STATISTICAL METHODS FOR BAYESIAN CLINICAL TRIAL DESIGN AND DNA METHYLATION DECONVOLUTION

Matthew A. Psioda

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.

Chapel Hill
2016

Approved by:

Joseph Ibrahim

Wei Sun

Mengjie Chen

Yun Li

Kathleen Dorsey

ABSTRACT

Matthew A. Psioda: Statistical Methods for Bayesian Clinical Trial Design and DNA Methylation Deconvolution
(Under the direction of Joseph Ibrahim and Wei Sun)

We consider the Bayesian clinical trial design problem in situations where a historical trial is available to inform the design and analysis of a future trial. Currently the FDA requires that all proposed designs exhibit *reasonable* type I error control. Traditionally, frequentist type I error control has been required. This is currently the case in the Center for Drug Evaluation and Research but no longer in the Center for Devices and Radiological Health. The requirement that a design exhibit frequentist type I error control necessitates that all prior information be discarded. We propose several Bayesian solutions that balance the need to control type I errors with the desire to utilize high quality prior information.

For scenarios where the historical trial informs the parameter being tested, we propose Bayesian versions of the type I error rate and power that are defined with respect to the posterior distribution for the parameters given the historical data and conditional on the respective hypothesis being true. We demonstrate that in designs that control the Bayesian type I error rate, meaningful amounts of prior information can be borrowed but that the size of the new trial must be relatively large to justify borrowing a large amount of historical information. We tailor our design methodology for survival applications using proportional hazards and cure rate models. We also develop Bayesian adaptive designs for large cardiovascular outcomes trials (CVOTs) which incorporate control information from a historical CVOT conducted in a similar patient population. We propose an all-or-nothing adaptive design utilizing the power prior as well as a dynamic borrowing adaptive design utilizing a novel extension of the joint power prior.

Separately, we present a statistical deconvolution method for DNA methylation data from bisulfite sequencing experiments. We propose a joint model for methylation data from a set of

heterogeneous tissue samples and another set of reference tissue samples. Unlike other methylation deconvolution methods, our method allows one to estimate the heterogeneous tissue composition and provides improved estimates of cell type-specific methylation levels through the process of deconvolution. We demonstrate our method using data from DNA mixture tissues and simulation studies.

ACKNOWLEDGMENTS

I would first like to thank the three people whose love and support has been indispensable to me throughout my life. To my wife, Ashley, I would like to say thank you for being there for me throughout this journey. There were many challenging times and I surely would have fallen short of my goals without your love and support. You are my rock and my best friend. To my mother, Linda, I would like to say thank you for listening, advising, comforting, and supporting me. I hope my reaching this point make you as proud of me as I am proud to be your son. To my father, Terry, I would like to say thank you for your unfailing example and for teaching me the value of honest hard work.

I would like to thank my advisors, Dr. Joseph Ibrahim and Dr. Wei Sun. I am truly grateful for all of the time they have shared with me over the past five years and for all of the doors they have helped me open. I would like to thank Dr. Mat Soukup for the mentorship given to me during my fellowship with the Food and Drug Administration. I would also like to thank Dr. Mengjie Chen, Dr. Kathleen Dorsey, and Dr. Yun Li for serving on my doctoral committee. I would like to thank the National Cancer Institute for supporting three years of my research through the *Biostatistics for Research in Genomics and Cancer* Training Grant (NCI grant 5T32CA106209-07) and the Food and Drug Administration for supporting one year.

I have made many good friends in the past five years. The closest of which are James Xenakis and Elizabeth Rowley. I would like to thank them both for their companionship, their help in all of our courses, and (most of all) for all the "wedgies".

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In this dissertation we present research on two important but otherwise unrelated topics. The first topic, which is treated with the most depth, is the problem of Bayesian clinical trial design in situations where a previously completed clinical trial (i.e. a historical trial) is available to inform the design and analysis of a future trial. The traditional frequentist approach to clinical trial design requires that the type I error rate be controlled for every possible null value of the parameters. However, such control is impossible if one wishes to incorporate *any* subjective prior information. We propose several Bayesian design solutions that balance the need to control type I errors with the desire to utilize high quality prior information for the purpose of decreasing the size or duration of a future trial.

In Chapters 3 and 4 we propose and apply Bayesian versions of the type I error rate and statistical power in the design of superiority trials for time-to-event data. These Bayesian operational characteristics are defined with respect to the posterior distribution for the parameters given the historical trial data and conditional on the respective hypothesis being true. Chapter 3 focuses on design (i.e. sample size determination) for time-to-event trials where a proportional hazards assumption is tenable. Chapter 4 focuses on designs for time-to-event trials where a fraction of the studied population is "cured" (i.e. immune to having the event).

