^{-3}	—
	BHM	-8.16×10^{-6}	—
1	FELM	3.99×10^{-6}	1.37×10^{-4}
	BMA	-2.77×10^{-3}	1.17×10^{-2}
	BHM	-3.60×10^{-3}	2.91×10^{-2}
2	FELM	-3.25×10^{-4}	1.38×10^{-4}
	BMA	-4.43×10^{-3}	8.70×10^{-3}
	BHM	-8.30×10^{-2}	2.41×10^{-2}
3	FELM	1.15×10^{-4}	-2.25×10^{-4}
	BMA	-5.54×10^{-3}	5.68×10^{-3}
	BHM	-1.37×10^{-2}	1.82×10^{-2}
4	FELM	-5.86×10^{-5}	-7.16×10^{-5}
	BMA	-7.72×10^{-3}	3.10×10^{-3}
	BHM	-2.08×10^{-2}	1.03×10^{-2}
5	FELM	—	-1.26×10^{-4}
	BMA	—	-2.34×10^{-4}
	BHM	—	-1.22×10^{-4}

[†] Treatment effect for alternative regions is 0.034

A.2.3 Rejection Rates, MSE, and Bias for Simulation Study 3

- Simulation Study 3:
  - Sample sizes of alternative regions are half the size of null regions
  - Alternative regions have treatment effects equal to 0.034

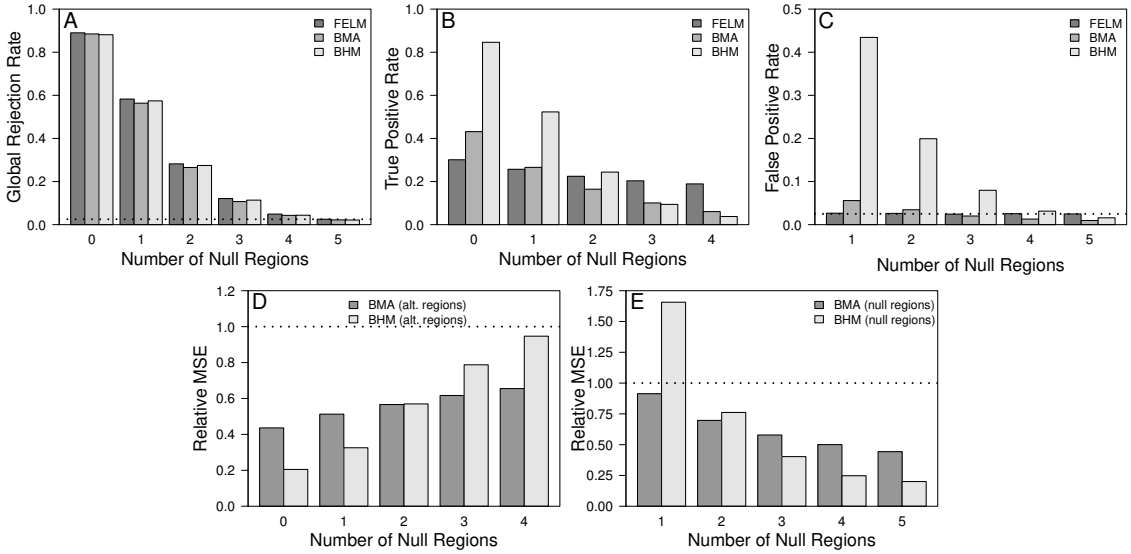

Figure A.2: Global rejection rates (*Panel A*), true positive rates for alternative regions (*Panel B*), false positive rates for null regions (*Panel C*), relative MSE (FELM as reference) for alternative regions (*Panel D*), and relative MSE for null regions (*Panel E*) for simulations with regional sample size allocation such that alternative regions are half the size of null regions. Alternative regions have a treatment effect of 0.034 L.

Table A.4: Bias of region-specific treatment effects for simulations where sample sizes of alternative regions are half the size of null regions.

Number of Null Regions	Method	Bias (Alternative[†] Regions)	Bias (Null Regions)
0	FELM	-2.10×10^{-5}	—
	BMA	-1.42×10^{-3}	—
	BHM	-8.16×10^{-6}	—
1	FELM	1.05×10^{-4}	5.64×10^{-4}
	BMA	-4.80×10^{-3}	6.93×10^{-3}
	BHM	-1.01×10^{-2}	2.11×10^{-2}
2	FELM	-1.87×10^{-5}	1.61×10^{-4}
	BMA	-7.99×10^{-3}	4.59×10^{-3}
	BHM	-1.81×10^{-2}	1.38×10^{-2}
3	FELM	-1.35×10^{-4}	-3.80×10^{-5}
	BMA	-1.14×10^{-2}	2.87×10^{-3}
	BHM	-2.43×10^{-2}	8.05×10^{-3}
4	FELM	-8.12×10^{-5}	5.18×10^{-5}
	BMA	-1.48×10^{-2}	1.47×10^{-3}
	BHM	-2.90×10^{-2}	3.68×10^{-3}
5	FELM	—	-1.26×10^{-4}
	BMA	—	-2.34×10^{-4}
	BHM	—	-1.22×10^{-4}

[†] Treatment effect for alternative regions is 0.034

A.2.4 Bias for Simulation Study 4

- Simulation Study 4:
  - Equal regional sample sizes
  - All regions have positive, heterogeneous treatment effects (see Table A.5)

Table A.5: Region-specific treatment effects for three scenarios.

Scenario	γ_1	γ_2	γ_3	γ_4	γ_5
1	0.017	0.026	0.034	0.043	0.051
2	0.017	0.017	0.017	0.034	0.034
3	0.017	0.034	0.034	0.051	0.051

Table A.6: Bias of region-specific treatment effects for simulations with heterogeneous treatment effects.

Region	Method	Scenario 1	Scenario 2	Scenario 3
1	FELM	-1.41×10^{-4}	-3.35×10^{-4}	1.11×10^{-4}
	BMA	5.47×10^{-3}	1.67×10^{-3}	6.59×10^{-3}
	BHM	1.63×10^{-2}	6.30×10^{-3}	1.91×10^{-2}
2	FELM	3.40×10^{-4}	-1.00×10^{-4}	-1.50×10^{-4}
	BMA	2.23×10^{-3}	1.76×10^{-3}	-1.49×10^{-4}
	BHM	7.85×10^{-3}	6.31×10^{-3}	3.12×10^{-3}
3	FELM	4.89×10^{-4}	1.99×10^{-5}	-4.55×10^{-4}
	BMA	-8.67×10^{-4}	1.88×10^{-3}	-3.56×10^{-4}
	BHM	3.39×10^{-4}	6.31×10^{-3}	3.10×10^{-3}
4	FELM	-1.869×10^{-4}	-3.63×10^{-5}	1.96×10^{-4}
	BMA	-5.11×10^{-3}	-5.48×10^{-3}	-7.10×10^{-3}
	BHM	-8.15×10^{-3}	-9.67×10^{-3}	-1.28×10^{-2}
5	FELM	9.66×10^{-5}	2.35×10^{-5}	-1.23×10^{-4}
	BMA	-8.40×10^{-3}	-5.46×10^{-3}	-7.25×10^{-3}
	BHM	-1.57×10^{-2}	-9.67×10^{-3}	-1.28×10^{-2}

A.3 Sensitivity Analysis

A.3.1 Vary Values of Hyperparameters in Regression Parameter Priors

- Equal regional sample sizes
- Model prior type: $\alpha_0 = 0$ used for uniform model priors
- True region-specific control means/intercepts: 0.082
- True mean difference under alternative: 0.034
- We vary the values of the region-specific control mean and the mean difference, considering cases when the assumed values are both half and double the true values specified for data generation.

Table A.7: Rejection rates and relative MSE (FELM as reference) for sensitivity analysis on hyperparameters in priors of regression parameters. Shaded cells correspond to scenarios with assumed true intercept and mean difference.

			——— Rejection Rates ———			— Relative MSE —	
Number of Null Regions	Assumed Intercept	Assumed mean difference	Global Rejection Rate	TPR (Alternative† Regions)	FPR (Null Regions)	(Alternative† Regions)	(Null Regions)
	0.082	0.034	0.887	0.415	—	0.43	—
	0.082	0.017	0.852	0.334	—	0.38	—
	0.082	0.068	0.894	0.469	—	0.43	—
	0.041	0.017	0.890	0.374	—	0.35	—
0	0.041	0.034	0.919	0.457	—	0.41	—
	0.041	0.068	0.927	0.516	—	0.41	—
	0.164	0.017	0.845	0.329	—	0.39	—
	0.164	0.034	0.881	0.408	—	0.44	—
	0.164	0.068	0.890	0.462	—	0.43	—
	0.082	0.034	0.717	0.340	0.056	0.50	0.77
	0.082	0.017	0.659	0.273	0.036	0.46	0.55
	0.082	0.068	0.731	0.376	0.077	0.50	0.86
	0.041	0.017	0.721	0.310	0.044	0.41	0.58
1	0.041	0.034	0.775	0.381	0.066	0.47	0.81
	0.041	0.068	0.788	0.419	0.090	0.47	0.91
	0.164	0.017	0.649	0.268	0.035	0.47	0.54
	0.164	0.034	0.709	0.335	0.055	0.51	0.77
	0.164	0.068	0.722	0.370	0.076	0.51	0.86
	0.082	0.034	0.473	0.279	0.037	0.58	0.66
	0.082	0.017	0.408	0.222	0.024	0.54	0.47
	0.082	0.068	0.492	0.300	0.047	0.59	0.70
	0.041	0.017	0.474	0.253	0.029	0.48	0.49
2	0.041	0.034	0.539	0.317	0.044	0.54	0.68
	0.041	0.068	0.555	0.337	0.057	0.55	0.74
	0.164	0.017	0.399	0.217	0.023	0.55	0.47
	0.164	0.034	0.464	0.274	0.036	0.59	0.65
	0.164	0.068	0.481	0.295	0.046	0.60	0.70
	0.082	0.034	0.239	0.225	0.025	0.67	0.57
	0.082	0.017	0.190	0.171	0.016	0.63	0.41
	0.082	0.068	0.253	0.234	0.030	0.70	0.59
	0.041	0.017	0.235	0.201	0.019	0.57	0.42
3	0.041	0.034	0.293	0.257	0.030	0.61	0.58
	0.041	0.068	0.308	0.269	0.037	0.64	0.61
	0.164	0.017	0.184	0.168	0.015	0.64	0.41
	0.164	0.034	0.233	0.221	0.025	0.68	0.57
	0.164	0.068	0.248	0.229	0.029	0.71	0.59
	0.082	0.034	0.082	0.162	0.016	0.82	0.50
	0.082	0.017	0.062	0.118	0.010	0.78	0.36
	0.082	0.068	0.090	0.161	0.018	0.89	0.51
	0.041	0.017	0.079	0.138	0.013	0.71	0.36
4	0.041	0.034	0.107	0.188	0.019	0.75	0.50
	0.041	0.068	0.114	0.189	0.022	0.81	0.51
	0.164	0.017	0.059	0.115	0.010	0.80	0.36
	0.164	0.034	0.080	0.159	0.016	0.83	0.50
	0.164	0.068	0.086	0.158	0.018	0.90	0.51
	0.082	0.034	0.023	—	0.009	—	0.43
	0.082	0.017	0.015	—	0.005	—	0.31
	0.082	0.068	0.026	—	0.010	—	0.43
	0.041	0.017	0.022	—	0.006	—	0.30
5	0.041	0.034	0.033	—	0.011	—	0.42
	0.041	0.068	0.036	—	0.012	—	0.42
	0.164	0.017	0.014	—	0.005	—	0.31
	0.164	0.034	0.022	—	0.009	—	0.44
	0.164	0.068	0.026	—	0.009	—	0.43

† Treatment effect for alternative regions is 0.034

A.3.2 Vary Values of Prior Model Probabilities

- Equal regional sample sizes
- Assumed control mean/intercept in regression coefficient priors: 0.10
- Assumed mean difference under alternative in regression coefficient priors: 0.04
- We vary the values of α_0 used in the prior model probabilities:
  - $\alpha_0 = 0$
  - $\alpha_0 = \pm 2$
  - $\alpha_0 = \pm 4$
  - $\alpha_0 = \pm 10$

Table A.8: Rejection rates and relative MSE (FELM as reference) for simulations with equal regional sample sizes. Compare BMA approach with $\alpha_0 \in \{0, \pm 2, \pm 4, \pm 10\}$ to FELM and BHM.

		——— Rejection Rates ———			— Relative MSE —	
Number of Null Regions	Method	Global Rejection Rate	TPR (Alternative† Regions)	FPR (Null Regions)	(Alternative† Regions)	(Null Regions)
0	FELM	0.895	0.301	—	1.00	—
	BHM	0.880	0.867	—	0.21	—
	BMA ($\alpha_0 = -10$)	0.891	0.881	—	0.19	—
	BMA ($\alpha_0 = -4$)	0.890	0.659	—	0.26	—
	BMA ($\alpha_0 = -2$)	0.889	0.513	—	0.35	—
	BMA ($\alpha_0 = 0$)	0.887	0.424	—	0.44	—
	BMA ($\alpha_0 = 2$)	0.886	0.379	—	0.51	—
	BMA ($\alpha_0 = 4$)	0.885	0.352	—	0.57	—
	BMA ($\alpha_0 = 10$)	0.882	0.308	—	0.71	—
1	FELM	0.731	0.300	0.030	1.00	1.00
	BHM	0.659	0.689	0.510	0.29	1.29
	BMA ($\alpha_0 = -10$)	0.723	0.705	0.644	0.29	1.39
	BMA ($\alpha_0 = -4$)	0.722	0.494	0.199	0.36	1.02
	BMA ($\alpha_0 = -2$)	0.720	0.396	0.097	0.44	0.86
	BMA ($\alpha_0 = 0$)	0.718	0.347	0.059	0.51	0.80
	BMA ($\alpha_0 = 2$)	0.716	0.325	0.047	0.57	0.78
	BMA ($\alpha_0 = 4$)	0.714	0.311	0.039	0.62	0.78
	BMA ($\alpha_0 = 10$)	0.710	0.289	0.030	0.73	0.80
2	FELM	0.491	0.304	0.027	1.00	1.00
	BHM	0.396	0.486	0.313	0.49	0.79
	BMA ($\alpha_0 = -10$)	0.484	0.470	0.404	0.54	0.86
	BMA ($\alpha_0 = -4$)	0.480	0.348	0.109	0.52	0.70
	BMA ($\alpha_0 = -2$)	0.477	0.300	0.056	0.55	0.67
	BMA ($\alpha_0 = 0$)	0.474	0.283	0.038	0.59	0.67
	BMA ($\alpha_0 = 2$)	0.472	0.278	0.032	0.63	0.69
	BMA ($\alpha_0 = 4$)	0.469	0.276	0.029	0.67	0.72
	BMA ($\alpha_0 = 10$)	0.464	0.275	0.024	0.76	0.78
3	FELM	0.259	0.307	0.027	1.00	1.00
	BHM	0.200	0.287	0.170	0.80	0.50
	BMA ($\alpha_0 = -10$)	0.248	0.247	0.209	0.92	0.50
	BMA ($\alpha_0 = -4$)	0.246	0.223	0.059	0.72	0.50
	BMA ($\alpha_0 = -2$)	0.243	0.220	0.033	0.68	0.54
	BMA ($\alpha_0 = 0$)	0.241	0.227	0.026	0.68	0.58
	BMA ($\alpha_0 = 2$)	0.239	0.236	0.024	0.70	0.62
	BMA ($\alpha_0 = 4$)	0.237	0.245	0.023	0.73	0.66
	BMA ($\alpha_0 = 10$)	0.233	0.262	0.022	0.79	0.75
4	FELM	0.092	0.298	0.027	1.00	1.00
	BHM	0.075	0.124	0.071	1.33	0.29
	BMA ($\alpha_0 = -10$)	0.087	0.088	0.078	1.51	0.27
	BMA ($\alpha_0 = -4$)	0.086	0.112	0.027	1.09	0.36
	BMA ($\alpha_0 = -2$)	0.084	0.138	0.018	0.92	0.44
	BMA ($\alpha_0 = 0$)	0.083	0.163	0.016	0.84	0.51
	BMA ($\alpha_0 = 2$)	0.082	0.184	0.017	0.82	0.57
	BMA ($\alpha_0 = 4$)	0.082	0.201	0.017	0.82	0.62
	BMA ($\alpha_0 = 10$)	0.081	0.235	0.019	0.83	0.73
5	FELM	0.028	—	0.027	—	1.00
	BHM	0.025	—	0.025	—	0.20
	BMA ($\alpha_0 = -10$)	0.026	—	0.024	—	0.19
	BMA ($\alpha_0 = -4$)	0.025	—	0.010	—	0.26
	BMA ($\alpha_0 = -2$)	0.025	—	0.009	—	0.35
	BMA ($\alpha_0 = 0$)	0.024	—	0.009	—	0.44
	BMA ($\alpha_0 = 2$)	0.023	—	0.011	—	0.51
	BMA ($\alpha_0 = 4$)	0.023	—	0.012	—	0.57
	BMA ($\alpha_0 = 10$)	0.022	—	0.016	—	0.70

† Treatment effect for alternative regions is 0.034

APPENDIX B: ADDITIONAL RESULTS FOR CHAPTER 4

B.1 Additional Details of the Laplace Approximation

Let $h_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell) = \log\{p(\boldsymbol{\theta}_\ell|\mathbf{D}, M_\ell)\}$. It can be shown that

$$
\begin{aligned}
h'_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell) &= \frac{dh_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell)}{d\boldsymbol{\theta}_\ell} \\
&= -\boldsymbol{\Sigma}_{0\ell}^{-1}(\boldsymbol{\theta}_\ell - \boldsymbol{\mu}_{0\ell}) + \sum_{d=1}^{D_\ell}\sum_{i\in\Omega_{\ell,d}}\sum_{k=1}^{K}\sum_{j=1}^{n_i}(\delta_{ijk}\nu_{ij}\boldsymbol{w}_{\ell ij}) \\
&\quad - \sum_{d=1}^{D_\ell}\sum_{i\in\Omega_{\ell,d}}\sum_{k=1}^{K}\left[\tilde{\eta}_{ik}\left\{\phi_{ik} + \sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\right\}^{-1} \times \left\{\sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\boldsymbol{w}_{\ell ij}\right\}\right],
\end{aligned}
$$

where $\hat{c}_{ijk} = \delta_{ijk}(y_{ij} - m_{k-1}) + \sum_{g=k+1}^{K}\delta_{ijg}(m_k - m_{k-1})$. Let $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_\ell} = -\left\{h''_{\boldsymbol{\theta}_\ell}(\hat{\boldsymbol{\theta}}_\ell)\right\}^{-1}$ where $\hat{\boldsymbol{\theta}}_\ell$ is the posterior mode of $\boldsymbol{\theta}_\ell$ and

$$
\begin{aligned}
h''_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell) &= \frac{d^2h_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_\ell)}{d\boldsymbol{\theta}_\ell d\boldsymbol{\theta}'_\ell} \\
&= -\boldsymbol{\Sigma}_{0\ell}^{-1} - \sum_{d=1}^{D_\ell}\sum_{i\in\Omega_{\ell,d}}\sum_{k=1}^{K}\tilde{\eta}_{ik}\left[\left\{\sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\boldsymbol{w}_{\ell ij}\boldsymbol{w}'_{\ell ij}\right\}\left\{\phi_{ik} + \sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\right\}^{-1}\right. \\
&\quad \left. - \left\{\sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\boldsymbol{w}_{\ell ij}\right\}\left\{\sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\boldsymbol{w}_{\ell ij}\right\}'\left\{\phi_{ik} + \sum_{j=1}^{n_i}\hat{c}_{ijk}\exp(\boldsymbol{w}'_{\ell ij}\boldsymbol{\theta}_\ell)\right\}^{-2}\right].
\end{aligned}
$$

B.2 Comparison Models in Simulation Studies

B.2.1 Cox Proportional Hazards Models

We consider two Cox proportional hazards models (CPHMs): one to estimate the global treatment effect γ_G and one to estimate the region-specific treatment effects $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S)$. For the first CPHM, we define λ_0 to be the baseline hazard where $\lambda_0 \sim \text{Gamma}(\eta_{\lambda 0}, \phi_{\lambda 0})$. Additionally, we define $\boldsymbol{\theta}_1^* = (\gamma_G, \boldsymbol{\beta}')'$ where $\boldsymbol{\beta}$ is a $p \times 1$ vector of covariate effects, and we assume $\boldsymbol{\theta}_1^* \sim N_{(p+1)}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Let $\boldsymbol{w}_{ij}$ be the $(p+1) \times 1$ vector corresponding to $\boldsymbol{\theta}_1^*$ for the jth subject in the ith region where the first element is a treatment indicator (1 for treatment and 0 for control) and the last p elements are optional covariates. Lastly, we define $\boldsymbol{\xi}_1^* = \{\lambda_0, \boldsymbol{\theta}_1^*\}$.

The likelihood for the first CPHM can be written as

$$\mathcal{L}(\boldsymbol{\xi}_1^* | \mathbf{D}) = \prod_{i=1}^{S} \prod_{j=1}^{n_i} \left[\left\{ \lambda_0 \exp\left(\boldsymbol{w}_{ij}' \boldsymbol{\theta}_1^* \right) \right\}^{\nu_{ij}} \times \exp\left\{ -\lambda_0 y_{ij} \exp\left(\boldsymbol{w}_{ij}' \boldsymbol{\theta}_1^* \right) \right\} \right],$$

where y_{ij} and ν_{ij} are as defined in Section 4.3.1. The full conditional distribution of $\lambda_0 | \boldsymbol{\theta}_1^*, \mathbf{D}$ is $\lambda_0 | \boldsymbol{\theta}_1^*, \mathbf{D} \sim \text{Gamma}(\eta_0^*, \phi_0^*)$, where

$$\eta_0^* = \sum_{i=1}^{S} \sum_{j=1}^{n_i} \nu_{ij} + \eta_{\lambda 0},$$

$$\phi_0^* = \phi_{\lambda 0} + \sum_{i=1}^{S} \sum_{j=1}^{n_i} y_{ij} \exp\left(\boldsymbol{w}_{ij}' \boldsymbol{\theta}_1^* \right).$$

The marginal posterior distribution of $\boldsymbol{\theta}_1^* | \mathbf{D}$ is

$$p(\boldsymbol{\theta}_1^* | \mathbf{D}) \propto \exp\left\{ \sum_{i=1}^{S} \sum_{j=1}^{n_i} \nu_{ij} \boldsymbol{w}_{ij}' \boldsymbol{\theta}_1^* \right\} \times \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\theta}_1^* - \boldsymbol{\mu}_1 \right)' \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\theta}_1^* - \boldsymbol{\mu}_1 \right) \right\} \times \left(\phi_0^* \right)^{-\eta_0^*}.$$

For the second CPHM, we define $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)$ to be the region-specific baseline hazards, where $\lambda_i \sim \text{Gamma}(\eta_{\lambda i}, \phi_{\lambda i})$, $i = 1, \ldots, S$. We define $\boldsymbol{\theta}_2^* = (\boldsymbol{\gamma}, \boldsymbol{\beta}')'$ and $\boldsymbol{\xi}_2^* = \{\boldsymbol{\lambda}, \boldsymbol{\theta}_2^*\}$, and

we now let $\boldsymbol{w}_{ij}$ be the $(S + p) \times 1$ vector corresponding to $\boldsymbol{\theta}_2^*$ where the first S elements are region-specific treatment indicators.

The likelihood for the second CPHM can be written as

$$\mathcal{L}(\boldsymbol{\xi}_2^*|\mathbf{D}) = \prod_{i=1}^{S}\prod_{j=1}^{n_i}\left[\left\{\lambda_i \exp\left(\boldsymbol{w}_{ij}'\boldsymbol{\theta}_2^*\right)\right\}^{\nu_{ij}} \times \exp\left\{-\lambda_i y_{ij}\exp\left(\boldsymbol{w}_{ij}'\boldsymbol{\theta}_2^*\right)\right\}\right].$$

The full conditional distribution of $\lambda_i|\boldsymbol{\theta}_2^*, \mathbf{D}$ is $\lambda_i|\boldsymbol{\theta}_2^*, \mathbf{D} \sim \text{Gamma}(\eta_i^*, \phi_i^*)$, $i = 1, \ldots, S$, where

$$\eta_i^* = \sum_{i=1}^{S}\sum_{j=1}^{n_i}\nu_{ij} + \eta_{\lambda i},$$

$$\phi_i^* = \phi_{\lambda i} + \sum_{i=1}^{S}\sum_{j=1}^{n_i} y_{ij}\exp\left(\boldsymbol{w}_{ij}'\boldsymbol{\theta}_2^*\right).$$

If we assume $\boldsymbol{\theta}_2^* \sim N_{(S+p)}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, then the marginal posterior distribution of $\boldsymbol{\theta}_2^*|\mathbf{D}$ is

$$p(\boldsymbol{\theta}_2^*|\mathbf{D}) \propto \exp\left\{\sum_{i=1}^{S}\sum_{j=1}^{n_i}\nu_{ij}\boldsymbol{w}_{ij}'\boldsymbol{\theta}_2^*\right\} \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}_2^* - \boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\theta}_2^* - \boldsymbol{\mu}_2\right)\right\} \times \prod_{i=1}^{S}\left(\phi_i^*\right)^{-\eta_i^*}.$$

We elicit the prior distributions as $\boldsymbol{\theta}_1^* \sim N_{(p+1)}(\mathbf{0}, 10{,}000\boldsymbol{I}_{p+1})$ and $\lambda_0 \sim \text{Gamma}(0.01, 0.01)$ for the first CPHM, and $\boldsymbol{\theta}_2^* \sim N_{(S+p)}(\mathbf{0}, 10{,}000\boldsymbol{I}_{S+p})$ and $\lambda_i \sim \text{Gamma}(0.01, 0.01)$, $i = \ldots, S$, for the second CPHM, where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. To test the global null hypothesis using simulation studies discussed in Section 4.3.5, we calculate the upper limit of a 95% credible interval for $\gamma_G|\mathbf{D}$ using the first CPHM, and we reject the null hypothesis if γ_0 is greater than this limit. Similarly, we reject the null hypothesis for each region-specific treatment effect using the second CPHM if the upper boundaries of the corresponding 95% credible intervals are below γ_0. For all simulation studies presented in Chapter 4 and Appendix B, we set $\gamma_0 = 0$.

B.2.2 Bayesian Hierarchical Model

We now consider a Bayesian hierarchical model (BHM), and we use the same model formulation and prior elicitation used by the FDA in a post hoc analysis of the LEADER trial data

(Rothmann, 2021). Let y_i^* denote the observed subgroup log hazard ratio for the ith region, $i = 1, \ldots, S$. We assume the following:

$$y_i^* \sim N(\mu_i, \sigma_i^2),$$

$$\mu_i \sim N(\mu, 1/\tau),$$

$$\mu \sim N(0, 16),$$

$$\tau \sim \text{Gamma}(0.001, 0.001).$$

For the ith region, σ_i^2 is estimated as the asymptotic variance of the log hazard ratio; i.e., $\sigma_i^2 = \frac{1}{\eta_{0i}} + \frac{1}{\eta_{1i}}$, where η_{0i} and η_{1i} correspond to the number of failures in the control group and the experimental treatment group, respectively (Lininger *et al.*, 1979). In Section B.5, we discuss alternative priors on the hierarchical standard deviation $\tau^{-\frac{1}{2}}$ as recommended by Gelman (2006).

We test the global treatment effect by calculating the 95% credible interval for $\mu|\mathbf{D}$, and we reject the null hypothesis for the global effect if the upper limit is below γ_0. Similarly, we test the ith region-specific treatment effect by calculating the 95% credible interval for $\mu_i|\mathbf{D}$, $i = 1, \ldots, S$, and we reject the null hypothesis if the upper limit is below γ_0.

B.3 Details of Data Generation in Simulation Studies

For each simulation study, we generate datasets designed to mirror the LEADER trial data. Marso *et al.* (2016) state that $N = 9340$ patients from $S = 4$ regions underwent randomization from September 2010 through April 2012 (i.e., 18 months). Additionally, they report the following details for the LEADER trial:

- 608 out of 4668 patients in the liraglutide group and 694 out of 4672 patients in the placebo group experienced a primary composite outcome.
- 96.8% of the patients completed a final visit, died, or had a primary outcome.
- The median time of exposure to liraglutide or placebo was 3.5 years.

In the data simulation process, we set the constant baseline hazard equal to 0.0386 and the dropout rate to 0.0082 (i.e., we assume theoretical dropout times are randomly sampled from an exponential distribution with mean 0.0082^{-1}). These two numbers were chosen such that the average number of events in the liraglutide and placebo group, the average percentage of subjects who either completed follow-up or experienced an event, and the median follow-up time from thousands of simulated datasets were approximately equal to the reported information above.

B.4 Additional Results for Primary Simulation Studies

B.4.1 First Simulation Study: Equal Regional Sample Sizes

Simulation study details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Prior elicitation:
  - Each element of $\boldsymbol{\mu}_{0\ell}$ equal to $\log(1.3)$, $\ell = 1, \ldots, L$
  - Diagonals of $\boldsymbol{\Sigma}_{0\ell}$ equal to Diag$\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$
  - $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$, $i = 1, \ldots, S$, $k = 1, \ldots, K$
  - $p(M_\ell) \propto e^{D_\ell \times \alpha_0}$, where $\alpha_0 = 0$

Table B.1: Bias of region-specific treatment effects for simulation study with equal regional sample sizes.

Number of Null Regions	Method	Bias (Alternative† Regions)	Bias (Null Regions)
0	CPHM	1.47×10^{-3}	—
	BMA	4.17×10^{-4}	—
	BHM	2.28×10^{-3}	—
1	CPHM	1.04×10^{-3}	-7.35×10^{-4}
	BMA	2.19×10^{-2}	-6.25×10^{-2}
	BHM	1.98×10^{-2}	-5.11×10^{-2}
2	CPHM	5.65×10^{-4}	1.08×10^{-4}
	BMA	4.19×10^{-2}	-3.98×10^{-2}
	BHM	3.66×10^{-2}	-3.28×10^{-2}
3	CPHM	3.02×10^{-4}	4.72×10^{-5}
	BMA	6.51×10^{-2}	-2.08×10^{-2}
	BHM	5.54×10^{-2}	-1.68×10^{-2}
4	CPHM	—	1.81×10^{-4}
	BMA	—	-2.56×10^{-5}
	BHM	—	2.76×10^{-4}

† Treatment hazard ratio for alternative regions is 0.868

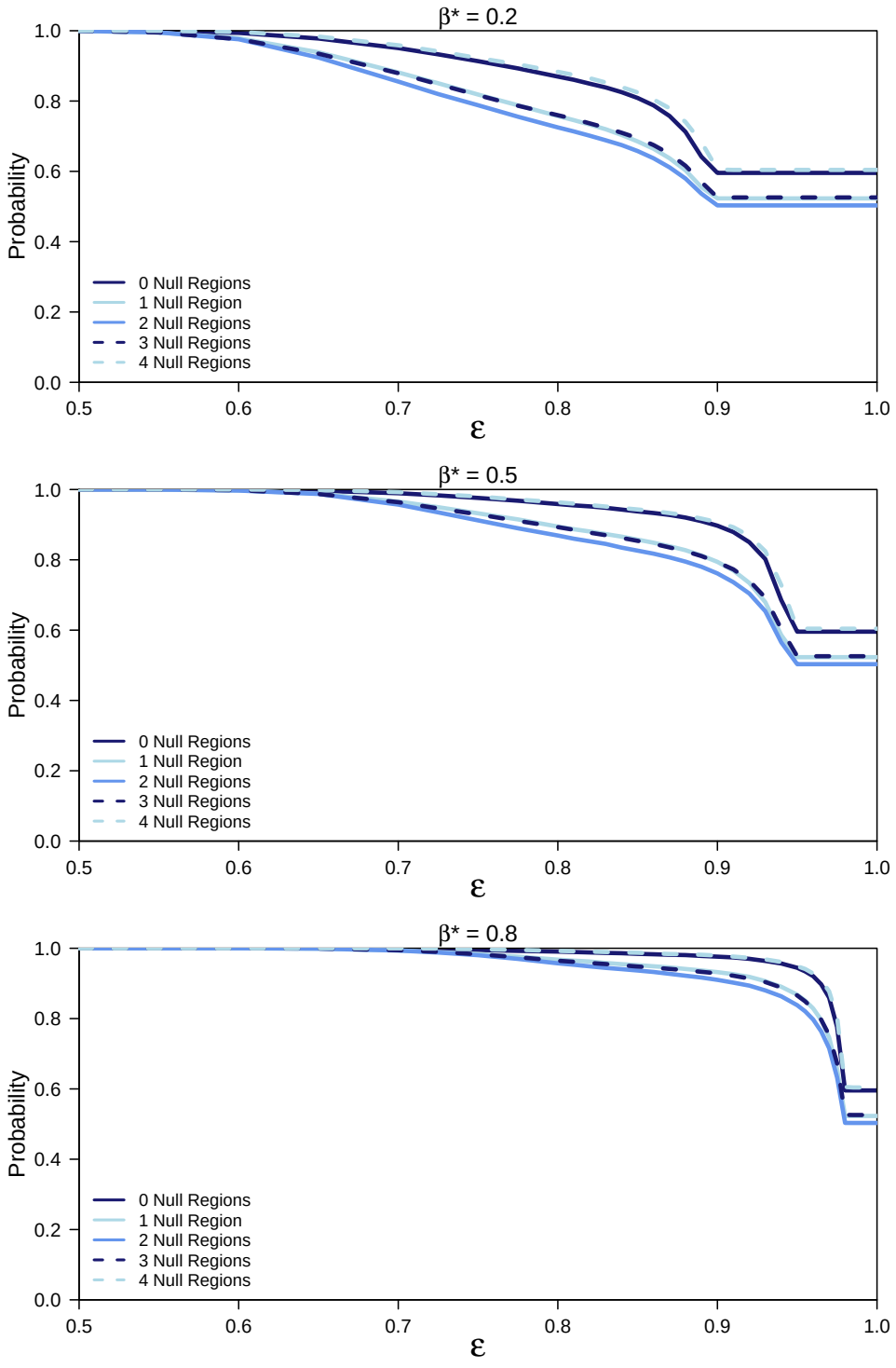

Figure B.1: Global consistency probabilities for varying values of the minimal clinically important regional difference ε and for $\beta^* \in \{0.2, 0.5, 0.8\}$.

B.4.2 Second Simulation Study: Unequal Regional Sample Sizes

Simulation study details:

- Regional sample sizes equal to original sample sizes in LEADER trial: sample sizes of 711, 3296, 2847, and 2486 for Asia, Europe, North America, and Rest of the World, respectively.
- Scenario 1: underlying treatment-to-placebo hazard ratio for all regions equal to 0.868 (i.e., treatment effect of -0.142)
- Scenario 2: underlying treatment-to-placebo hazard ratios of 0.62, 0.82, 1.01, and 0.83 for Asia, Europe, North America, and Rest of the World, respectively.
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Prior elicitation:
  - Each element of $\boldsymbol{\mu}_{0\ell}$ equal to $\log(1.3)$, $\ell = 1, \ldots, L$
  - Diagonals of $\boldsymbol{\Sigma}_{0\ell}$ equal to $\text{Diag}\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$
  - $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$, $i = 1, \ldots, S$, $k = 1, \ldots, K$
  - $p(M_\ell) \propto e^{D_\ell \times \alpha_0}$, where $\alpha_0 = 0$

Table B.2: Bias of region-specific treatment effects for simulation study with differing regional sample sizes and equal treatment-to-placebo hazard ratios.

Method	Bias			
	Asia ($n_1 = 711$, $\gamma_1 = -0.142$)	Europe ($n_2 = 3296$, $\gamma_2 = -0.142$)	North America ($n_3 = 28470$, $\gamma_3 = 0.142$)	Rest of the World ($n_4 = 248$, ($\gamma_4 = -0.142$)
CPHM	-2.10×10^{-3}	1.32×10^{-3}	1.78×10^{-3}	-3.60×10^{-4}
BMA	-1.09×10^{-3}	-2.34×10^{-4}	1.00×10^{-4}	-7.81×10^{-4}
BHM	1.28×10^{-3}	1.65×10^{-3}	2.06×10^{-3}	8.98×10^{-4}

Table B.3: Bias of region-specific treatment effects for simulation study with differing regional sample sizes and differing treatment-to-placebo hazard ratios.

Method	Bias			
	Asia ($n_1 = 711$, $\gamma_1 = -0.478$)	Europe ($n_2 = 3296$, $\gamma_2 = -0.198$)	North America ($n_3 = 28470$, $\gamma_3 = 0.010$)	Rest of the World ($n_4 = 248$, ($\gamma_4 = -0.186$)
CPHM	-5.65×10^{-3}	2.60×10^{-3}	-6.49×10^{-4}	1.15×10^{-3}
BMA	1.63×10^{-1}	2.30×10^{-2}	-7.10×10^{-2}	1.77×10^{-2}
BHM	1.88×10^{-1}	1.57×10^{-2}	-5.63×10^{-2}	1.14×10^{-2}

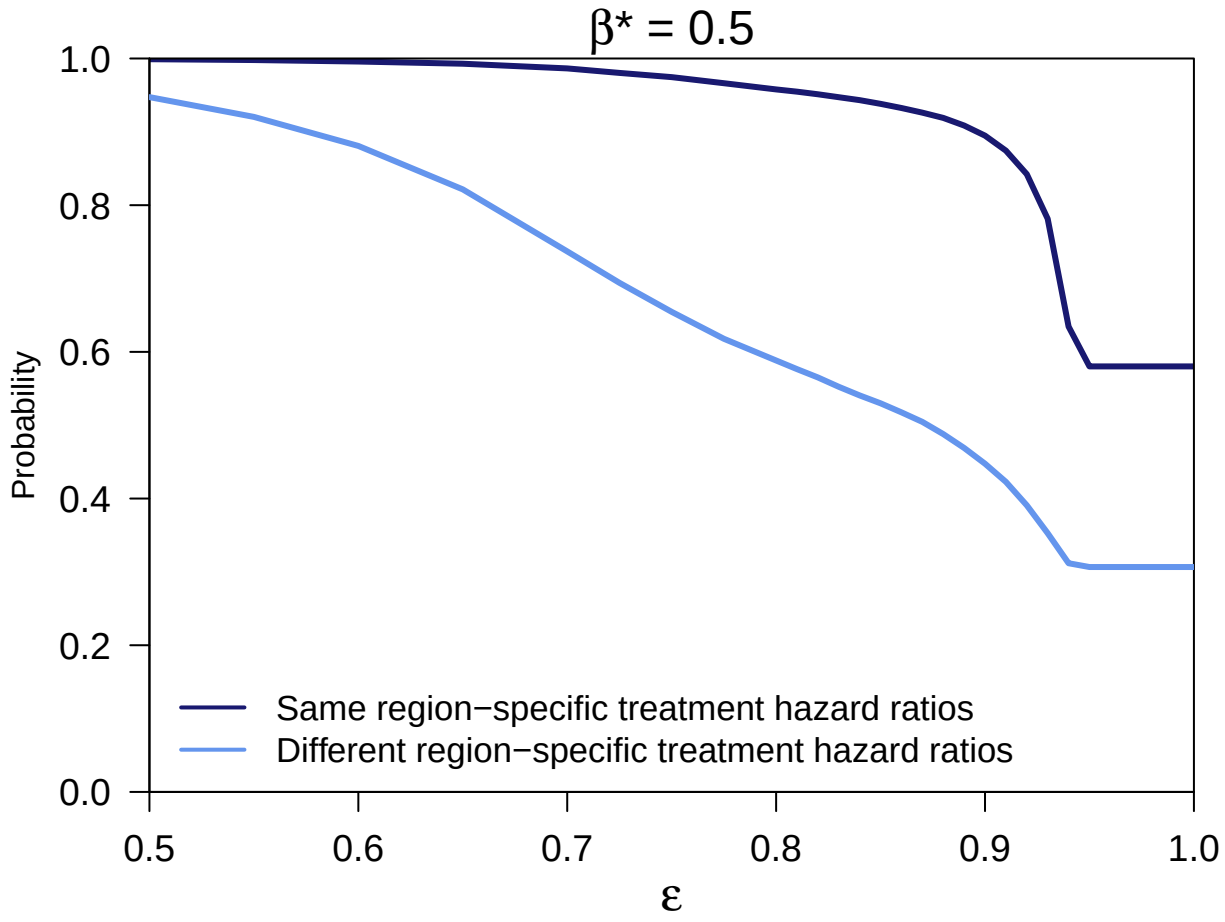

Figure B.2: Global consistency probabilities for varying values of the minimal clinically important regional difference ε and for $\beta^* = 0.5$.