In Chapter 5 we consider the design of large cardiovascular outcomes trials (CVOTs) utilizing data from control subjects from a previously completed CVOT (i.e. a historical CVOT). Borrowing information only through the control arm presents unique challenges. The benefit of randomization in the new trial is essentially lost and it becomes important to adjust for prognostic factors to help ensure subjects in the two trials are exchangeable. We develop two Bayesian adaptive designs to address a potential lack of exchangeability: an all-or-nothing borrowing approach using the basic power prior (Ibrahim and Chen, 2000) and an adaptive borrowing approach using a novel restricted maximal borrowing power prior which is a special case of the joint power prior (Ibrahim and Chen,

2000).

In Chapter 6 we switch our focus to the second topic: statistical deconvolution of DNA methylation levels. In this chapter we present a deconvolution method for DNA methylation data from bisulfite sequencing experiments. We propose a joint model for methylation data from a set of heterogeneous tissue samples and a set of reference tissue samples. Unlike other deconvolution methods, our methods allows one to estimate the heterogeneous tissue composition and provides improved estimates of cell type-specific methylation levels through the process of deconvolution. Our method allows one to assess methylation at the cell type level whereas commonly used reference-free methods do not. We demonstrate our method using real data from DNA mixture experiments and from simulation studies.

CHAPTER 2: LITERATURE REVIEW

2.1 Bayesian Clinical Trial Design

There is a growing demand for principled ways to incorporate knowledge from previously completed clinical trials in the design and analysis of future trials. Regulatory authorities, such as the Food and Drug Administration (FDA), are increasingly embracing Bayesian methodology for this purpose. For example, in 2010 the FDA issued guidance on using Bayesian methods in medical device clinical trials (Food and Drug Administration, 2010). Recently, additional guidance has been issued that discusses using Bayesian methods to extrapolate information from adult trials to pediatric populations (Food and Drug Administration, 2016). At the Center for Devices and Radiological Health, companies are encouraged to take advantage of good prior information on the safety and effectiveness of their investigational devices through formal Bayesian analysis (Pennello and Thompson, 2007). The FDA sees great promise in using Bayesian methods for incorporation of prior information as well as for conducting adaptive trials (Campbell, 2011).

There is an extensive literature on Bayesian sample size determination methods. Details in book format can be found in Spiegelhalter et al. (2004). Somewhat outdated reviews are given by Adcock (1997) and Pezeshk (2003). Brief summaries of more recent work focusing on traditional Bayesian operating characteristics are as follows:

Wang and Gelfand (2002) formalized a general Bayesian framework that can be used to calculate the required sample size such that certain characteristics of the posterior meet a specified criteria based on an assumed sampling model for the data, sampling prior for the parameters, and fitting prior for the analysis. An example criterion is the average posterior variance criterion (APVC). For this criterion, one seeks a sample size n such that $\mathrm{E}\left[\mathrm{var}\left(h(\boldsymbol{\theta})\middle|\mathbf{y}_n\right)\right] \leq \epsilon$ for some chosen $\epsilon > 0$, where $h(\boldsymbol{\theta})$ is some functional of interest and $\mathbf{y}_n$ is the data comprised of n observations. The

expectation is with respect to the prior predictive distribution for the data defined by the sampling model for the data and sampling prior for the parameters. The authors present two sample size determination examples. One is based on a survival model with censoring and the other is based on a logistic regression model. This simulation-based framework has been widely adopted for Bayesian sample size determination.

Pham-Gia and Turkkan (2003) present a computationally-intensive method to calculate the exact sample size required so that the expected length of the highest posterior density (HPD) interval for the difference in two binomial proportions is sufficiently small. M'Lan et al. (2008) investigate binomial sample size determination using generalized versions of the average length and average coverage criteria, median length and median coverage criteria, and the worst outcome criterion (Joseph et al., 1995). They compare sample sizes derived from highest posterior density intervals and equal-tailed credible intervals. In some cases, they develop closed-form sample size formulas.

De Santis (2004) considers the problem of choosing the sample size for testing hypotheses using Bayes factors. The predictive criterion proposed for determining the sample size is maximizing the probability of obtaining substantial evidence in favor of the true hypotheses (or, equivalently minimizing the probabilities of having either misleading or weak evidence). The method is developed for the normally distributed data and applied to the design of a bladder cancer clinical trial.

Inoue et al. (2005) explore parallels between Bayesian and frequentist methods for determining sample size. The authors provide a simple but general framework for investigating the relationship between the two approaches, based on identifying mappings to connect the Bayesian and frequentist inputs necessary to obtain the same sample size. They highlight a somewhat surprising "approximate functional correspondence" between power-based and information-based optimal sample sizes.

De Santis (2006) presents a robust Bayesian approach to the sample size determination problem based on the lower bound, upper bound, and range of posterior quantities of interest. These characteristics (e.g. the lower bound) are obtained by varying the prior over a class of distributions. Specifically, the authors aim to select an appropriate sample size that guarantees the researcher

will observe a small value of the range and either a sufficiently large lower bound or a sufficiently small upper bound for the posterior quantity of interest.