B.5 Additional Simulation Studies

B.5.1 Sample Sizes of Null Regions Half the Size of Alternative Regions

Simulation study details:

- Sample sizes of null regions half the size of alternative regions
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Same prior elicitation discussed in Section 4.4.1

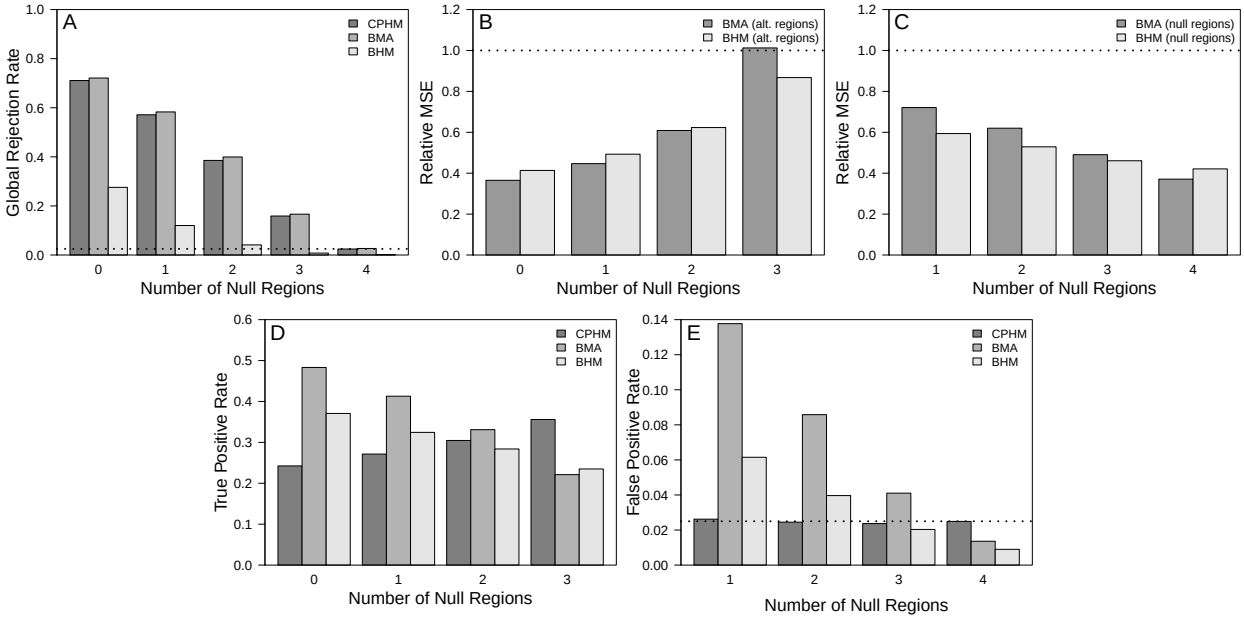

Figure B.3: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study where sample sizes of null regions are half the size of alternative regions.

B.5.2 Sample Sizes of Null Regions Double the Size of Alternative Regions

Simulation study details:

- Sample sizes of null regions double the size of alternative regions
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Same prior elicitation discussed in Section 4.4.1

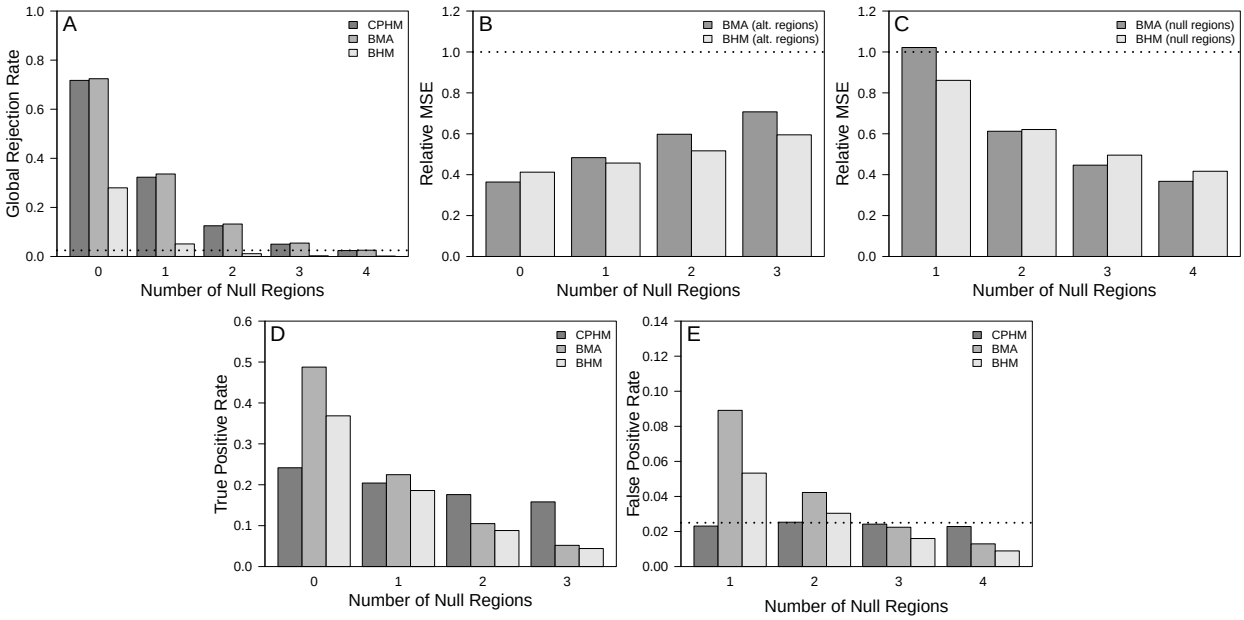

Figure B.4: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study where sample sizes of null regions are double the size of alternative regions.

B.5.3 Non-Constant Baseline Hazard

Simulation study details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying piecewise baseline hazard of 0.02, 0.035, and 0.055 over the intervals (0, 2], (2, 3.75], and (3.75, 5], respectively
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Same prior elicitation discussed in Section 4.4.1

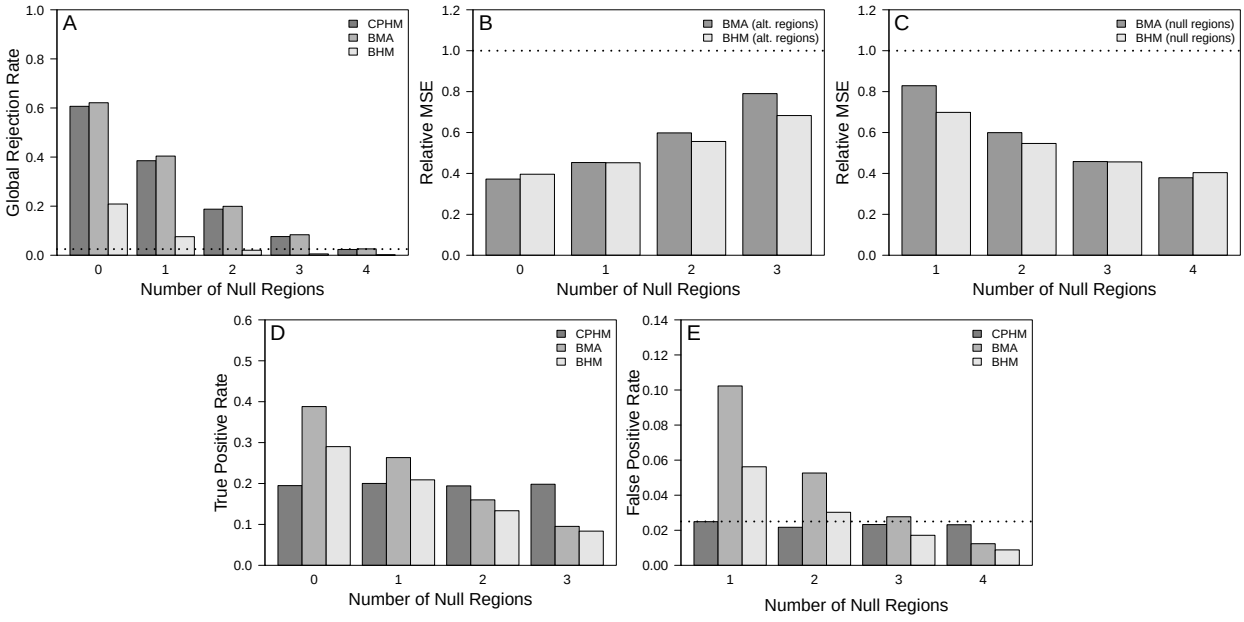

Figure B.5: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study with a non-constant baseline hazard.

B.5.4 Comparison of BHMs with Different Priors on Hierarchical Parameters

We investigated the prior choice on the hierarchical precision/standard deviation parameter by comparing the version of the BHM used in the FDA's analysis which had a Gamma$(0.001, 0.001)$ prior on the hierarchical precision τ (Rothmann, 2021) to two additional BHMs, one with a Uniform$(0, 100)$ prior on the hierarchical standard deviation $\tau^{-\frac{1}{2}}$ and one with a Half-Cauchy$(\text{df} = 1, \text{peak} = 0, \text{scale} = 10)$ prior on $\tau^{-\frac{1}{2}}$. The priors on $\tau^{-\frac{1}{2}}$ for these additional BHMs were chosen according to the recommendations of Gelman (2006), and the priors for all other parameters were the same across the BHMs (see Section B.2.2 for details). To compare these models, we repeated the first two simulation studies.

For both simulation studies, all three BHMs were fit using the `rjags` package in R (Plummer *et al.*, 2022) with four chains, 1000 burn-in iterations per chain, and $100{,}000$ post-burn-in iterations per chain that were thinned by every fifth iteration. We assessed the convergence of all parameters using the potential scale reduction factor (Gelman and Rubin, 1992) and the multivariate potential scale reduction factor (Brooks and Gelman, 1998).

B.5.4.1 First Simulation Study: Equal Regional Sample Sizes

Across all five scenarios, the average potential scale reduction factor for the hierarchical variance τ^{-1} was approximately 1.09 for the BHM with the gamma prior (BHM-G), 1.12 for the BHM with the uniform prior (BHM-U), and 1.05 for the BHM with the half-Cauchy prior (BHM-HC). The potential scale reduction factors for all other parameters were approximately 1.00 with each model. With each simulated datasets across the scenarios, the multivariate potential scale reduction factor ranged between 1.000 and 1.001 for the BHM-G, 1.000 and 1.045 for the BHM-U, and 1.000 and 1.006 for the BHM-HC, indicating good convergence.

The results from the first repeated simulation study are shown in Figure B.6. The BHM-HC resulted in global rejection rates that were marginally lower than the BHM-G, while the BHM-U resulted in drastically lower global rejection rates than the other two BHMs. All three BHMs showed similar patterns in MSE across scenarios compared to the BMA approach. As discussed by Gelman (2006), the uniform prior on $\tau^{-\frac{1}{2}}$ supports larger values of the standard deviation and thus less information borrowing when the number of groups is small, as is often the case with MRCTs.

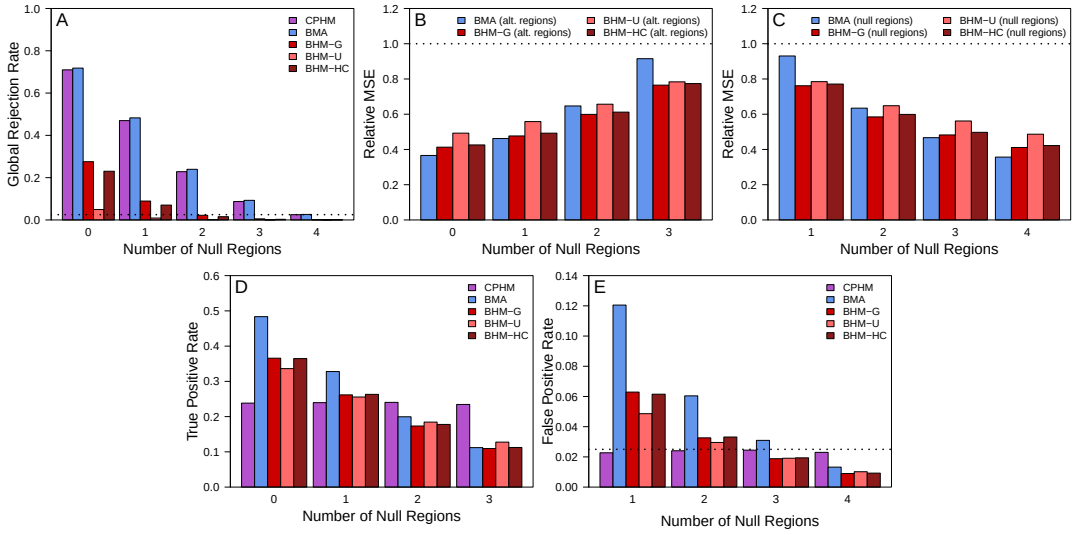

Figure B.6: Comparison of BHMs with respect to global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for the simulation study with equal regional sample sizes.

B.5.4.2 Second Simulation Study: Unequal Regional Sample Sizes

Across the two scenarios, the average potential scale reduction factor for the hierarchical variance τ^{-1} was approximately 1.08 for the BHM-G, 1.12 for the BHM-U, and 1.05 for the BHM-HC. The potential scale reduction factors for all other parameters were approximately 1.00 with each model. With each simulated datasets across the scenarios, the multivariate potential scale reduction factor ranged between 1.000 and 1.001 for the BHM-G, 1.000 and 1.035 for the BHM-U, and 1.000 and 1.004 for the BHM-HC, indicating good convergence.

The results from the second repeated simulation study are shown in Figure B.7. Similar to the first repeated simulation study, the BHM-HC resulted in global rejection rates that were slightly lower than the BHM-G, while the BHM-U again resulted in drastically lower global rejection rates than the other two BHMs. All three BHMs showed similar patterns in MSE for all regions in both scenarios compared to the BMA approach.

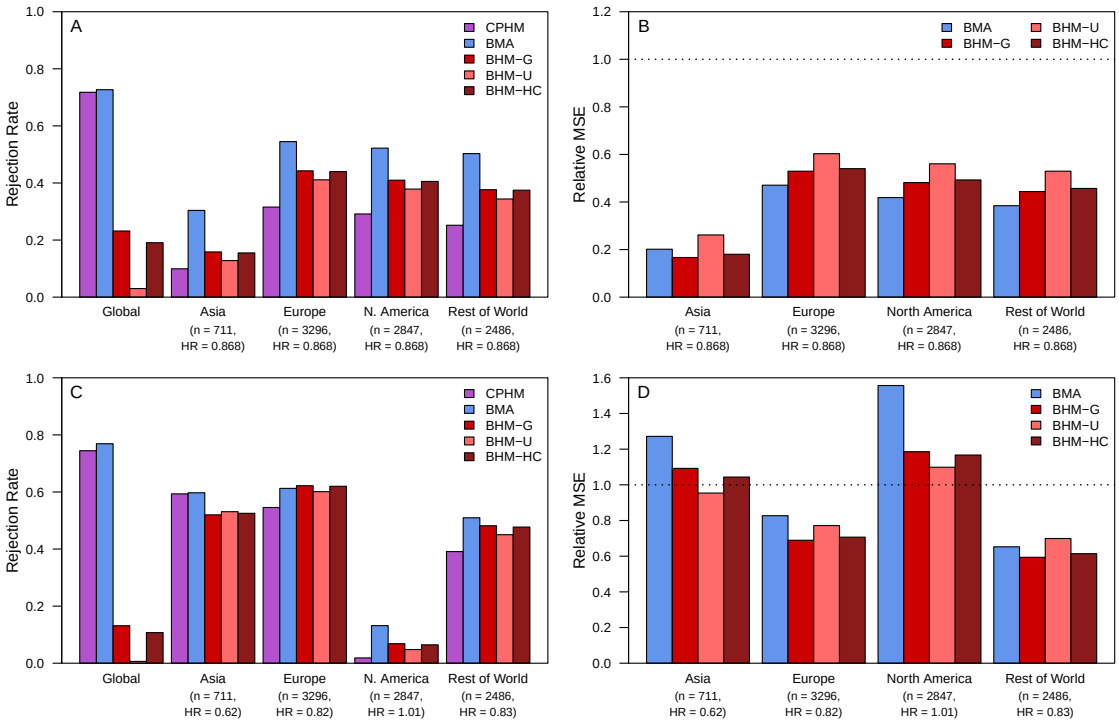

Figure B.7: Comparison of BHMs with respect to rejection rates (*Panel A*) and MSE relative to CPHM (*Panel B*) for the scenario with all treatment-to-placebo hazard ratios equal to 0.868, and rejection rates (*Panel C*) and relative MSE (*Panel D*) for the scenario with differing treatment-to-placebo hazard ratios.

B.6 Sensitivity Analyses for Simulation Studies

B.6.1 Change Number of Intervals K

Simulation study details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K \in \{4, 8, 12, 16\}$
- Prior elicitation:
  - Each element of $\boldsymbol{\mu}_{0\ell}$ equal to $\log(1.3)$, $\ell = 1, \ldots, L$
  - Diagonals of $\boldsymbol{\Sigma}_{0\ell}$ equal to Diag$\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$
  - $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$, $i = 1, \ldots, S$, $k = 1, \ldots, K$
  - $p(M_\ell) \propto e^{D_\ell \times \alpha_0}$, where $\alpha_0 = 0$

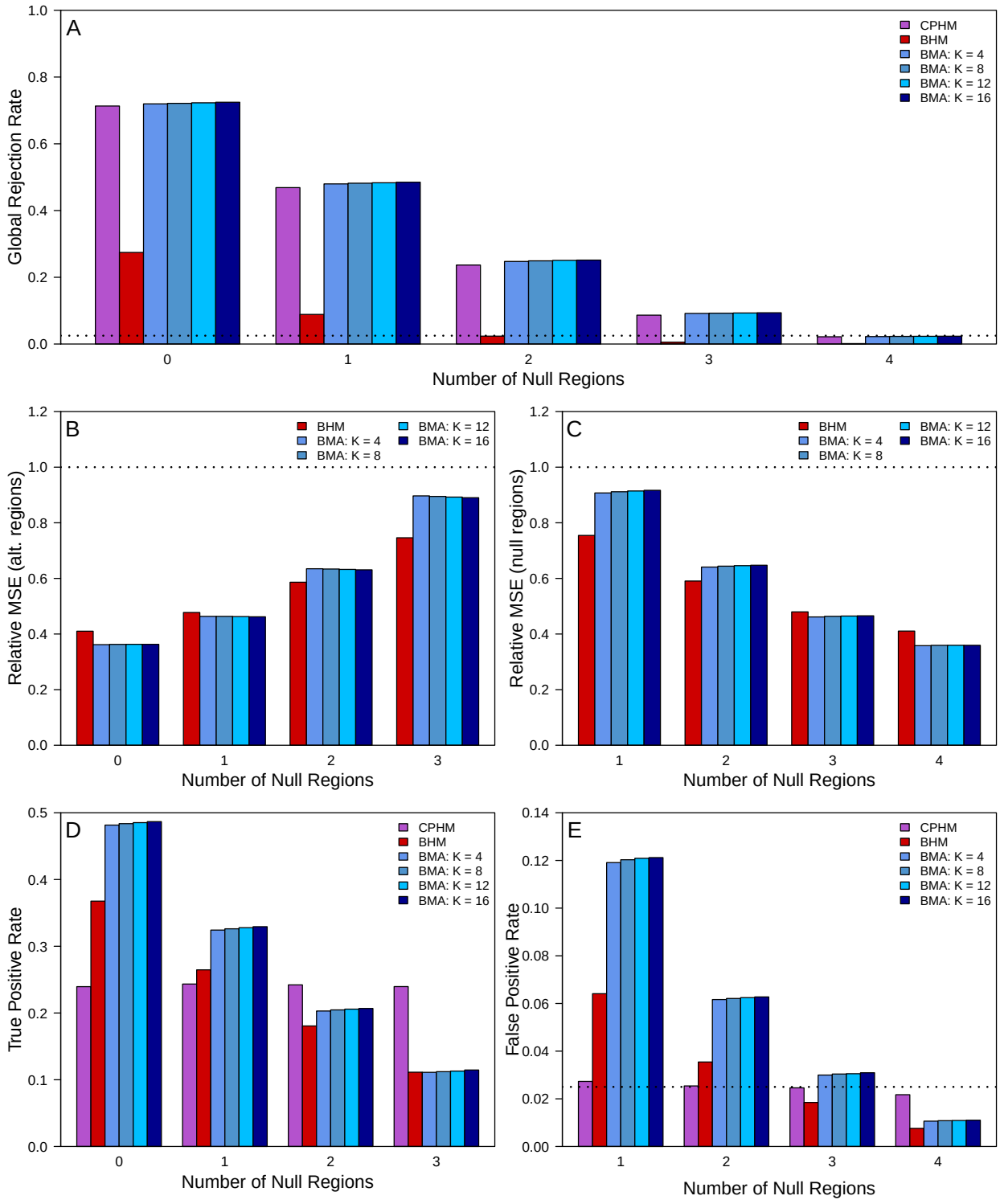

Figure B.8: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for sensitivity analysis with varying values of K.

B.6.2 Change Elicitation of Prior Distributions on Regression Effects

Simulation study details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Prior elicitation:
  - Each element of $\boldsymbol{\mu}_{0\ell}$ chosen to be equal, where $e^{\boldsymbol{\mu}_{0\ell}} \in \{0.7, 1.05, 1.3, 1.5\}$, $\ell = 1, \ldots, L$
  - Diagonals of $\boldsymbol{\Sigma}_{0\ell}$ equal to Diag$\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$
  - $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$, $i = 1, \ldots, S$, $k = 1, \ldots, K$
  - $p(M_\ell) \propto e^{D_\ell \times \alpha_0}$, where $\alpha_0 = 0$

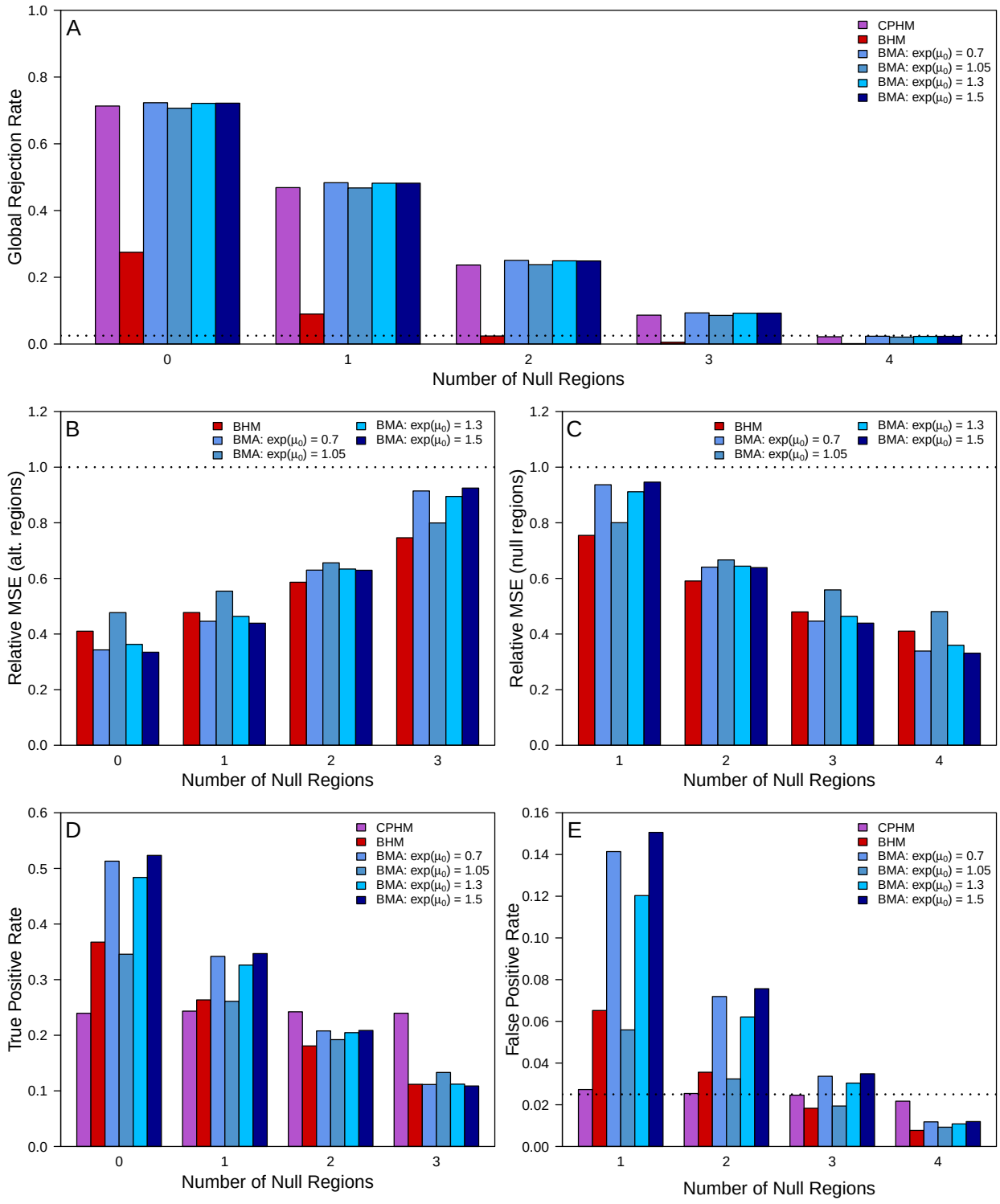

Figure B.9: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for sensitivity analysis with varying values of $\boldsymbol{\mu}_{0\ell}$.

B.6.3 Change Elicitation of α_0

Simulation study details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Underlying baseline hazard of 0.0386
- Number of time intervals in piecewise constant baseline hazard: $K = 8$
- Prior elicitation:
  - Each element of $\boldsymbol{\mu}_{0\ell}$ equal to $\log(1.3)$, $\ell = 1, \ldots, L$
  - Diagonals of $\boldsymbol{\Sigma}_{0\ell}$ equal to Diag$\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$
  - $\eta_{ik} = 0.01$ and $\phi_{ik} = 0.01$, $i = 1, \ldots, S$, $k = 1, \ldots, K$
  - $p(M_\ell) \propto e^{D_\ell \times \alpha_0}$, where $\alpha_0 \in \{0, \pm 0.5, \pm 1, \pm 2, \pm 5\}$

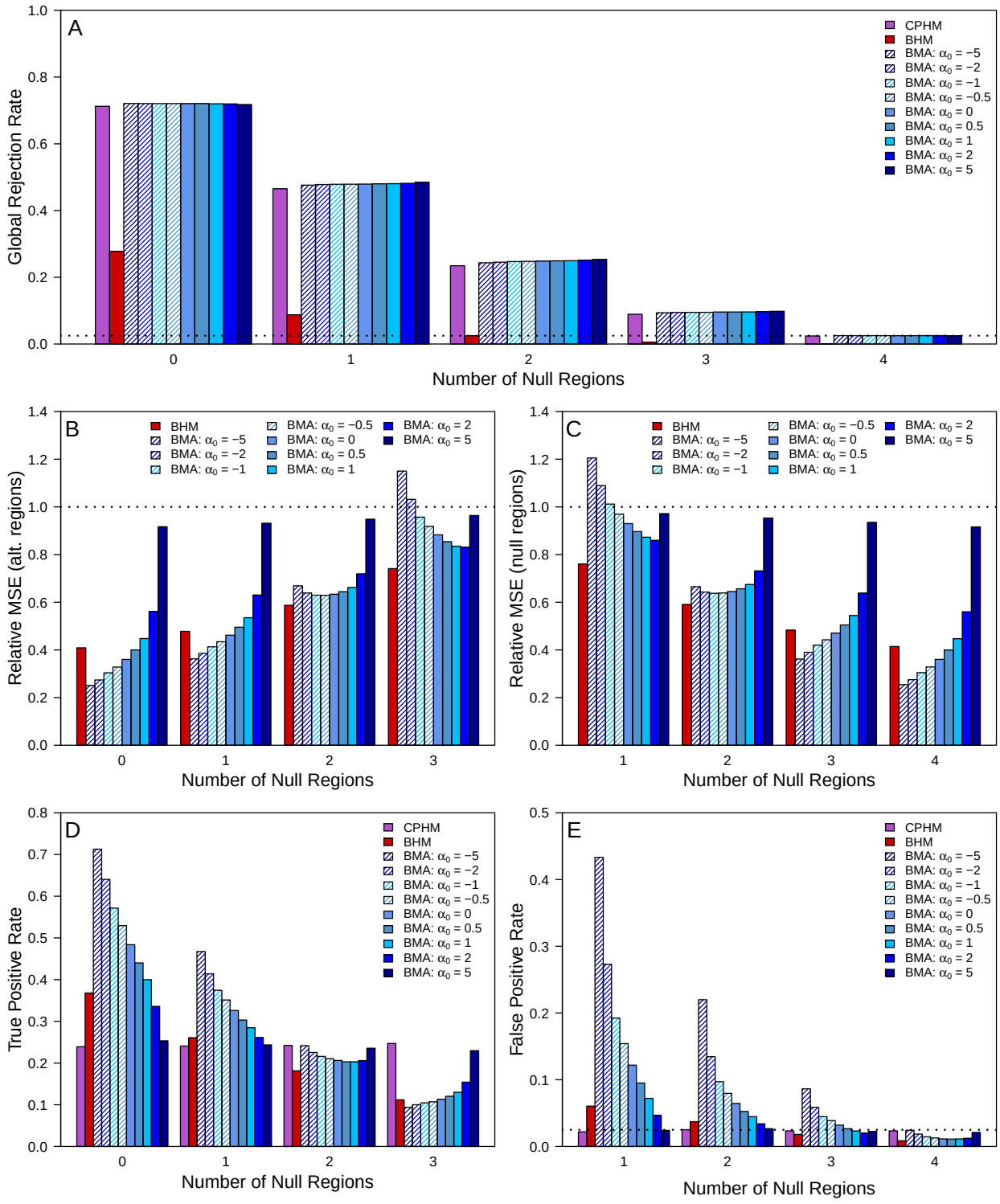

Figure B.10: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for sensitivity analysis with varying values of α_0.

B.7 Additional Information on the BHM Parameters in the LEADER Trial Data Analysis

To better understand the fit of the BHM in the analysis of the LEADER trial data, we examine the convergence diagnostics and summary statistics of the hierarchical parameters. The BHM was fit using the `rjags` package in R (Plummer *et al.*, 2022) with four chains, 1000 burn-in iterations per chain, and 100,000 post-burn-in iterations per chain that were thinned by every fifth iteration. Posterior summary statistics and the potential scale reduction factors (PSRFs) (Gelman and Rubin, 1992) for the hierarchical mean μ and hierarchical precision τ are reported in Table B.4, and trace plots are displayed in Figures B.11 and B.12, respectively. The PSRFs for all model parameters and the multivariate PSRF (Brooks and Gelman, 1998) were each approximately 1.000, indicating good convergence.

Table B.4: Posterior summary statistics for the hierarchical parameters from the BHM.

Hierarchical Parameter	Posterior Mean	Posterior SD	95% Credible Interval	PSRF Point Est.	PSRF 95% C.I. Upper Bound
Mean μ	-0.157	0.120	(-0.399, 0.045)	1.000	1.000
Precision τ	259.8	432.8	(4.03, 1515.64)	1.000	1.000

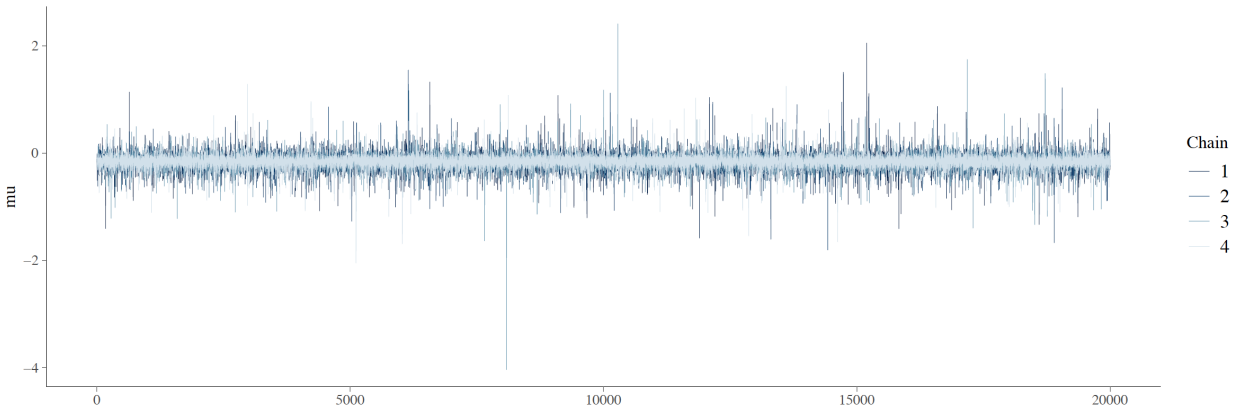

Figure B.11: Trace plot for μ from BHM in analysis of the LEADER trial data.

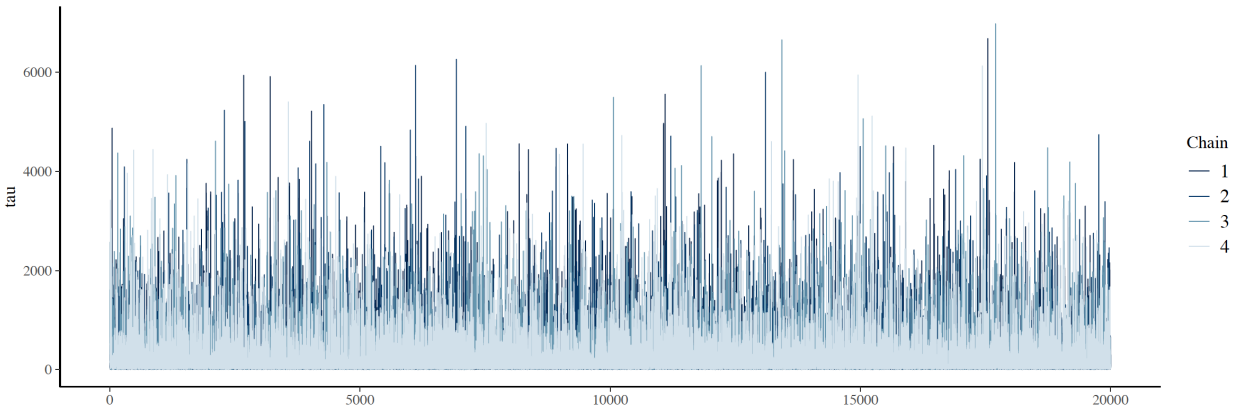

Figure B.12: Trace plot for τ from BHM in analysis of the LEADER trial data.