Clarke and Yuan (2006) give asymptotic expressions for the expected value (under a fixed parameter) of certain types of functionals of the posterior density. They obtain simple inequalities which can be solved to give minimal sample sizes needed for various estimation goals. The authors verify that the asymptotic bounds give good approximations to the expected values of the functionals they approximate.

De Santis (2007) presents a general sample size determination method with a goal of choosing the minimal sample size that guarantees probabilistic control on the performance of quantities that are derived from the posterior distribution and used for inference. The author illustrates how the class of power priors (Ibrahim and Chen, 2000) can be fruitfully employed to deal with lack of homogeneity between historical data and observations of the upcoming experiment. The authors discuss the need for discounting prior information and evaluating the effect of heterogeneity on the optimal sample size. Some of the most popular Bayesian operational characteristics are reviewed and their use is illustrated in several examples.

Brutti et al. (2008) consider determination of a sample size that guarantees a large posterior probability that an unknown parameter of interest (e.g. a treatment effect) is greater than a chosen threshold. The authors argue that a straightforward sample size criterion is to select the minimal number of observations so that the predictive probability of a successful trial is sufficiently large. In the paper the authors address sensitivity to prior assumptions by proposing a robust version of the sample size criterion. Specically, instead of using a single prior distribution, the authors consider a class of plausible priors for the parameter of interest. Robust sample sizes are then selected by looking at the predictive distribution of the lower bound of the posterior probability that the unknown parameter is greater than a chosen threshold. The authors consider classes of ϵ-contamination priors (Berger and Berliner, 1986) for the flexibility and mathematical tractability.

Most of the aforementioned works have focused on computing sample size based on "traditional" Bayesian operating characteristics (i.e. the APVC). As noted above, regulatory bodies such as the FDA require evaluation of type I error rates and power as part of a study design proposal so it

is natural to plan a Bayesian clinical trial where the prior is calibrated to ensure adequate type I error control and power. The idea that these operating characteristics should be considered for Bayesian clinical trial designs is not new (Rubin et al., 1984; Box, 1980). A variety of authors have approached Bayesian sample size determination from the perspective of statistical power and type I error rates with some attempting to incorporate subjective prior information in the design. Brief summaries of selected works are as follows:

Spiegelhalter and Freedman (1986) develop sample size determination method which takes into account prior clinical opinion about the treatment difference. This method adopts the point of clinical equivalence (determined by interviewing the study participants) as the null hypothesis. Decision rules at the end of the study are based on whether the interval estimate of the treatment difference includes the null hypothesis. The prior distribution is used to predict the probabilities of making the decisions to select one of the two treatments or to reserve final judgment. The authors advocate that the sample size be chosen to control the predicted probability of reserving final judgment. An example is presented using normally distributed data from a cancer clinical trial.

Brown et al. (1987) present a sample size determination method that uses the results of a historical study to predict the outcome of a subsequent trial in a setting where the response is binary. The author's method can be easily generalized to other two-sample cases for which power can be calculated in closed-form (e.g. exponential survival), and to one-sample cases such as demonstrating a minimal response rate. Bayesian methods are used to obtain a posterior distribution based on an analysis of the historical study and this posterior is used as the "state of knowledge" for the parameters of interest. The probability that the future study will demonstrate treatment superiority is obtained by using the posterior distribution to compute the average power over the parameter space.

Weiss (1997) presents an approach to sample size determination that aims to make it a priori probable that the Bayes factor will be greater than a given cut-off of prespecified size. Similar to Brown et al. (1987), the approach permits the propagation of uncertainty in quantities which are unknown (i.e. the true effect size) and permits the computation of power and type I error

rates either conditionally (i.e. conditional on the effect size being larger than a certain value) or unconditionally (based on the prior distribution for the effect size). The approach is illustrated for a one-sided and for a two-sided alternative hypothesis for continuous data with a normal prior.

Chen et al. (2011) develop a sample size determination method specifically for the design of non-inferiority clinical trials using subject-level historical control data. The authors compare a hierarchical modeling approach for information borrowing (which necessitates multiple historical studies) with the power prior approach of Ibrahim and Chen (2000). The authors consider use of the basic power prior which fixes the amount of information borrowed a priori as well as the normalized power prior (Duan, 2005; Duan et al., 2006) which attempts to dynamically adjust the amount of information borrowed from the historical study based on the level of agreement between the two data sets. We note that the historical studies only provide information on the control group in the case study and the type I error and power are computed under the implicit assumption that the generative model for the new study data is consistent with the historical data. Hence the type I error control demonstrated by the method is not as stringent as traditional frequentist type I error control. The normalized power prior is a special case of the joint power prior proposed by Ibrahim and Chen (2000) and it was developed to address some shortcomings of the joint power prior that were described in Duan (2005).