APPENDIX C: ADDITIONAL RESULTS FOR CHAPTER 5

C.1 Additional Details of the Maximization Algorithm

C.1.1 Full Conditional Distributions for Model $M_{\ell,\ell'}$

- Piecewise constant baseline hazards λ_{iq}, $i = 1, \ldots, S$, $q = 1, \ldots, Q$:

$$\lambda_{iq}|\boldsymbol{\alpha}, \boldsymbol{\theta}_{Y,\ell'}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'} \sim \text{Gamma}\left(\tilde{\eta}_{iq}, \tilde{\phi}_{\ell' iq}\right),$$

  where

$$\tilde{\eta}_{iq} = \eta_{iq} + \sum_{j=1}^{n_i} \delta_{ijq}\nu_{ij},$$

$$\tilde{\phi}_{\ell' iq} = \phi_{iq} + \sum_{j=1}^{n_i}\left[\left\{\delta_{ijq}(y_{ij} - m_{q-1}) + \sum_{g=q+1}^{Q}\delta_{ijg}(m_q - m_{q-1})\right\}\exp(\boldsymbol{\alpha}^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}_{Y,\ell' ij}^\intercal \boldsymbol{\theta}_{Y,\ell'})\right].$$

- Survival regression effects $\boldsymbol{\theta}_{Y,\ell'}$:

  Let $\Omega_{\ell',d}$ be the set of region labels for all regions that share the dth distinct treatment effect for some survival submodel ℓ', $d = 1, \ldots, D_{Y,\ell'}$.

$$p(\boldsymbol{\theta}_{Y,\ell'}|\boldsymbol{\alpha}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'}) \propto \left\{\prod_{d=1}^{D_{Y,\ell'}}\prod_{i\in\Omega_{\ell',d}}\prod_{q=1}^{Q}\tilde{\phi}_{\ell' iq}^{-\tilde{\eta}_{iq}}\exp\left(\sum_{j=1}^{n_i}\delta_{ijq}\nu_{ij}\boldsymbol{w}_{Y,\ell' ij}^\intercal\boldsymbol{\theta}_{Y,\ell'}\right)\right\} \times p(\boldsymbol{\theta}_{Y,\ell'}|M_{\ell,\ell'})$$

- Longitudinal regression effects $\boldsymbol{\theta}_{X,\ell}$:

  Let $\Omega_{\ell,d}$ be the set of region labels for all regions that share the dth distinct treatment effect for some longitudinal submodel ℓ, $d = 1, \ldots, D_{X,\ell}$, and let $\boldsymbol{W}_{X,\ell ij}$ be the subject-specific design matrix in the longitudinal submodel where each row corresponds to a different

longitudinal observation for the jth subject in the ith regions, $i = 1, \ldots, S$, $j = 1, \ldots, n_i$.

$$\boldsymbol{\theta}_{X,\ell} | \tau, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'} \sim N\left(\tilde{\boldsymbol{\mu}}_{X,\ell}, \tau^{-1}\tilde{\boldsymbol{\Sigma}}_{X,\ell}\right),$$

where

$$\tilde{\boldsymbol{\mu}}_{X,\ell} = \tilde{\boldsymbol{\Sigma}}_{X,\ell}\left\{\sum_{d=1}^{D_{X,\ell}}\sum_{i\in\Omega_{\ell,d}}\sum_{j=1}^{n_i}\boldsymbol{W}_{X,\ell ij}^{\intercal}\left(\boldsymbol{X}_{ij} - \boldsymbol{z}(t)^{\intercal}\boldsymbol{b}_{ij}\right) + \boldsymbol{\Sigma}_{X,\ell}^{-1}\boldsymbol{\mu}_{X,\ell}\right\},$$

$$\tilde{\boldsymbol{\Sigma}}_{X,\ell} = \left(\sum_{d=1}^{D_{X,\ell}}\sum_{i\in\Omega_{\ell,d}}\sum_{j=1}^{n_i}\boldsymbol{W}_{X,\ell ij}^{\intercal}\boldsymbol{W}_{X,\ell ij} + \boldsymbol{\Sigma}_{X,\ell}^{-1}\right)^{-1}.$$

- Precision of longitudinal likelihood τ:

$$\tau | \boldsymbol{\theta}_{X,\ell}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'} \sim \text{Gamma}\left(\tilde{\eta}_{\tau}, \tilde{\phi}_{\ell\tau}\right),$$

where

$$\tilde{\eta}_{\tau} = \frac{1}{2}\left\{\sum_{d=1}^{D_{X,\ell}}\sum_{i\in\Omega_{\ell,d}}\sum_{j=1}^{n_i}K_{ij} + \text{length}(\boldsymbol{\theta}_{X,\ell}) + \eta_{\tau}\right\},$$

$$\begin{aligned}\tilde{\phi}_{\ell\tau} =& \frac{1}{2}\left\{\sum_{d=1}^{D_{X,\ell}}\sum_{i\in\Omega_{\ell,d}}\sum_{j=1}^{n_i}\left(\boldsymbol{X}_{ij} - \boldsymbol{W}_{X,\ell ij}\boldsymbol{\theta}_{X,\ell} - \boldsymbol{z}(t)^{\intercal}\boldsymbol{b}_{ij}\right)^{\intercal}\left(\boldsymbol{X}_{ij} - \boldsymbol{W}_{X,\ell ij}\boldsymbol{\theta}_{X,\ell} - \boldsymbol{z}(t)^{\intercal}\boldsymbol{b}_{ij}\right)\right. \\ &\left. + \left(\boldsymbol{\theta}_{X,\ell} - \boldsymbol{\mu}_{X,\ell}\right)^{\intercal}\Sigma_{X,\ell}^{-1}\left(\boldsymbol{\theta}_{X,\ell} - \boldsymbol{\mu}_{X,\ell}\right) + \phi_{\tau}\right\}.\end{aligned}$$

- Association parameters $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha} | \boldsymbol{\theta}_{Y,\ell'}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'}) \propto \left\{\prod_{d=1}^{D_{Y,\ell'}}\prod_{i\in\Omega_{\ell',d}}\prod_{q=1}^{Q}\tilde{\phi}_{\ell' iq}^{-\tilde{\eta}_{iq}}\exp\left(\sum_{j=1}^{n_i}\delta_{ijq}\nu_{ij}\boldsymbol{\alpha}^{\intercal}\boldsymbol{b}_{ij}\right)\right\} \times p(\boldsymbol{\alpha} | M_{\ell,\ell'})$$

- Inverse of covariance matrix of random effects $\boldsymbol{G}^{-1}$:

$$\boldsymbol{G}^{-1}|\boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'} \sim \text{Wishart}_r \left(\nu_0 + N, \left(\boldsymbol{C}_0^{-1} + \sum_{i=1}^{S} \sum_{j=1}^{n_i} \boldsymbol{b}_{ij} \boldsymbol{b}_{ij}^{\intercal} \right)^{-1} \right)$$

- Subject-specific random effects $\boldsymbol{b}_{ij}$, $i = 1, \ldots, S$, $j = 1, \ldots, n_i$:

$$\begin{aligned} p(\boldsymbol{b}_{ij}|\boldsymbol{\alpha}, \boldsymbol{\theta}_{Y,\ell'}, \boldsymbol{\theta}_{X,\ell}, \tau, \boldsymbol{G}, \boldsymbol{b}_{-(ij)}, \mathbf{D}, M_{\ell,\ell'}) \propto & \left\{ \prod_{q=1}^{Q} \tilde{\phi}_{\ell' iq}^{-\tilde{\eta}_{iq}} \exp\left(\delta_{ijq} \nu_{ij} \boldsymbol{\alpha}^{\intercal} \boldsymbol{b}_{ij}\right) \right\} \\ & \times \exp\left\{ -\frac{\tau}{2} \left(\boldsymbol{X}_{ij} - \boldsymbol{W}_{X,\ell ij} \boldsymbol{\theta}_{X,\ell} - \boldsymbol{z}(t)^{\intercal} \boldsymbol{b}_{ij} \right)^{\intercal} \left(\boldsymbol{X}_{ij} - \boldsymbol{W}_{X,\ell ij} \boldsymbol{\theta}_{X,\ell} - \boldsymbol{z}(t)^{\intercal} \boldsymbol{b}_{ij} \right) \right\} \\ & \times p(\boldsymbol{b}_{ij}|\boldsymbol{G}, M_{\ell,\ell'}) \end{aligned}$$

C.1.2 Algorithm Details

Let $M_{\ell,\ell'}$ denote the joint model with longitudinal submodel ℓ and survival submodel ℓ'. Our goal is to estimate the posterior modes of the full conditional distributions of the parameters and random effects. The algorithm steps are as follows:

1. Obtain initial values of all parameters.

   (a) Set $\boldsymbol{\theta}_{X,\ell}^{(0)}$, $\tau^{(0)}$, and $\boldsymbol{G}^{-1(0)}$ equal to the maximum likelihood estimates (MLEs) and $\boldsymbol{b}^{(0)}$ equal to the empirical best linear unbiased predictors from a linear mixed model corresponding to longitudinal submodel ℓ.

   (b) Set $\boldsymbol{\theta}_{Y,\ell'}^{(0)}$ equal to the MLEs from a Cox proportional hazards model corresponding to survival submodel ℓ'.

   (c) Set $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$ for the first model, or set $\boldsymbol{\alpha}^{(0)}$ equal to the estimated posterior mode of $\boldsymbol{\alpha}$ from the previous model.

2. Update the current estimates of all parameters and random effects for iteration m by evaluating the modes of the full conditional distributions (see Section A.1).

   (a) Update $\boldsymbol{\alpha}^{(m)}$ by evaluating the mode of $p\left(\boldsymbol{\alpha}^{(m)}|\boldsymbol{\theta}_{Y,\ell}^{(m-1)}, \boldsymbol{b}^{(m-1)}, \mathbf{D}, M_{\ell,\ell'}\right)$.

   (b) Update $\boldsymbol{\theta}_{Y,\ell}^{(m)}$ by evaluating the mode of $p\left(\boldsymbol{\theta}_{Y,\ell}^{(m)}|\boldsymbol{\alpha}^{(m)}, \boldsymbol{b}^{(m-1)}, \mathbf{D}, M_{\ell,\ell'}\right)$.

   (c) Update $\tau^{(m)}$ from $p\left(\tau^{(m)}|\boldsymbol{\theta}_{X,\ell}^{(m-1)}, \boldsymbol{b}^{(m-1)}, \mathbf{D}, M_{\ell,\ell'}\right)$ (gamma distribution).

   (d) Update $\boldsymbol{\theta}_{X,\ell}^{(m)}$ from $p\left(\boldsymbol{\theta}_{X,\ell}^{(m)}|\tau^{(m)}, \boldsymbol{b}^{(m-1)}, \mathbf{D}, M_{\ell,\ell'}\right)$ (multivariate normal distribution).

   (e) Update $\boldsymbol{G}^{-1(m)}$ from $p\left(\boldsymbol{G}^{-1(m)}|\boldsymbol{b}^{(m-1)}, \mathbf{D}, M_{\ell,\ell'}\right)$ (Wishart distribution).

   (f) Update $\boldsymbol{b}_{ij}^{(m)}$ (the random effects for the jth subject from the ith region) by evaluating the mode of $p\left(\boldsymbol{b}_{ij}^{(m)}|\boldsymbol{\alpha}^{(m)}, \boldsymbol{\theta}_{Y,\ell}^{(m)}, \tau^{(m)}, \boldsymbol{\theta}_{X,\ell}^{(m)}, \boldsymbol{G}^{(m)}, \boldsymbol{b}_{i'j'<ij}^{(m)}, \boldsymbol{b}_{i'j'>ij}^{(m-1)}, \mathbf{D}, M_{\ell,\ell'}\right)$.

3. Repeat Step 2 until convergence of all parameters and random effects.

C.2 Hessian Matrices Used in the Laplace Approximations

C.2.1 Hessian of Full Conditional Distribution of Association Parameters (α)

Let $h_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \log\{p(\boldsymbol{\alpha}|\boldsymbol{\theta}_{Y,\ell'}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'})\}$ and $\boldsymbol{\Psi}_{\boldsymbol{\alpha}} = -\left\{\ddot{h}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\right\}^{-1}$, where

$$\begin{aligned}
\ddot{h}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) &= \frac{d^2 h_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})}{d\boldsymbol{\alpha} d\boldsymbol{\alpha}^\intercal} \\
&= -\boldsymbol{\Sigma}_\alpha^{-1} - \sum_{i=1}^{S}\sum_{q=1}^{Q} \tilde{\eta}_{iq} \left[\left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq}\hat{g}_{\ell' ij}\boldsymbol{b}_{ij}\boldsymbol{b}_{ij}^\intercal \right\} \left\{ \phi_{iq} + \sum_{j=1}^{n_i} \hat{c}_{ijq}\hat{g}_{\ell' ij} \right\}^{-1} \right. \\
&\quad \left. - \left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq}\hat{g}_{\ell' ij}\boldsymbol{b}_{ij} \right\} \left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq}\hat{g}_{\ell' ij}\boldsymbol{b}_{ij} \right\}^\intercal \left\{ \phi_{iq} + \sum_{j=1}^{n_i} \hat{c}_{ijq}\hat{g}_{\ell' ij} \right\}^{-2} \right],
\end{aligned}$$

where

- $\tilde{\eta}_{iq} = \eta_{iq} + \sum_{j=1}^{n_i} \delta_{ijq}\nu_{ij}$,
- $\hat{c}_{ijq} = \delta_{ijq}(y_{ij} - m_{q-1}) + \sum_{g=q+1}^{Q} \delta_{ijg}(m_q - m_{q-1})$,
- $\hat{g}_{\ell' ij} = \exp(\boldsymbol{\alpha}^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}_{Y,\ell' ij}^\intercal \boldsymbol{\theta}_{Y,\ell'})$.

C.2.2 Hessian of Full Conditional Distribution of Survival Regression Effects (θ_Y)

Let $\Omega_{\ell',d}$ be the set of region labels for all regions that share the dth distinct treatment effect for some survival submodel ℓ', $d = 1, \ldots, D_{Y,\ell'}$. Let $h_{\boldsymbol{\theta}_{Y,\ell'}}(\boldsymbol{\theta}_{Y,\ell'}) = \log\{p(\boldsymbol{\theta}_{Y,\ell'}|\boldsymbol{\alpha}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'})\}$ and $\boldsymbol{\Psi}_{\boldsymbol{\theta}_{Y,\ell'}} = -\left\{\ddot{h}_{\boldsymbol{\theta}_{Y,\ell'}}(\boldsymbol{\theta}_{Y,\ell'})\right\}^{-1}$, where

$$
\begin{aligned}
\ddot{h}_{\boldsymbol{\theta}_{Y,\ell'}}(\boldsymbol{\theta}_{Y,\ell'}) &= \frac{d^2 h_{\boldsymbol{\theta}_{Y,\ell'}}(\boldsymbol{\theta}_{Y,\ell'})}{d\boldsymbol{\theta}_{Y,\ell'} d\boldsymbol{\theta}_{Y,\ell'}^{\intercal}} \\
&= -\boldsymbol{\Sigma}_{Y,\ell'}^{-1} - \sum_{d=1}^{D_{Y,\ell'}} \sum_{i \in \Omega_{\ell',d}} \sum_{q=1}^{Q} \tilde{\eta}_{iq} \left[\left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \boldsymbol{w}_{Y,\ell' ij} \boldsymbol{w}_{Y,\ell' ij}^{\intercal} \right\} \left\{ \phi_{iq} + \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \right\}^{-1} \right. \\
&\quad \left. - \left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \boldsymbol{w}_{Y,\ell' ij} \right\} \left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \boldsymbol{w}_{Y,\ell' ij} \right\}^{\intercal} \left\{ \phi_{iq} + \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \right\}^{-2} \right],
\end{aligned}
$$

where

- $\tilde{\eta}_{iq} = \eta_{iq} + \sum_{j=1}^{n_i} \delta_{ijq} \nu_{ij}$,
- $\hat{c}_{ijq} = \delta_{ijq}(y_{ij} - m_{q-1}) + \sum_{g=q+1}^{Q} \delta_{ijg}(m_q - m_{q-1})$,
- $\hat{g}_{\ell' ij} = \exp(\boldsymbol{\alpha}^{\intercal} \boldsymbol{b}_{ij} + \boldsymbol{w}_{Y,\ell' ij}^{\intercal} \boldsymbol{\theta}_{Y,\ell'})$.

C.2.3 Hessian of Likelihood Multiplied by Priors for Fixed Effects $\left(\xi^*_{\ell,\ell'}\right)$

Let $h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'}) = \log\left\{p(\mathbf{D}|\boldsymbol{\xi}^*_{\ell,\ell'}, \boldsymbol{b}, \mathbf{D}, M_{\ell,\ell'})p(\boldsymbol{\xi}^*_{\ell,\ell'}|M_{\ell,\ell'})\right\}$, $\boldsymbol{\Psi}_{\boldsymbol{\xi}^*_{\ell,\ell'}} = -\left\{\ddot{h}_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})\right\}^{-1}$, and

$$\ddot{h}_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'}) = \begin{bmatrix} \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\alpha} d\boldsymbol{\alpha}^\intercal} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\alpha} d\boldsymbol{\theta}_{Y,\ell'}^\intercal} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{Y,\ell'} d\boldsymbol{\alpha}^\intercal} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{Y,\ell'} d\boldsymbol{\theta}_{Y,\ell'}^\intercal} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{X,\ell} d\boldsymbol{\theta}_{X,\ell}^\intercal} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{X,\ell} d\tau} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tau d\boldsymbol{\theta}_{X,\ell}^\intercal} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tau^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{G}^{-1} d\boldsymbol{G}^{-1}} \end{bmatrix},$$

where

$$\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\alpha} d\boldsymbol{\alpha}^\intercal} = \frac{d^2 h_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})}{d\boldsymbol{\alpha} d\boldsymbol{\alpha}^\intercal} \text{ (see Section B.1),}$$

$$\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{Y,\ell'} d\boldsymbol{\theta}_{Y,\ell'}^\intercal} = \frac{d^2 h_{\boldsymbol{\theta}_{Y,\ell'}}(\boldsymbol{\theta}_{Y,\ell'})}{d\boldsymbol{\theta}_{Y,\ell'} d\boldsymbol{\theta}_{Y,\ell'}^\intercal} \text{ (see Section B.2),}$$

$$\begin{aligned} \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{Y,\ell'} d\boldsymbol{\alpha}^\intercal} &= \left(\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\alpha} d\boldsymbol{\theta}_{Y,\ell'}^\intercal}\right)^\intercal \\ &= -\sum_{d=1}^{D_{Y,\ell'}} \sum_{i \in \Omega_{\ell',d}} \sum_{q=1}^{Q} \tilde{\eta}_{iq} \left[\left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \boldsymbol{w}_{Y,\ell' ij} \boldsymbol{b}_{ij}^\intercal \right\} \left\{ \phi_{iq} + \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \right\}^{-1} \right. \\ &\quad \left. - \left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \boldsymbol{w}_{Y,\ell' ij} \right\} \left\{ \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \boldsymbol{b}_{ij} \right\}^\intercal \left\{ \phi_{iq} + \sum_{j=1}^{n_i} \hat{c}_{ijq} \hat{g}_{\ell' ij} \right\}^{-2} \right], \end{aligned}$$

with

- $\tilde{\eta}_{iq} = \eta_{iq} + \sum_{k=1}^{n_i} \delta_{ikq}\nu_{ik}$,
- $\hat{c}_{ikq} = \delta_{ikq}(y_{ik} - m_{q-1}) + \sum_{g=q+1}^{Q} \delta_{ikg}(m_q - m_{q-1})$,
- $\hat{g}_{\ell' ij} = \exp(\boldsymbol{\alpha}^\intercal \boldsymbol{b}_{ij} + \boldsymbol{w}_{Y,\ell' ij}^\intercal \boldsymbol{\theta}_{Y,\ell'})$.