Hobbs et al. (2011) propose the commensurate power prior which is another extension of the normalized power prior (Duan, 2005; Duan et al., 2006). The commensurate power prior is quite different from the normalized power prior in that it does not assume a common parameter in the models for the historical and new study. Instead the new study parameter is assumed to be normally distributed about the historical study parameter. The precision parameter in the normal component of the commensurate power prior is called the commensurability parameter and this parameter is incorporated into the prior for the parameter that controls the amount of information borrowed (i.e. the power on the historical study likelihood). The authors compare the frequentist performance of several informative prior choices using simulations and present an example using a colon cancer trial that illustrates the adaptive borrowing approach in the case of a linear model. In general the method does not lead to frequentist type I error control. However, the authors further compare calibrated versions of their commensurate power prior and other informative priors (including the

normalized power prior) where all methods control the type I error rate at the nominal level. The power function with the largest area under the curve was based on the author's commensurate power prior. One draw back of the authors approach is that is requires normally distributed data to apply without approximation. Hobbs et al. (2012) propose extensions to the commensurate power prior and evaluate the frequentist and Bayesian properties using general and generalized linear mixed regression models. The authors provide an example analysis of a colon cancer trial comparing time-to-disease progression using a Weibull regression model. Use of the commensurate power prior (as well as the normalized power prior and joint power prior) are closely related to meta-analysis. Determining a sample size for a future study using these meta-analytic priors is challenging due to the computational burden of fitting models with them. A complete review of the basic power prior and its generalizations can be found in Ibrahim et al. (2015).

Ibrahim et al. (2012b) develop a Bayesian meta-analysis framework using survival regression models to assess whether the size of a clinical development program is adequate to evaluate a particular safety endpoint. Their sample size determination methodology for meta-analysis clinical trial design focuses on controlling the type I error rate and providing adequate statistical power. Using the partial borrowing power prior (i.e. the basic power prior where not all parameters are shared between the new and historical study), the authors incorporate the historical survival meta-data into the statistical design. The authors develop a simulation-based algorithm for estimating relevant operational characteristics of the Bayesian meta-analysis trial design. The proposed methodology is applied to the design of a phase 2/3 development program including a non-inferiority clinical trial for CV risk assessment in type II diabetes mellitus (T2DM). In Chen et al. (2014a) this methodology is extended to a group sequential clinical trial design framework. In both papers the authors show that minimal information can be borrowed without inflating the type I error rate (e.g. that one can borrow $\leq 2\%$ of the available information). The version of type I error rate that the authors attempt to control is essentially the same as in Chen et al. (2011) with the only difference being that information is being borrowed on the parameter being tested. This key difference makes information borrowing essentially impossible.

Chen et al. (2014b) propose a Bayesian approach for the design of superiority clinical trials using recurrent events frailty regression models. Historical recurrent events data from an already

completed trial are incorporated into the design via the partial borrowing power prior. The authors develop a simulation-based algorithm for computing various design quantities such as the type I error rate and power. Similar to Ibrahim et al. (2012b) and Chen et al. (2014a), the results demonstrate that little information can be borrowed if the design is required to control type I error in a traditional sense.

2.2 Bayesian Analysis of Proportional Hazards and Cure Rate Models

Bayesian analysis of proportional hazards and cure rate models has been considered by many authors. We briefly review the relevant literature in this section. A definitive text on Bayesian survival analysis was provided by Ibrahim et al. (2001b). This text includes extensive discussion on Bayesian analysis of proportional hazards models as well as cure rate models.

Bayesian analysis of the Cox partial likelihood (Cox, 1975) dates back to the work of Kalbfleisch (1978) who was the first to give a fully Bayesian justification. A definitive work on Bayesian justifications of the partial likelihood was given by Sinha et al. (2003). Therein the authors establish new (both naive and formal) Bayesian justifications for the Cox partial likelihood and its various modifications. The authors consider cases with time-dependent covariates, time-varying regression parameters, continuous time survival data, and grouped survival data. In addition, the authors present a Bayesian justification of a modified partial likelihood for handling ties.

Ibrahim and Chen (1998) propose a class of informative prior distributions for a proportional hazards model. A novel construction of the prior is developed based on the notion of the availability of historical data and using the power prior. This work discussed using the power prior several years before the work of Ibrahim and Chen (2000), wherein the power prior was formalized. The authors took a semi-parametric approach for modeling the baseline hazard which has become standard for Bayesian proportional hazards models. They also developed an efficient Gibbs sampling framework for model fitting.