Let $\Omega_{\ell,d}$ denote the set of region labels for all regions that share the dth distinct treatment effect for some longitudinal submodel ℓ, $d = 1, \ldots, D_{X,\ell}$. Then

$$\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{X,\ell} d\boldsymbol{\theta}_{X,\ell}^\intercal} = -\tau \tilde{\boldsymbol{\Sigma}}_{X,\ell}^{-1},$$

$$\text{with } \tilde{\boldsymbol{\Sigma}}_{X,\ell} = \left(\sum_{d=1}^{D_{X,\ell}} \sum_{i \in \Omega_{\ell,d}} \sum_{j=1}^{n_i} \boldsymbol{W}_{X,\ell ij}^\intercal \boldsymbol{W}_{X,\ell ij} + \boldsymbol{\Sigma}_{X,\ell}^{-1} \right)^{-1},$$

$$\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tau^2} = -\frac{1}{2\tau^2} \left(\sum_{d=1}^{D_{X,\ell}} \sum_{i \in \Omega_{\ell,d}} \sum_{j=1}^{n_i} K_{ij} \right) - \frac{1}{\tau^2} \left(\frac{\eta_\tau}{2} - 1 \right),$$

$$\begin{aligned} \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{\theta}_{X,\ell} d\tau} &= \left(\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tau d\boldsymbol{\theta}_{X,\ell}^\intercal} \right)^\intercal \\ &= -\tilde{\boldsymbol{\Sigma}}_{X,\ell}^{-1} (\boldsymbol{\theta}_{X,\ell} - \tilde{\boldsymbol{\mu}}_{X,\ell}) \\ &= 0 \text{ if } \boldsymbol{\theta}_{X,\ell} \text{ is optimized to equal } \tilde{\boldsymbol{\mu}}_{X,\ell}, \end{aligned}$$

$$\text{with } \tilde{\boldsymbol{\mu}}_{X,\ell} = \tilde{\boldsymbol{\Sigma}}_{X,\ell} \left\{ \sum_{d=1}^{D_{X,\ell}} \sum_{i \in \Omega_{\ell,d}} \sum_{j=1}^{n_i} \boldsymbol{W}_{X,\ell ij}^\intercal \left(\boldsymbol{X}_{ij} - \boldsymbol{z}(t)^\intercal \boldsymbol{b}_{ij} \right) + \boldsymbol{\Sigma}_{X,\ell}^{-1} \boldsymbol{\mu}_{X,\ell} \right\}.$$

If only random intercepts are included, then let $\boldsymbol{b}_{ij} = b_{0ij}$, $\boldsymbol{C}_0^{-1} = \tilde{C}_{011}$, and $\boldsymbol{G}^{-1} = \tilde{g}_{11}$. Then

$$\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{G}^{-1} d\boldsymbol{G}^{-1}} = -\frac{1}{2\tilde{g}_{11}^2} (\nu_0 - r + N - 1).$$

If both random intercepts and random slopes are included, then let

$$\boldsymbol{b}_{ij} = \begin{bmatrix} b_{0ij} \\ b_{1ij} \end{bmatrix}, \quad \boldsymbol{C}_0^{-1} = \begin{bmatrix} \tilde{C}_{011} & \tilde{C}_{012} \\ \tilde{C}_{012} & \tilde{C}_{022} \end{bmatrix}, \quad \boldsymbol{G}^{-1} = \begin{bmatrix} \tilde{g}_{11} & \tilde{g}_{12} \\ \tilde{g}_{12} & \tilde{g}_{22} \end{bmatrix}.$$

It follows that

$$\frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\boldsymbol{G}^{-1} d\boldsymbol{G}^{-1}} = \begin{bmatrix} \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{11}^2} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{11} d\tilde{g}_{22}} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{11} d\tilde{g}_{12}} \\ \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{11} d\tilde{g}_{22}} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{22}^2} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{22} d\tilde{g}_{12}} \\ \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{11} d\tilde{g}_{12}} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{22} d\tilde{g}_{12}} & \frac{d^2 h_{\boldsymbol{\xi}^*_{\ell,\ell'}}(\boldsymbol{\xi}^*_{\ell,\ell'})}{d\tilde{g}_{12}^2} \end{bmatrix}$$

$$= -\frac{(\nu_0 - r + N - 1)}{(\tilde{g}_{11}\tilde{g}_{22} - 2\tilde{g}_{12})^2} \begin{bmatrix} \frac{\tilde{g}_{22}^2}{2} & \tilde{g}_{12} & -\tilde{g}_{22} \\ \tilde{g}_{12} & \frac{\tilde{g}_{11}^2}{2} & -\tilde{g}_{11} \\ -\tilde{g}_{22} & -\tilde{g}_{11} & 2 \end{bmatrix}.$$

C.3 Details of Data Generation in Simulation Studies

For each simulation study, we generate datasets designed to mirror the LEADER trial data. Marso et al. (2016) state that $N = 9340$ patients from $S = 4$ regions underwent randomization from September 2010 through April 2012 (i.e., 18 months).

In the data simulation process, we imitate the 18-month enrollment period and maximum follow-up time of 60 months, and we set the dropout rate to 0.0082 (i.e., we assume theoretical dropout times are randomly sampled from an exponential distribution with mean 0.0082^{-1}). For the survival data, we set the baseline hazard to increase from 0.0026 to 0.0045 at a constant rate of 0.0001 per 3-month period. We sample the longitudinal observations for each visit from a normal distribution with a common standard deviation of $\sigma = 0.886$ and means specific to each visit and treatment group (see Table C.1).

Table C.1: Means of simulated longitudinal HbA1c values for each visit by treatment group.

Visit	1	2	3	4	5	6	7	8	9	10	11
Month	0	3	6	12	18	24	30	36	42	48	54
Placebo	8.7	8.2	8.1	8.0	7.9	7.9	7.9	7.9	7.9	7.9	8.0
Treatment	8.7	7.2	7.2	7.3	7.3	7.4	7.4	7.5	7.5	7.5	7.6

C.4 Comparison Models in Simulation Studies

C.4.1 Cox Proportional Hazards Models

We consider two Cox proportional hazards models (CPHMs): one to estimate the global treatment effect γ_G and one to estimate the region-specific treatment effects $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S)$. For the first CPHM, we define λ_0 to be the baseline hazard where $\lambda_0 \sim \text{Gamma}(\eta_{\lambda 0}, \phi_{\lambda 0})$. Additionally, we define $\boldsymbol{\theta}_1^* = (\gamma_G, \boldsymbol{\beta}^\intercal)^\intercal$ where $\boldsymbol{\beta}$ is a $p \times 1$ vector of covariate effects, and we assume $\boldsymbol{\theta}_1^* \sim N_{(p+1)}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Let $\boldsymbol{w}_{ij}$ be the $(p+1) \times 1$ vector corresponding to $\boldsymbol{\theta}_1^*$ for the jth subject in the ith region where the first element is a treatment indicator (1 for treatment and 0 for control) and the last p elements are optional covariates. Lastly, we define $\boldsymbol{\xi}_1^* = \{\lambda_0, \boldsymbol{\theta}_1^*\}$.

The likelihood for the first CPHM can be written as

$$\mathcal{L}(\boldsymbol{\xi}_1^* | \mathbf{D}) = \prod_{i=1}^{S} \prod_{j=1}^{n_i} \left[\left\{ \lambda_0 \exp\left(\boldsymbol{w}_{ij}^\intercal \boldsymbol{\theta}_1^* \right) \right\}^{\nu_{ij}} \times \exp\left\{ -\lambda_0 y_{ij} \exp\left(\boldsymbol{w}_{ij}^\intercal \boldsymbol{\theta}_1^* \right) \right\} \right],$$

where y_{ij} and ν_{ij} are as defined in Section 5.3.1. The full conditional distribution of $\lambda_0 | \boldsymbol{\theta}_1^*, \mathbf{D}$ is $\lambda_0 | \boldsymbol{\theta}_1^*, \mathbf{D} \sim \text{Gamma}(\eta_0^*, \phi_0^*)$, where

$$\eta_0^* = \sum_{i=1}^{S} \sum_{j=1}^{n_i} \nu_{ij} + \eta_{\lambda 0},$$

$$\phi_0^* = \phi_{\lambda 0} + \sum_{i=1}^{S} \sum_{j=1}^{n_i} y_{ij} \exp\left(\boldsymbol{w}_{ij}^\intercal \boldsymbol{\theta}_1^* \right).$$

The marginal posterior distribution of $\boldsymbol{\theta}_1^* | \mathbf{D}$ is

$$p(\boldsymbol{\theta}_1^* | \mathbf{D}) \propto \exp\left\{ \sum_{i=1}^{S} \sum_{j=1}^{n_i} \nu_{ij} \boldsymbol{w}_{ij}^\intercal \boldsymbol{\theta}_1^* \right\} \times \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\theta}_1^* - \boldsymbol{\mu}_1 \right)^\intercal \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\theta}_1^* - \boldsymbol{\mu}_1 \right) \right\} \times \left(\phi_0^* \right)^{-\eta_0^*}.$$

For the second CPHM, we define $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)$ to be the region-specific baseline hazards, where $\lambda_i \sim \text{Gamma}(\eta_{\lambda i}, \phi_{\lambda i})$, $i = 1, \ldots, S$. We define $\boldsymbol{\theta}_2^* = (\boldsymbol{\gamma}, \boldsymbol{\beta}^\intercal)^\intercal$ and $\boldsymbol{\xi}_2^* = \{\boldsymbol{\lambda}, \boldsymbol{\theta}_2^*\}$, and

we now let $\boldsymbol{w}_{ij}$ be the $(S+p)\times 1$ vector corresponding to $\boldsymbol{\theta}_2^*$ where the first S elements are region-specific treatment indicators.

The likelihood for the second CPHM can be written as

$$\mathcal{L}(\boldsymbol{\xi}_2^*|\mathbf{D}) = \prod_{i=1}^{S}\prod_{j=1}^{n_i}\left[\left\{\lambda_i \exp\left(\boldsymbol{w}_{ij}^{\intercal}\boldsymbol{\theta}_2^*\right)\right\}^{\nu_{ij}} \times \exp\left\{-\lambda_i y_{ij}\exp\left(\boldsymbol{w}_{ij}^{\intercal}\boldsymbol{\theta}_2^*\right)\right\}\right].$$

The full conditional distribution of $\lambda_i|\boldsymbol{\theta}_2^*, \mathbf{D}$ is $\lambda_i|\boldsymbol{\theta}_2^*, \mathbf{D} \sim \text{Gamma}(\eta_i^*, \phi_i^*)$, $i = 1, \ldots, S$, where

$$\eta_i^* = \sum_{i=1}^{S}\sum_{j=1}^{n_i}\nu_{ij} + \eta_{\lambda i},$$

$$\phi_i^* = \phi_{\lambda i} + \sum_{i=1}^{S}\sum_{j=1}^{n_i} y_{ij}\exp\left(\boldsymbol{w}_{ij}^{\intercal}\boldsymbol{\theta}_2^*\right).$$

If we assume $\boldsymbol{\theta}_2^* \sim N_{(S+p)}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, then the marginal posterior distribution of $\boldsymbol{\theta}_2^*|\mathbf{D}$ is

$$p(\boldsymbol{\theta}_2^*|\mathbf{D}) \propto \exp\left\{\sum_{i=1}^{S}\sum_{j=1}^{n_i}\nu_{ij}\boldsymbol{w}_{ij}^{\intercal}\boldsymbol{\theta}_2^*\right\} \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}_2^* - \boldsymbol{\mu}_2\right)^{\intercal}\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\theta}_2^* - \boldsymbol{\mu}_2\right)\right\} \times \prod_{i=1}^{S}\left(\phi_i^*\right)^{-\eta_i^*}.$$

We elicit the prior distributions as $\boldsymbol{\theta}_1^* \sim N_{(p+1)}(\mathbf{0}, 10{,}000\boldsymbol{I}_{p+1})$ and $\lambda_0 \sim \text{Gamma}(0.01, 0.01)$ for the first CPHM, and $\boldsymbol{\theta}_2^* \sim N_{(S+p)}(\mathbf{0}, 10{,}000\boldsymbol{I}_{S+p})$ and $\lambda_i \sim \text{Gamma}(0.01, 0.01)$, $i = \ldots, S$, for the second CPHM, where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. To test the global null hypothesis using simulation studies discussed in Section 5.3.4, we calculate the upper limit of a 95% credible interval for $\gamma_G|\mathbf{D}$ using the first CPHM, and we reject the null hypothesis if γ_0 is greater than this limit. Similarly, we reject the null hypothesis for each region-specific treatment effect using the second CPHM if the upper boundaries of the corresponding 95% credible intervals are below γ_0. For all simulation studies presented in the per, we set $\gamma_0 = 0$.

C.4.2 Survival-Only Bayesian Model Averaging Approach

We also consider the Bayesian model averaging approach for survival data only (BMA-S) as proposed by Bean, Ibrahim, and Psioda (2022). We partition the time axis into $Q = 8$ intervals where each interval contains an approximately equal number of observed events, and we assume each region has a unique constant baseline hazard in each interval. If we consider a model space with L models and let M_ℓ denote the ℓth model, then we can define $\boldsymbol{\xi}_\ell = \{\boldsymbol{\lambda}, \boldsymbol{\gamma}_\ell\}$ to be the model parameters under M_ℓ, $\ell = 1, \ldots, L$, where $\boldsymbol{\lambda}$ is the $Q \times S$ constant baseline hazards and $\boldsymbol{\gamma}_\ell$ is the vector of D_ℓ distinct region-specific treatment effects (i.e., log of the hazard ratio).

For the prior elicitation under M_ℓ, we set

$$p(\boldsymbol{\xi}_\ell|M_\ell) = p(\boldsymbol{\gamma}_\ell|M_\ell) \times \left\{ \prod_{i=1}^{S} \prod_{q=1}^{Q} p(\lambda_{iq}|M_\ell) \right\},$$

where

$$\boldsymbol{\gamma}_\ell|M_\ell \sim N_{D_\ell}(\boldsymbol{\mu}_{0\ell}, \boldsymbol{\Sigma}_{0\ell}),$$

$$\lambda_{iq}|M_\ell \sim \text{Gamma}(\eta_{iq}, \phi_{iq}).$$

We set each element of $\boldsymbol{\mu}_{0\ell}$ to equal $\log(1.3)$ and $\boldsymbol{\Sigma}_{0\ell} = \text{Diag}\{(10 \times |\boldsymbol{\mu}_{0\ell}|)^2\}$, $\ell = 1, \ldots, L$, and we choose $\eta_{iq} = 0.01$ and $\phi_{iq} = 0.01$, $i = 1, \ldots, S$, $q = 1, \ldots, Q$. Lastly, we assume uniform prior model probabilities; i.e., $p(M_\ell) \propto 1$.

Let γ denote either the global treatment effect or a region-specific treatment effect, and set $\gamma_0 = 0$. We test $H_0 : \gamma \geqslant \gamma_0$ versus $H_1 : \gamma < \gamma_0$ by calculating

$$P(\gamma < \gamma_0|\mathbf{D}) = \sum_{\ell=1}^{L} P(\gamma < \gamma_0|\mathbf{D}, M_\ell) p(M_\ell|\mathbf{D}).$$

C.5 Additional Simulation Studies

C.5.1 Equal Sample Sizes with $\alpha \in \{0, 0.15\}$

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Association parameter: $\alpha = 0$

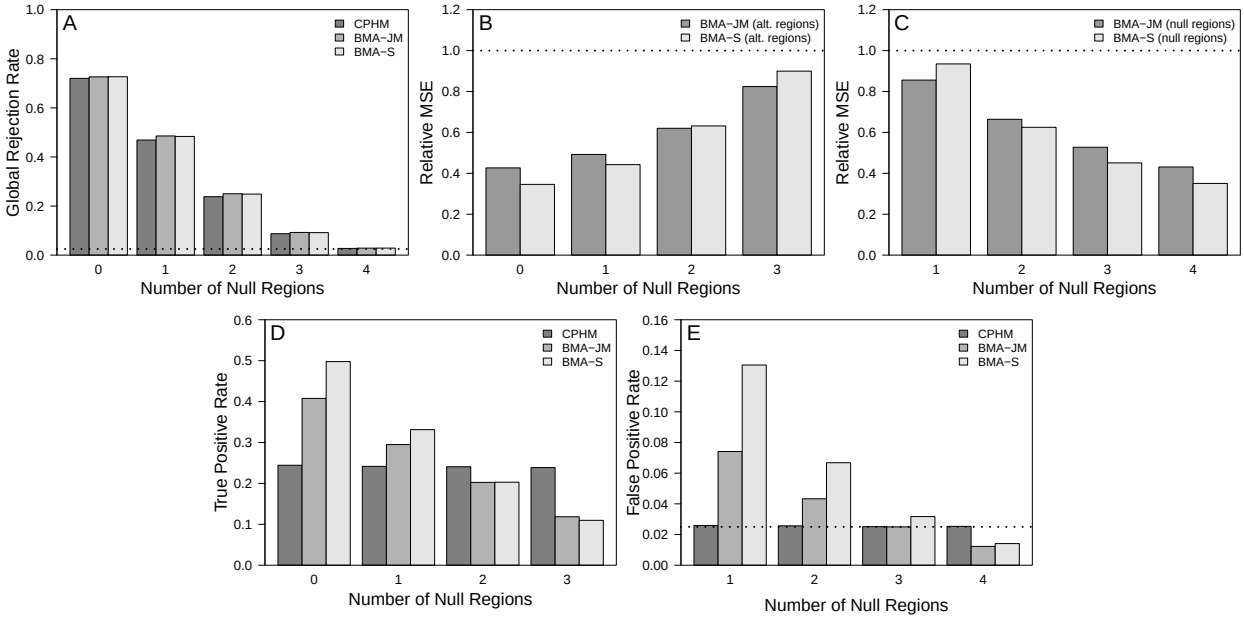

Figure C.1: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study with equal regional sample sizes and $\alpha = 0$.

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Association parameter: $\alpha = 0.15$

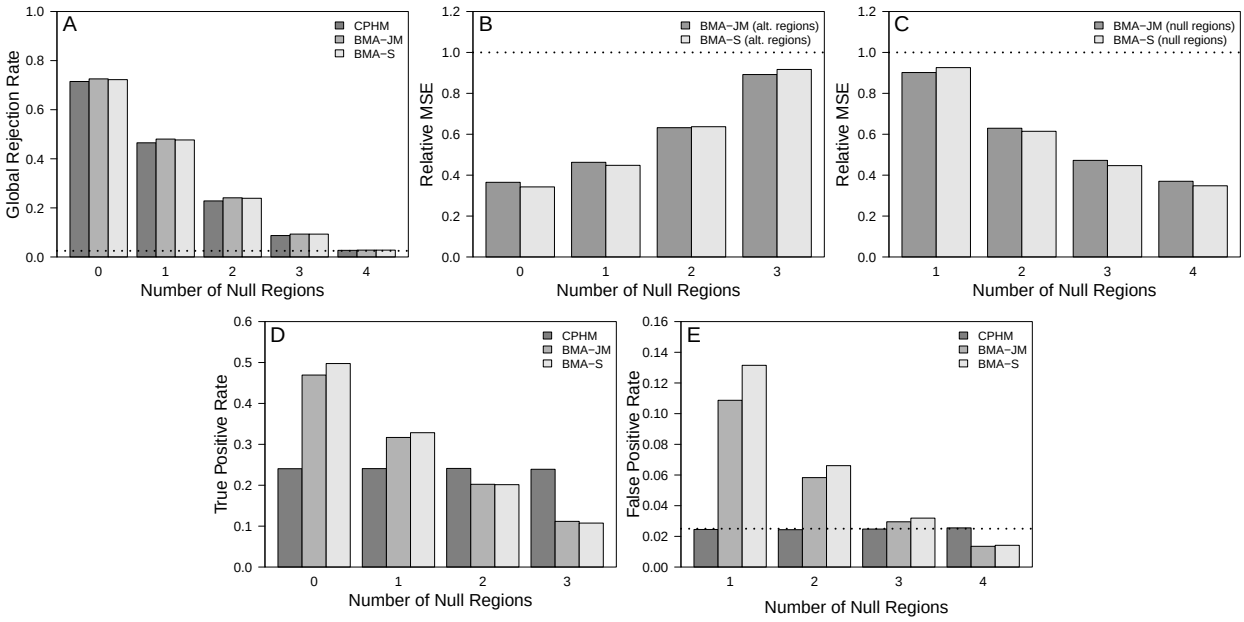

Figure C.2: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study with equal regional sample sizes and $\alpha = 0.15$.