Cure rate models date back to Berkson and Gage (1952) where the mixture cure rate model was formalized. The authors develop a simple mixture model for survival data where an unknown fraction of subjects are "cured" and the remaining complement have survival times governed by an

exponential survival distribution. The class of *promotion time* cure rate models are an important subclass of mixture cure rate models. The promotion time cure rate model was originally proposed by Yakovlev et al. (1993). These models have a natural biological motivation based on a latent competing risk framework where metastasis competent tumor cells (that survive treatment) define the competing risks.

Bayesian analysis of promotion time cure rate models was considered first by Chen et al. (1999). The authors provide a thorough discussion of the natural motivation and interpretation of the model and derive several novel properties of it. In particular, the authors show that the model has a proportional hazards structure on the marginal subhazard when the cured fraction is modeled as a function of covariates. They also establish a mathematical connection with the standard mixture cure rate model. They discuss prior elicitation in some detail and propose classes of non-informative and informative prior distributions (the power prior being among them). Several theoretical properties of the proposed priors and resulting posteriors are derived.

Ibrahim et al. (2001a) extend the Bayesian version of the promotion time cure rate model given in Chen et al. (1999). Specifically, the authors propose a semi-parametric cure rate model with a smoothing parameter that controls the degree of parametricity in the right tail of the survival distribution. The authors argue that such a parameter is important for these models and can influence posterior estimates. Several novel properties of the proposed model are derived and a class of informative priors based on historical data is proposed (i.e. the power prior). A case study involving a melanoma clinical trial is discussed in detail to demonstrate the proposed methodology.

Chen et al. (2002a) provide an interesting comparison of several Bayesian models for time-to-event data. They consider a piecewise exponential baseline hazard proportional hazards model, a fully parametric cure rate model, and a semi-parametric cure rate model. For each model, they derive the likelihood function and examine some of its properties for carrying out Bayesian inference with non-informative priors. They also examine model identifiably issues and give conditions which guarantee identifiably. The authors compare the performance of the models based on the conditional predictive ordinate (CPO) which is defined using the posterior predictive distribution for each data point based on the posterior with that data point left out. The authors perform a detailed case

study using the E1690 clinical trial (Kirkwood et al., 2000).

Ibrahim et al. (2012a) consider a joint analysis of the E1684 clinical trial (Kirkwood et al., 1996) and the E1690 clinical trial (Kirkwood et al., 2000) using cure rate models and the power prior. This was an interesting application of Bayesian methods to synthesize information from two clinical trials but it did not address how to design a future trial given information from a previously completed clinical trial.

2.3 DNA Methylation Deconvolution

The concept of DNA methylation deconvolution arises in two types of studies. In studies aimed at identifying associations between DNA methylation and phenotypes of interest (e.g. comparing diseased cases and healthy controls), researchers generally measure methylation levels using heterogeneous tissue samples, such as blood, which can be collected easily. The tissue samples are heterogeneous in the sense that they are comprised of several (possibly many) functionally distinct cell types with each having a potentially distinct DNA methylation signature that may or may not be associated with the disease phenotype. For example, blood contains many different types of white blood cells, including neutrophils, basophils, eosinophils, lymphocytes, and monocytes. In addition to white blood cells, organs such as the liver can contribute significantly to the circulating pool of DNA in blood (Lo et al., 1998). Presumably, each cell type represented in a heterogeneous tissue contributes DNA for the methylation measurement in accordance with its relative abundance in the tissue sample. If the cell types comprising a set of heterogeneous tissue samples have functionally different methylation signatures, as they do in blood (Reinius et al., 2012; Houseman et al., 2012), it is critically important to control for cell type composition in analyses that attempt to establish associations between tissue-level methylation and phenotypes of interest. This fact is increasingly appreciated and it has been argued that many purported associations between age and DNA methylation can be attributed to tissue composition bias that was poorly addressed in the analysis (Jaffe and Irizarry, 2014).

In studies aimed at identifying associations between tissue composition and phenotypes of interest, DNA methylation can be used to facilitate estimation of heterogeneous tissue composition.

In this setting, the DNA methylation levels of the constituent cell types are not of direct interest. Rather, the cell type-specific DNA methylation levels simply offer a means to estimate the heterogeneous tissue's composition. For example, Sun et al. (2015) performed a tissue composition study in hepatocellular carcinoma (HCC) and observed increased levels of liver DNA in plasma from cancer patients compared to healthy controls. Tissue composition studies require estimates for the DNA methylation levels of the constituent cell types. These estimates typically comes from reference tissue samples that are comprised predominately of one cell type though they are generally not perfectly homogeneous (Reinius et al. (2012), supplemental table 2). The reference tissues are constructed through a procedure such as fluorescence activated cell sorting (FACS), which can be laborious. As a result, reference tissues often have few replicates (biological or technical). In the case of bisulfite sequencing data, there are often no replicates of any kind.

Broadly speaking, one may define DNA methylation deconvolution as the process of estimating cell type-specific methylation levels and/or heterogeneous tissue composition using DNA methylation levels measured from heterogeneous tissue samples. The statistical approaches developed for these studies generally fall into two classes: reference-based and reference-free methods. Reference-based methods can be utilized for both types of studies described above whereas reference-free methods are generally only useful for studies attempting to associate tissue-level methylation with phenotypes of interest.

Reference-based methods use a set of reference samples from homogeneous cell type tissues corresponding to the cell types thought to be in the heterogeneous tissues. The reference tissue samples serve as the basis for deconvolution. Houseman et al. (2012) proposed the most widely used method to estimate cell-type composition in blood. The method was designed for the Infinium ®HumanMethylation27 BeadChip and a set of reference samples was published along with the statistical methodology. In the publication, the authors identified a subset of probes that were informative for the white blood cell types and these probes were used in the deconvolution process. Jaffe and Irizarry (2014) extended the method by augmenting the selected probe set based on the Infinium ®HumanMethylation450 BeadChip, which is newer and more popular. The Houseman reference-based method is similar in spirit to a regression calibration method and is based on a set of three linear models. Following Houseman et al. (2012), the model for the reference cell type

tissue methylation levels is

$$\mathbf{Y}_{0h} = \boldsymbol{\beta}_0 \mathbf{w}_{0h} + \boldsymbol{\epsilon}_{0h}$$

where $\mathbf{Y}_{0h}$ is an $m \times 1$ vector of methylation values corresponding to homogeneous cell type tissue sample h, $\mathbf{w}_{0h}$ is a $d_0 \times 1$ ANOVA type covariate vector that identifies the cell type to which a measurement corresponds, $\boldsymbol{\beta}_0$ is a $m \times d_0$ matrix of average methylation values for the d_0 cell types at the m loci, and $\boldsymbol{\epsilon}_{0h}$ is a vector of errors. The model for the heterogeneous tissue methylation levels is

$$\mathbf{Y}_{1i} = \boldsymbol{\beta}_1 \mathbf{z}_{1i} + \boldsymbol{\epsilon}_{1i}$$

where $\mathbf{Y}_{1i}$ is a $m \times 1$ vector of methylation values corresponding to heterogeneous tissue sample i, $\mathbf{z}_{1i}$ is a $d_1 \times 1$ covariate vector, $\boldsymbol{\beta}_1$ is a $m \times d_1$ matrix, and $\boldsymbol{\epsilon}_{1i}$ is a vector of errors. The surrogacy model is given by

$$\boldsymbol{\beta}_1 = \mathbf{1}_m \boldsymbol{\gamma}_0^T + \boldsymbol{\beta}_0 \boldsymbol{\Gamma} + \mathbf{U}$$

where $\boldsymbol{\Gamma}$ is a $d_0 \times d_1$ matrix that summarizes the associations between the rows of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, $\boldsymbol{\gamma}_0$ is a $d_1 \times 1$ vector and $\mathbf{U}$ is a matrix of errors. The author shows that, under some conditions, one can estimate the cell type compositions of the heterogeneous tissues from this model. The key assumption is that $\mathrm{E}\left[\mathbf{Y}_{1i} \mid \mathbf{z}_{1i}\right]$ can be written as

$$\mathrm{E}\left[\mathbf{Y}_{1i} \mid \mathbf{z}_{1i} = \mathbf{z}\right] \quad = \quad \boldsymbol{\xi}^{(\mathbf{z})} + \sum_{k=1}^{d_0} \mathbf{b}_{0,k} \omega_k^{(\mathbf{z})}$$

where $\boldsymbol{\beta}_0 = (\mathbf{b}_{0,1}, \ldots, \mathbf{b}_{0,d_0})$, $\omega_k^{(\mathbf{z})}$ is the cell type k mixture fraction in the tissue, and $\boldsymbol{\xi}^{(\mathbf{z})}$ is an $m \times 1$ vector that is orthogonal to the column space of $\boldsymbol{\beta}_0$. The authors show that the orthogonality assumption can be verified to some degree through sensitivity analysis but that bias results when it does not hold. As noted above, the extension of this method by Jaffe and Irizarry (2014) was simply to select an appropriate set of m probes from the newest Infinium ®BeadChip technology. While Houseman's method attempts to account for the fact that reference samples provide estimates of the

true average cell type methylation levels, it does not produce refined estimates of those methylation levels by incorporating information from the heterogeneous tissues.