C.5.2 Unequal Sample Sizes with $\alpha \in \{0, 0.15, 1.0\}$

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Regional sample sizes equal to original sample sizes in LEADER trial: sample sizes of 711, 3296, 2847, and 2486 for Asia, Europe, North America, and Rest of the World, respectively.
- Scenario 1: underlying treatment-to-placebo hazard ratio for all regions equal to 0.868 (i.e., treatment effect of -0.142)
- Scenario 2: underlying treatment-to-placebo hazard ratios of 0.62, 0.82, 1.01, and 0.83 for Asia, Europe, North America, and Rest of the World, respectively.
- Association parameter: $\alpha = 0$

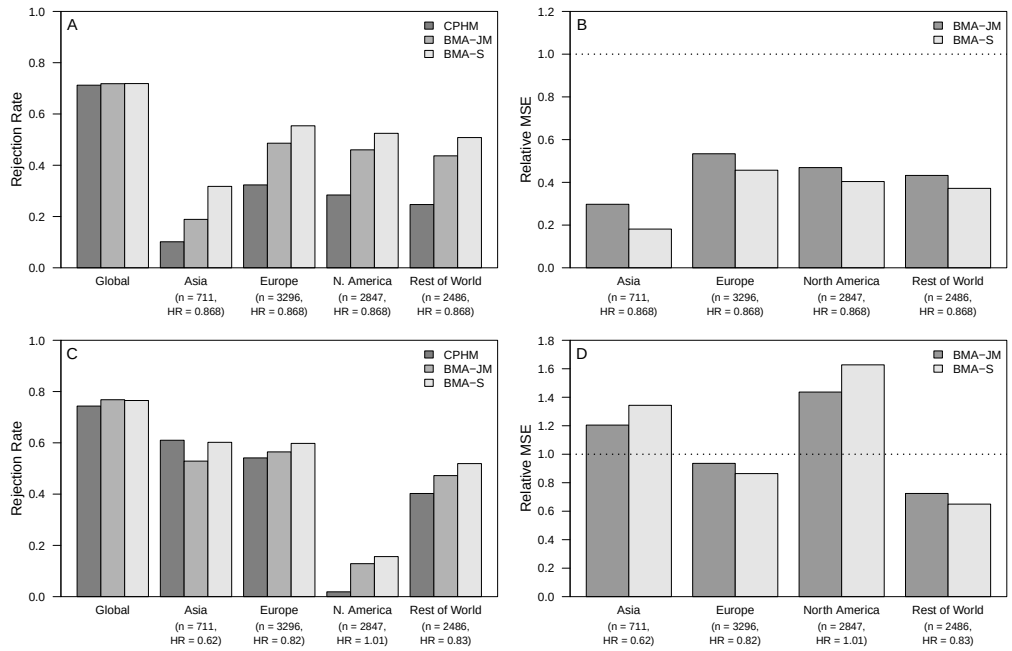

Figure C.3: Rejection rates (*Panel A*) and MSE relative to CPHM (*Panel B*) for the scenario with all treatment-to-placebo hazard ratios equal to 0.868, and rejection rates (*Panel C*) and relative MSE (*Panel D*) for the scenario with differing treatment-to-placebo hazard ratios. Both scenarios consider unequal regional sample sizes and $\alpha = 0$.

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Regional sample sizes equal to original sample sizes in LEADER trial: sample sizes of 711, 3296, 2847, and 2486 for Asia, Europe, North America, and Rest of the World, respectively.
- Scenario 1: underlying treatment-to-placebo hazard ratio for all regions equal to 0.868 (i.e., treatment effect of -0.142)
- Scenario 2: underlying treatment-to-placebo hazard ratios of 0.62, 0.82, 1.01, and 0.83 for Asia, Europe, North America, and Rest of the World, respectively.
- Association parameter: $\alpha = 0.15$

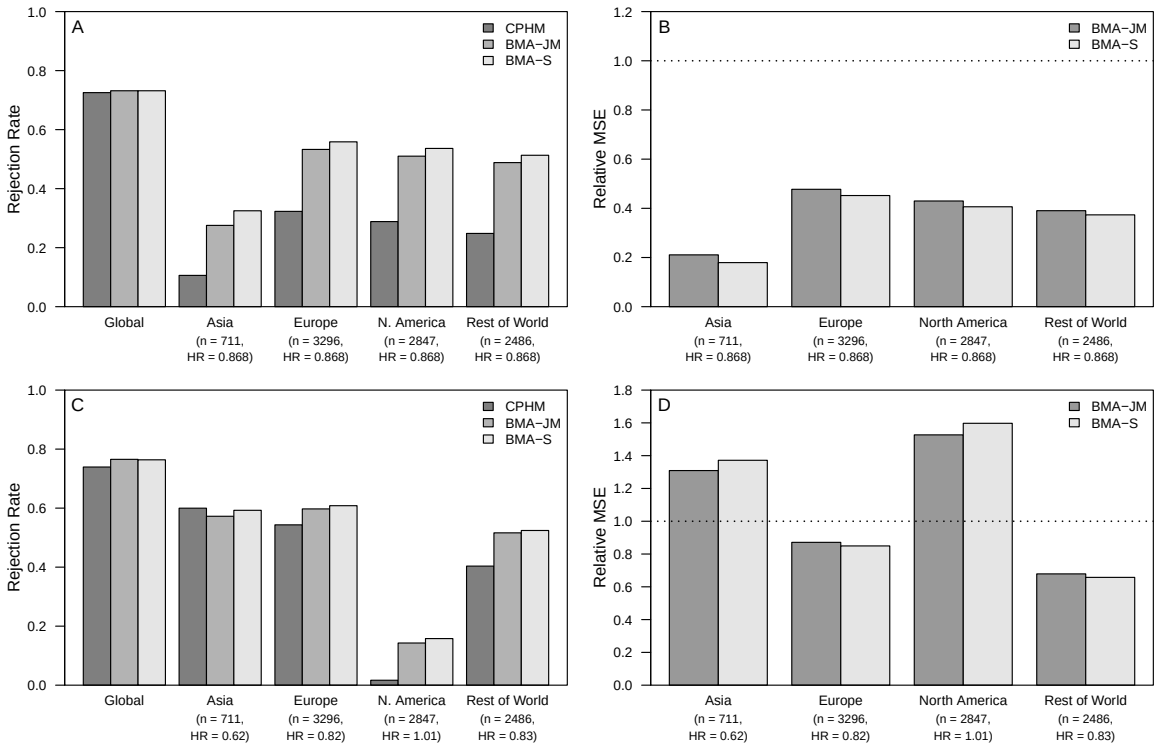

Figure C.4: Rejection rates (*Panel A*) and MSE relative to CPHM (*Panel B*) for the scenario with all treatment-to-placebo hazard ratios equal to 0.868, and rejection rates (*Panel C*) and relative MSE (*Panel D*) for the scenario with differing treatment-to-placebo hazard ratios. Both scenarios consider unequal regional sample sizes and $\alpha = 0.15$.

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Regional sample sizes equal to original sample sizes in LEADER trial: sample sizes of 711, 3296, 2847, and 2486 for Asia, Europe, North America, and Rest of the World, respectively.
- Scenario 1: underlying treatment-to-placebo hazard ratio for all regions equal to 0.868 (i.e., treatment effect of -0.142)
- Scenario 2: underlying treatment-to-placebo hazard ratios of 0.62, 0.82, 1.01, and 0.83 for Asia, Europe, North America, and Rest of the World, respectively.
- Association parameter: $\alpha = 1.0$

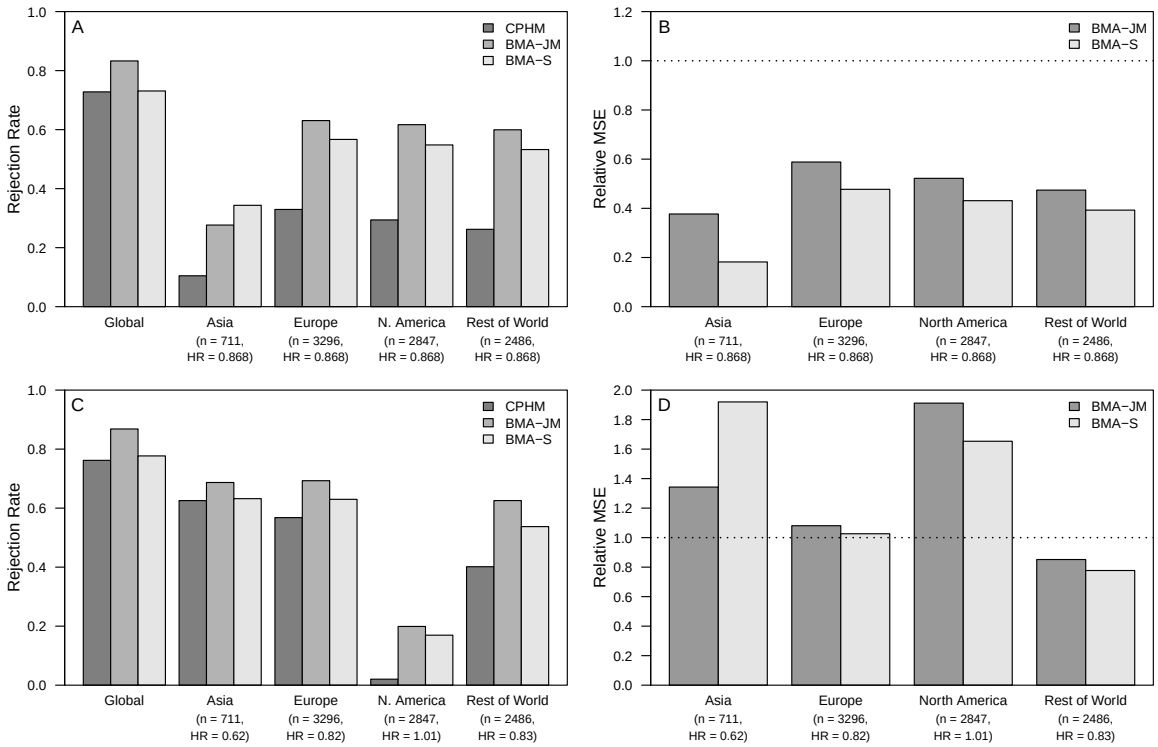

Figure C.5: Rejection rates (*Panel A*) and MSE relative to CPHM (*Panel B*) for the scenario with all treatment-to-placebo hazard ratios equal to 0.868, and rejection rates (*Panel C*) and relative MSE (*Panel D*) for the scenario with differing treatment-to-placebo hazard ratios. Both scenarios consider unequal regional sample sizes and $\alpha = 1.0$.

C.5.3 Sample Sizes of Null Regions Half the Size of Alternative Regions

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Sample sizes of null regions half the size of alternative regions
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Association parameter: $\alpha = 0.5$

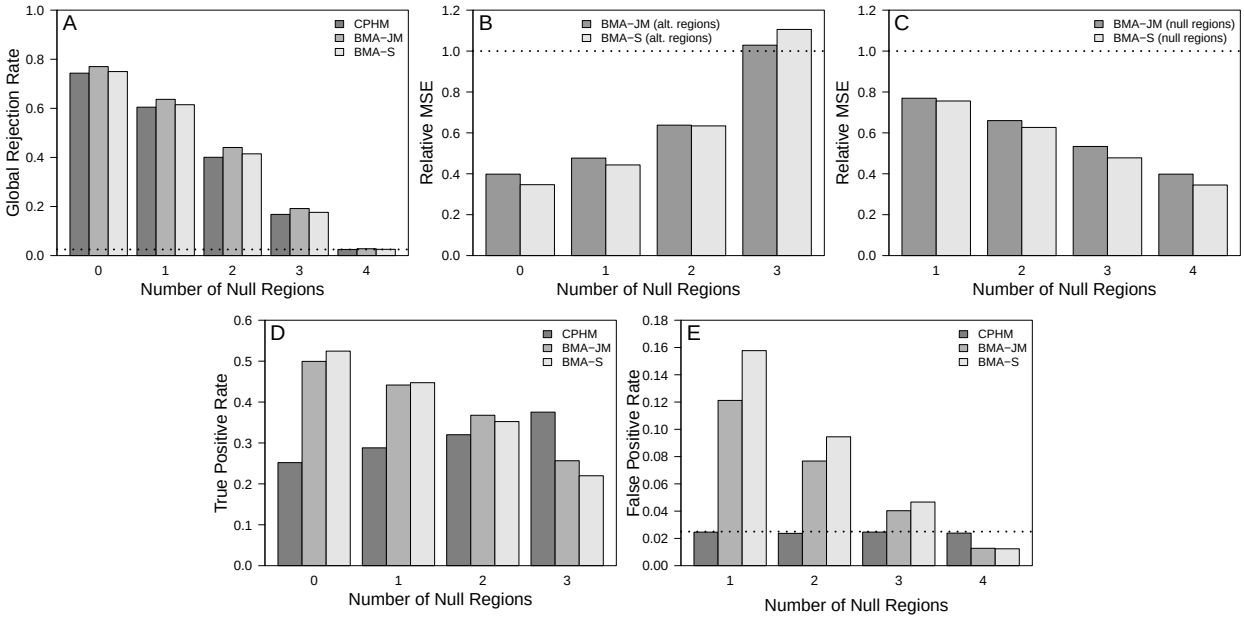

Figure C.6: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study where sample sizes of null regions are half the size of alternative regions.

C.5.4 Sample Sizes of Null Regions Double the Size of Alternative Regions

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Sample sizes of null regions double the size of alternative regions
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Association parameter: $\alpha = 0.5$

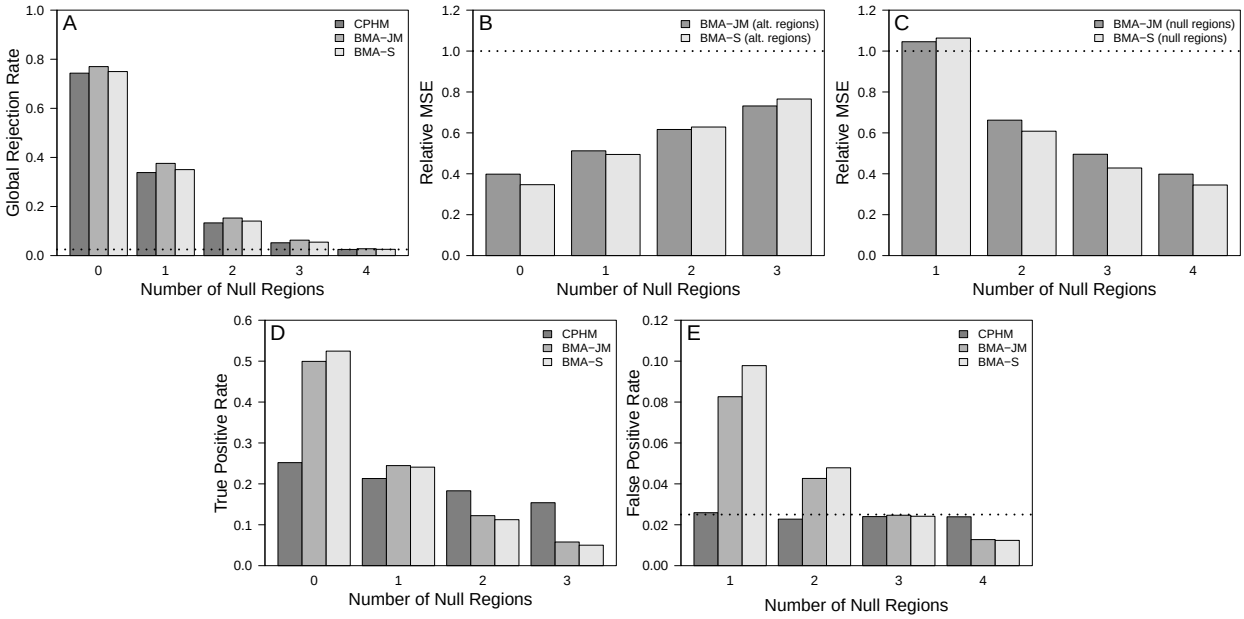

Figure C.7: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for simulation study where sample sizes of null regions are double the size of alternative regions.

C.5.5 Sensitivity Analysis: Change Prior Model Probabilities

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Number of time intervals in piecewise constant baseline hazard: $Q = 8$
- Association parameter: $\alpha = 0.5$
- Prior model probabilities: $p(M_{\ell,\ell'}) \propto \exp\left(a_X D_{X,\ell} + a_Y D_{Y,\ell'}\right)$, where $a_X, a_Y \in \{-1, 0, 1\}$

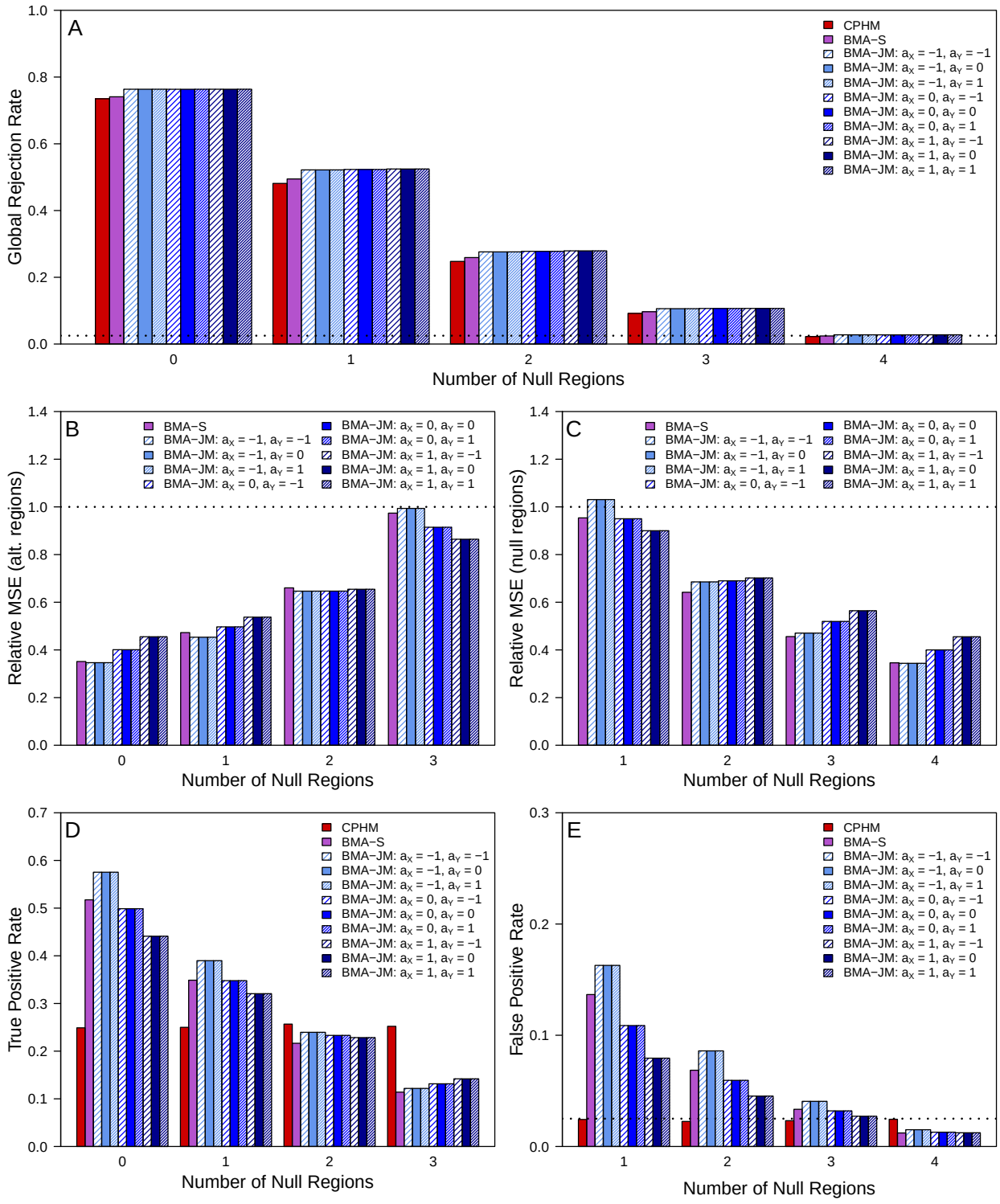

Figure C.8: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for sensitivity analysis with varying values of a_X and a_Y.