In a study of methylation levels in rheumatoid arthritis patients, Liu et al. (2013) attempted to adjust for cell type composition using estimates from the Houseman referenced-based method in a simple linear regression model. Jaffe and Irizarry (2014) demonstrated this may be a poor approach. Jaffe and Irizarry (2014) also consider a surrogate variables approach, named RUV (Gagnon-Bartsch and Speed, 2012), that out performed the naive method of adjustment used by Liu et al. (2013). Application of the RUV method still resulted in significant confounding compared to a gold standard approach based on FACS. The RUV approach attempts to remove unwanted variation (hence the name) based on a set of negative control probes. Note that the RUV approach was not designed to solve the deconvolution problem and so the poor performance is not surprising. Similar approaches to that used by Liu et al. (2013) have been proposed by others. For example, Montaño et al. (2013) and Guintivano et al. (2013) propose methods specifically for brain tissue that use data from the CHARM array (Irizarry et al., 2008).

The second class of statistical methods are reference-free. This class of methods has the same goal as the reference-based methods used by Liu et al. (2013), Montaño et al. (2013), and Guintivano et al. (2013) in that they attempt to provide a valid comparison of tissue-level methylation by adjusting for cell type composition. However, as suggested by their name, reference-free methods do so without requiring reference samples. Houseman et al. (2014) employ a singular value decomposition (SVD) to separate cell-composition effects from the effect of interest. Zou et al. (2014) use a similar approach but the role of dependent and independent variables are reversed. The potential utility of reference-free methods is obvious. They do not require the potentially burdensome step of obtaining reference tissue samples. However, reference-free methods cannot provide information about cell type-specific methylation levels which is ultimately the level of granularity that is of greatest biological importance. Moreover, the reference-free methods are designed to detect so-called direct effects which may not fit the biological process under study. Houseman et al. (2014), propose the model

$$\mathbf{Y} \quad = \quad \mathbf{B}\mathbf{X}^T + \mathbf{M}\boldsymbol{\Omega}^T + \mathbf{E}$$

$$\mathbf{\Omega} \quad = \quad \mathbf{X\Gamma} + \mathbf{\Xi}$$

where $\mathbf{Y}$ is an $m \times n$ matrix of methylation values, $\mathbf{B}$ is an $m \times p$ matrix of direct epigenetic effects, $\mathbf{\Gamma}$ is a $p \times k$ coefficient matrix representing cell-proportion effects, $\mathbf{X}$ is a $n \times p$ matrix of covariates, $\mathbf{M}$ is an $m \times k$ matrix of cell type-specific mean methylation values (falling between 0 and 1), $\mathbf{\Omega}$ is an $n \times k$ matrix of of subject-specific cell proportions, $\mathbf{E}$ is an $m \times n$ matrix of errors, and $\mathbf{\Xi}$ is an $n \times k$ matrix of errors. The authors state that in adjusted epigenome-wide association studies (EWAS), $\mathbf{B}$ is of direct interest. These direct affects can be interpreted as covariate effects that influence methylation levels in all cell types in the heterogeneous tissue equally and hence are not mediated by cell composition. However, since the covariates may also be associated with compositional differences, it is important to correct for tissue composition in the analysis. While such direct affects may indeed exist, we find it more reasonable to posit that covariates (e.g. disease phenotype) may influence the methylation level for one or more cell types in the tissue. Such cell type-specific methylation effects are harder to study but may be more biologically meaningful. The reference-free method proposed by Houseman et al. (2014) was used by Charlton et al. (2014) in the study of Wilms tumors. Houseman et al. (2015) provide an review-like discussion of DNA methylation deconvolution using the Infinium ®BeadChip data.

Recently deconvolution using bisulfite sequencing (BS-seq) data has been studied. This type of data presents unique challenges. For methylation data collected using Infinium ®BeadChip technology, multiple replicates for reference tissues are typically available (e.g. data from Houseman et al. (2012) and Reinius et al. (2012)). However, for BS-seq data, reference tissues generally do not have replication making the method of Houseman et al. (2012) impossible to apply with adaptation.

Sun et al. (2015) use a quadratic programming approach (Van den Meersche et al., 2009) to solve a simultaneous system of equations constructed using methylation estimates from heterogeneous tissue samples using 14 reference tissue samples: liver, lungs, esophagus, heart, pancreas, colon, small intestines, adipose tissues, adrenal glands, brain, T cells, B cells, neutrophils, and placenta. The majority of these data were collected as a part of the NIH Roadmap Epigenomics project (Bernstein et al., 2010) and do not have replicates (biological or technical) although some of the reference tissue samples are pooled from multiple donors. Using these reference tissues the authors

identified a set of 5,820 genomic regions (500bp in length) that were informative of the 14 reference tissue types. For each region the methylation level in each of the samples was estimated as the percentage of all CpGs that were methylated. Then, for heterogeneous tissue sample $s = 1, ..., S$, the following linear system was solved

$$\mathbf{Y}_s \quad = \quad \mathbf{M}\boldsymbol{\rho}_s$$

where $\mathbf{Y}_s$ is a $m \times 1$ vector of estimated methylation levels in heterogeneous tissue sample s at the m genomic regions, $\mathbf{M}$ is the $m \times k$ matrix of estimated methylation levels for the k reference samples considered in a given deconvolution problem (k was generally 5-6, not 14) and $\boldsymbol{\rho}_s$ was the unknown $k \times 1$ vector of tissue composition fractions that satisfied the constraints that $0 \leq \rho_{s,j} \leq 1$ for $j = 1, ..., k$ and $\sum_{j=1}^{k} \rho_{s,j} = 1$. The authors first demonstrated their method was able to recover the composition fractions of DNA mixture tissues. Then, the authors demonstrated their method on a real data set comprised of plasma samples from 15 pregnant women (five from each trimester) and validated their ability to estimates the placental composition in the mother's blood plasma using paternally inherited fetal SNP alleles that were not possessed by the mother.