C.5.6 Sensitivity Analysis: Change Number of Time Intervals

We follow the same simulation setup, model details, and prior elicitations discussed in Section 4.2 of the main paper, along with the following details:

- Equal regional sample sizes
- Underlying treatment-to-placebo hazard ratio for alternative regions equal to 0.868 (i.e., treatment effect of -0.142)
- Association parameter: $\alpha = 0.5$
- Number of time intervals in piecewise constant baseline hazard: $Q \in \{5, 8, 12\}$

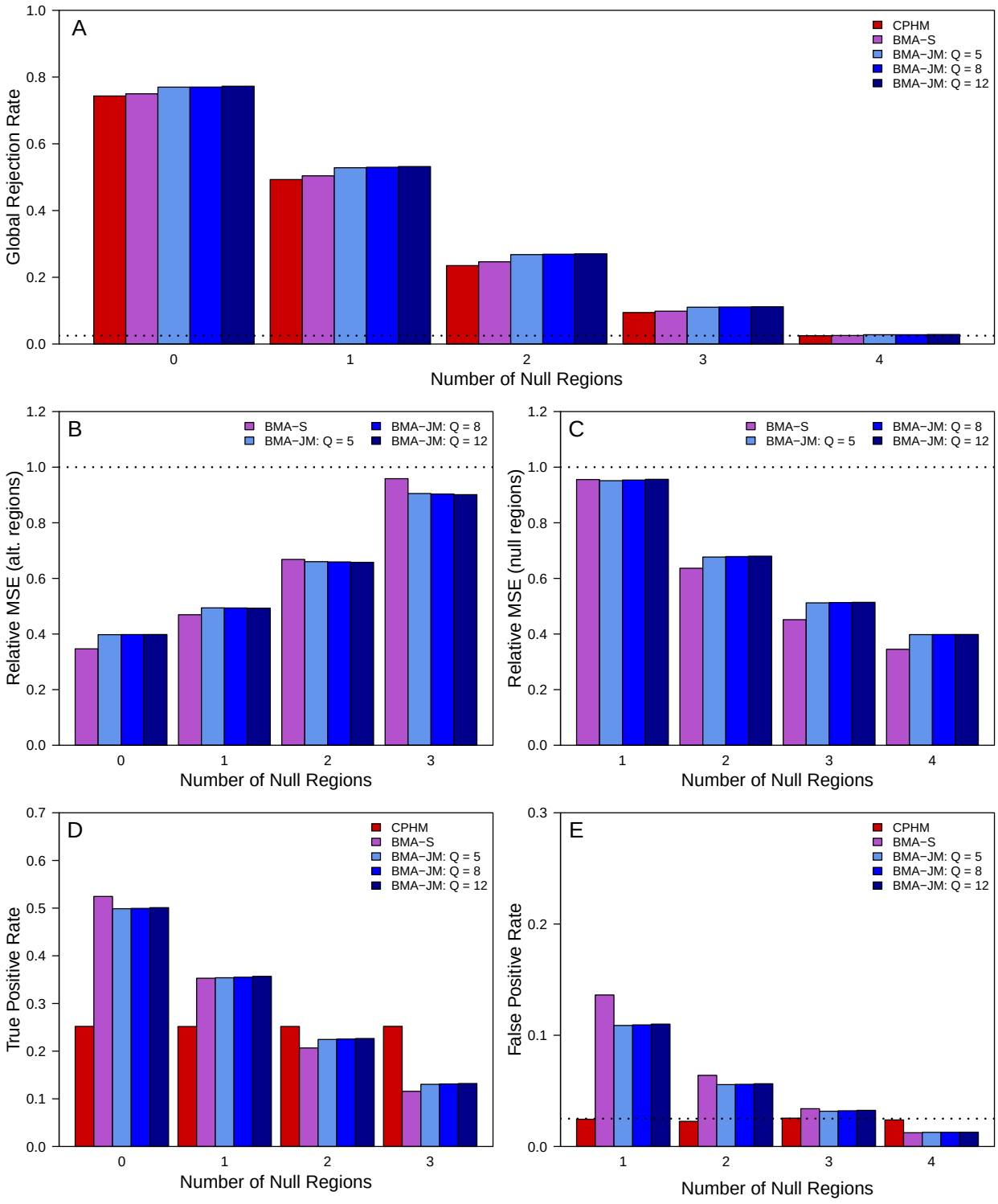

Figure C.9: Global rejection rates (*Panel A*), relative MSE (CPHM as reference) for alternative regions (*Panel B*), relative MSE for null regions (*Panel C*), true positive rates for alternative regions (*Panel D*), and false positive rates for null regions (*Panel E*) for sensitivity analysis with varying values of Q.

C.6 Additional Results for the LEADER Trial Data Analysis

C.6.1 Treatment Effects on the Longitudinal Marker HbA1c

Table C.2: Posterior summary statistics for the global treatment effects $\gamma_{X,G}$ (main effect and treatment-by-time interactions) on the longitudinal marker HbA1c. The knots for the linear splines are at $t \in \{3, 18\}$.

Type of Global Treatment Effect	Posterior Mean	Posterior SD	Probability $P(\gamma_{X,G} < 0 \mid \mathbf{D})$	Probability $P(\gamma_{X,G} > 0 \mid \mathbf{D})$
Main Effect	0.067	0.021	0.325	0.675
Interaction with Time $t > 0$	-0.351	0.009	1.000	0.000
Interaction with Time $t > 3$	0.381	0.010	0.000	1.000
Interaction with Time $t > 18$	-0.022	0.003	1.000	0.000

Table C.3: Posterior summary statistics for the region-specific treatment effects $\gamma_{X,i}$ (main effect and treatment-by-time interactions) on the longitudinal marker HbA1c. The knots for the linear splines are at $t \in \{3, 18\}$.

Region / Type of Treatment Effect	Posterior Mean	Posterior SD	Probability $P(\gamma_{X,i} < 0 \mid \mathbf{D})$	Probability $P(\gamma_{X,i} > 0 \mid \mathbf{D})$
Asia				
Main Effect	0.130	0.020	0.002	0.998
Interaction with Time $t > 0$	-0.402	0.008	1.000	0.000
Interaction with Time $t > 3$	0.434	0.009	0.000	1.000
Interaction with Time $t > 18$	-0.025	0.002	1.000	0.000
Europe				
Main Effect	-0.037	0.024	0.881	0.119
Interaction with Time $t > 0$	-0.269	0.010	1.000	0.000
Interaction with Time $t > 3$	0.295	0.011	0.000	1.000
Interaction with Time $t > 18$	-0.018	0.003	1.000	0.000
North America				
Main Effect	0.123	0.020	0.004	0.996
Interaction with Time $t > 0$	-0.393	0.008	1.000	0.000
Interaction with Time $t > 3$	0.425	0.009	0.000	1.000
Interaction with Time $t > 18$	-0.025	0.002	1.000	0.000
Rest of the World				
Main Effect	0.131	0.020	0.000	1.000
Interaction with Time $t > 0$	-0.404	0.008	1.000	0.000
Interaction with Time $t > 3$	0.436	0.009	0.000	1.000
Interaction with Time $t > 18$	-0.025	0.002	1.000	0.000

C.6.2 Marginal Posterior Model Probabilities by Submodel

Table C.4: Marginal posterior model probabilities (PMPs) for models corresponding to different partitions of regions into sets (survival submodel).

	Set Assignment				Number of	
Submodel	Asia	Europe	North America	Rest of World	Distinct Effects	Marginal PMP
1	1	1	1	1	1	0.218
2	1	1	1	2	2	0.032
3	1	1	2	1	2	0.170
4	1	1	2	2	2	0.114
5	1	2	1	1	2	0.067
6	1	2	1	2	2	0.083
7	1	2	2	1	2	0.073
8	1	2	2	2	2	0.244

Table C.5: Marginal posterior model probabilities (PMPs) for models corresponding to different partitions of regions into sets (longitudinal submodel).

	Set Assignment				Number of	
Submodel	Asia	Europe	North America	Rest of World	Distinct Effects	Marginal PMP
1	1	1	1	1	1	0.000
2	1	1	1	2	2	0.012
3	1	1	2	1	2	0.000
4	1	1	2	2	2	0.003
5	1	2	1	1	2	0.907
6	1	2	1	2	2	0.000
7	1	2	2	1	2	0.078
8	1	2	2	2	2	0.000

REFERENCES

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics*, **20**, 260–279.

Bean, N. W., Ibrahim, J. G., and Psioda, M. A. (2021). Bayesian multiregional clinical trials using model averaging. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxab027.

Bean, N. W., Ibrahim, J. G., and Psioda, M. A. (2022). Bayesian model averaging for multi-regional clinical trials with a time-to-event endpoint. Submitted for publication.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221–228.

Chen, J., Zheng, H., Quan, H., Li, G., Gallo, P., Ouyang, S. P., Binkowitz, B., Ting, N., Tanaka, Y., Luo, X., and Ibia, E. (2013). Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. *Clinical Trials*, **10**, 842–851.

Chi, Y.-Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, **62**, 432–445.

Chiang, C. and Hsiao, C. F. (2019). Use of interval estimations in design and evaluation of multiregional clinical trials with continuous outcomes. *Statistical Methods in Medical Research*, **28**, 2179–2195.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220.

Cunanan, K. M., Iasonos, A., Shen, R., M., and Gonen (2019). Variance prior specification for a basket trial design using Bayesian hierarchical modeling. *Clinical Trials*, **216**, 142–153.

Donohue, J. F. (2005). Minimal clinically important differences in COPD lung function. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, **2**, 111–124.

Esbjerg, S. and Ogenstad, S. (2012). Statistical analysis plan LEADER (Trial ID: EX2211-3748). Technical document, Novo Nordisk.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Biosciences*, **15**, 1663–1685.

FDA (2008). Guidance for industry: Diabetes mellitus—evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. Guidance document, U. S. Food and Drug Administration.

Gabry, J., Ali, I., Brilleman, S., Novik, J. B., AstraZeneca, of Columbia University, T., Wood, S., Team, R. C. D., Bates, D., Maechler, M., and et al. (2022). rstanarm (package version 2.21.3). https://cran.r-project.org/web/packages/rstanarm/.

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

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Gould, A. L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*, **34**, 2181–2195.

Gressani, O. and Lambert, P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics & Data Analysis*, **124**, 151–167.

GSK (2014a). Clinical study report redact (GSK study ID: 200820). Technical document.

GSK (2014b). Protocol (GSK study ID: 200820). Technical document.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics - Theory and Methods*, **19**, 2745–2756.

Hayashi, N. and Itoh, Y. (2017). A re-examination of Japanese sample size calculation for multi-regional clinical trial evaluating survival endpoint. *Japanese Journal of Biometrics*, **38**, 79–92.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors). *Statistical Science*, **14**, 382–417.

Huang, Q., Chen, G., Yuan, Z., and Lan, K. K. G. (2012). Design and sample size considerations for simultaneous global drug development program. *Journal of Biopharmaceutical Statistics*, **22**, 1060–1073.

Huang, W. S., Hung, H. N., Hamasaki, T., and Hsiao, C. F. (2017). Sample size determination for a specific region in multiregional clinical trials with multiple co-primary endpoints. *PLoS One*, **12**.

Hung, H. M. J., Wang, S.-J., and ONeill, R. T. (2010). Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics*, **9**, 173–178.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis (Springer Series in Statistics)*. New York: Springer-Verlag.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine studies. *Statistica Sinica*, **14**, 863–883.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, **28**, 2796–2801.

ICH (1998). E5(R1) Ethnic factors in the acceptability of foreign clinical data. Guidance document, International Council for Harmonisation (ICH).

ICH (2017). E17 General principles for planning and design of multi-regional clinical trials. Guidance document, International Council for Harmonisation (ICH).

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kawai, N., Chuang-Stein, C., Komiyama, O., and Li, Y. (2008). An approach to rationalize partitioning sample size into individual regions in a multiregional trial. *Drug Information Journal*, **42**, 139–147.

Ko, F.-S. (2020). The issue about sample size for survival analysis considering the interaction of unrecognized heterogeneity and treatment. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2020.1864827.

Ko, F. S., Tsou, H. H., Liu, J. P., and Hsiao, C. F. (2010). Sample size determination for a specific region in a multiregional trial. *Journal of Biopharmaceutical Statistics*, **20**, 870–885.

Lan, K. K. G. and Pinheiro, J. (2012). Combined estimation of treatment effects under a discrete random effects model. *Statistics in Biosciences*, **4**, 235–244.

Lan, K. K. G., Pinheiro, J., and Chen, F. (2014). Designing multiregional trials under the discrete random effects model. *Journal of Biopharmaceutical Statistics*, **24**, 415–428.

Li, G., Quan, H., and Lan, G. K. K. (2021). Analysis models for multi-regional clinical trials. In *Simultaneous Global New Drug Development*. Chapman & Hall/CRC.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673–687.

Lininger, L., Gail, M. H., Green, S. B., and Byar, D. P. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika*, **66**, 419–428.

Liu, J. T., Tsou, H. H., Lan, K. K. G., Chen, C. T., Lai, Y. H., Chang, W. J., Chyng-Shyan, T., and Hsiao, C.-F. (2016). Assessing the consistency of the treatment effect under the discrete random effects model in multiregional clinical trials. *Statistics in Medicine*, **35**, 2301–2314.

Mahaffey, K. W., Wojdyla, D. M., Carroll, K., Becker, R. C., Storey, R. F., Angiolillo, D. J., Held, C., Cannon, C. P., James, S., Pieper, K. S., Horrow, J., Harrington, R. A., Wallentin, L., and PLATO Investigators (2011). Ticagrelor compared with clopidogrel by geographic region in the Platelet Inhibition and Patient Outcomes (PLATO) trial. *Circulation*, **124**, 544–554.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F. E., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M., and Buse, J. B. (2016). Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, **375**, 311–322.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, **38**, 514–528.

MHLW (2007). Basic concepts for joint international clinical trials: Notification No.0928010. Guidance document. https://www.pmda.go.jp/files/000157900.pdf.

Muller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, **20**, 260–278.

Niekerk, J. V., Bakka, H., Rue, H., and Schenk, O. (2021). New frontiers in Bayesian modeling using the INLA package in R. *Journal of Statistical Software*, **100**, 1–28.

Nielsen, H. K., DeChiaro, S., and Goldman, B. (2021). Evaluation of consistency of treatment response across regions—the LEADER trial in relation to the ICH E17 guideline. *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2021.662775.

NIH (2014). Study evaluating the efficacy and safety of fluticasone furoate/vilanterol inhalation powder (FF/VI) compared with vilanterol inhalation powder (VI) in subjects with chronic obstructive pulmonary disease (COPD). https://clinicaltrials.gov/ct2/show/NCT02105974.

Pankratz, V. S., de Andrade, M., and Therneau, T. M. (2005). Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. *Genetic Epidemiology*, **28**, 97–109.

Plummer, M., Stukalov, A., and Denwood, M. (2022). rjags (package version 4-13). https://cran.r-project.org/web/packages/rjags/.

Psioda, M. and Alt, E. (2022). bmabasket (package version 0.1.2). https://cran.r-project.org/web/packages/bmabasket/.

Psioda, M. A., Xu, J., Jiang, Q., Ke, C., Yang, Z., and Ibrahim, J. G. (2021). Bayesian adaptive basket trial design using model averaging. *Biostatistics*, **22**, 19–34.

Quan, H., Li, M., Chen, J., Gallo, P., Binkowitz, B., Ibia, E., Tanaka, Y., Ouyang, S. P., Luo, X., Li, G., Menjoge, S., Talerico, S., and Ikeda, K. (2010a). Assessment of consistency of treatment effects in multiregional clinical trials. *Therapeutic Innovation and Regulatory Science*, **44**, 617–632.

Quan, H., Zhao, P. L., Zhang, J., Roessner, M., and Aizawa, K. (2010b). Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. *Pharmaceutical Statistics*, **9**, 100–112.

Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J., and Zhao, P. (2013). Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. *Statistics in Medicine*, **32**, 1691–1706.

Quan, H., Mao, X., Chen, J., Shih, W. J., Ouyang, S. P., Zhang, J., Zhao, P.-L., and Binkowitz, B. (2014). Multi-regional clinical trial design and consistency assessment of treatment effects. *Statistics in Medicine*, **33**, 2191–2205.

Raskin, P. and Mora, P. F. (2010). Glycaemic control with liraglutide: the phase 3 trial programme. *International Journal of Clinical Practice*, **64**, 21–27.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016–1022.

Rizopoulos, D. (2020). JMbayes (package version 0.8-85). https://cran.r-project.org/web/packages/JMbayes/.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*, **71**, 637–654.

Rothmann, M. (2021). Use of Bayesian hierarchical models in the presentation of subgroup analyses. DIA Bayesian Scientific Working Group Webinar Series. https://www.fda.gov/media/154292/download.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

Siler, T. M., Nagai, A., Scott-Wilson, C. A., Midwinter, D. A., and Crim, C. (2017). A randomised, phase III trial of once-daily fluticasone furoate/vilanterol 100/25 μg versus once-daily vilanterol 25 μg to evaluate the contribution on lung function of fluticasone furoate in the combination in patients with COPD. *Respiratory Medicine*, **123**, 8–17.

Sivaganesan, S., Laud, P. W., and Muller, P. (2011). A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine*, **30**, 312–323.

Song, S. Y., Chee, D., and Kim, E. (2019). Strategic inclusion of regions in multiregional clinical trials. *Clinical Trials*, **16**, 98–105.

Sweeting, M. J. and Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, **53**, 750–763.

Tanaka, Y., Mak, C., Burger, B., Ibia, E. O., Rabbia, M., Joshua, J., and et al. (2011). Points to consider in defining region for a multiregional clinical trial. *Therapeutic Innovation & Regulatory Science*, **45**, 575–585.

Teng, Z., Chen, Y. F., and Chang, M. (2017). Unified additional requirement in consideration of regional approval for multiregional clinical trials. *Journal of Biopharmaceutical Statistics*, **27**, 903–917.

Teng, Z., Lin, J., and Zhang, B. (2018). Practical recommendations for regional consistency evaluation in multi-regional clinical trials with different endpoints. *Statistics in Biopharmaceutical Research*, **10**, 50–56.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Tsong, Y., Chang, W. J., Dong, X., and Tsou, H. H. (2012). Assessment of regional treatment effect in a multiregional clinical trial. *Journal of Biopharmaceutical Statistics*, **22**, 1019–1036.

Uesaka, H. (2009). Sample size allocation to regions in a multiregional trial. *Journal of Biopharmaceutical Statistics*, **19**, 580–594.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, **83**, 447–452.

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, **96**, 895–905.

Wolfinger, R. D. and Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed models. *Computational Statistics & Data Analysis*, **25**, 465–490.

Wong, C. H., Siah, K. W., and Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters [published correction appears in Biostatistics. 2019 Apr 1;20(2):366]. *Biostatistics*, **20**, 273–286.

Wu, Y. J., Tan, T. S., Chow, S. C., and Hsiao, C. F. (2014). Sample size estimation of multiregional clinical trials with heterogeneous variability across regions. *Journal of Biopharmaceutical Statistics*, **24**, 254–271.

Xu, Y., Muller, P., Tsimberidou, A. M., and Berry, D. (2019). A nonparametric Bayesian basket trial design. *Biometrical Journal*, **61**, 1160–1174.

Yusuf, S. and Wittes, J. (2016). Interpreting geographic variations in results of randomized, controlled trials. *New England Journal of Medicine*, **375**, 2263–2271.