Ziller et al. (2013) use a similar quadratic programming algorithm to deconvolve *in silico* mixtures of HUES64 and hippocampus whole genome bisulfite sequencing (WGBS) libraries (i.e. heterogeneous tissues constructed by combining reads from sequenced HUES64 and hippocampus libraries). The work of Sun et al. (2015) and Ziller et al. (2013) both focus on statistical deconvolution as a method for estimating the composition of a heterogeneous tissue. The quadratic programming approaches does not take into account the fact that the reference tissues provide estimates of average methylation level for a tissue type and so these approach could suffer from some degree of measurement error bias. The quadratic programming approach used by Sun et al. (2015) and Ziller et al. (2013) has been previously proposed for gene expression data by Gong et al. (2011) and others.

CHAPTER 3: BAYESIAN DESIGN OF A SURVIVAL TRIAL UNDER A PROPORTIONAL HAZARDS ASSUMPTION USING HISTORICAL DATA

3.1 Introduction

There is a growing demand for principled ways to incorporate knowledge from previously completed clinical trials in the design and analysis of future trials. Regulatory authorities, such as the Food and Drug Administration (FDA), are increasingly embracing Bayesian methodology for this purpose. For example, in 2010 the FDA issued guidance on using Bayesian methods in medical device clinical trials (Food and Drug Administration, 2010). Recently additional guidance has been issued that discusses using Bayesian methods to extrapolate information from adult trials to pediatric populations (Food and Drug Administration, 2016).

We consider Bayesian design of a superiority survival trial in a situation where a previously completed clinical trial (i.e. a historical trial) is available to inform the design and analysis of the new trial. During design one must identify appropriate values of the controllable trial characteristics such as the number of subjects to enroll and the number of events required for analysis. An appropriate statistical methodology must also be proposed for hypothesis testing. These choices are made to ensure the trial design has desirable operating characteristics. For confirmatory trials, the primary operating characteristics of interest are the type I error rate and statistical power. Currently the Food and Drug Administration (FDA) requires that all proposed trial designs demonstrate *reasonable* type I error control. Traditionally, frequentist type I error control has been required. This is currently the case for the Center for Drug Evaluation and Research but no longer for the Center for Devices and Radiological Health where fully Bayesian approaches are common. For a design to exhibit Frequentist type I error control, the type I error rate cannot exceed some pre-specified threshold for *any* possible null value of the parameters. The requirement to have frequentist type I error control is not an issue for objective Bayesian designs (i.e. designs utilizing non-informative

priors that are designed to yield good frequentist operating characteristics). In contrast, Bayesian hypothesis testing using informative priors directly conflicts with the frequentist notion of type I error control. As our empirical results will demonstrate, ensuring frequentist type I error control necessitates that all prior information be discarded. See Appendix A.1 for a proof of this result in a sample case.

As a solution to this predicament, we propose a design methodology that yields *Bayesian* type I error control. Unlike the traditional frequentist approach, our Bayesian approach controls a weighted average type I error rate with weights determined by the posterior distribution of the parameters given the historical data and conditional on the null hypothesis being true. Such an approach is sensible when one has pertinent information about the null and alternative hypotheses (in the form of data from a previously completed trial) and when one still wants to protect against type I errors in a equitable way. The Center for Devices and Radiological Health will consider designs that control a Bayesian version of type I error when the historical data are of high quality (Pennello and Thompson, 2007). We demonstrate that in a design that has Bayesian type I error control, meaningful amounts of prior information can be incorporated into the design and analysis of the new trial. However, borrowing the prior information is not free. When the historical data are informative about the parameter being tested, the size of the new trial must be relatively large to justify borrowing a large fraction of the available information. We also introduce a complementary Bayesian version of statistical power, defined as a weighted average statistical power with weights determined by the posterior distribution of the parameters given the historical data and conditional on the alternative hypothesis being true. We demonstrate that designing a trial to have adequate Bayesian power can lead to a much more conservative sample size than designing a trial to have adequate power based on fixed values of the parameters, such as their posterior means given the historical data.

The trial design methodology we present in this chapter is essentially a Bayesian sample size determination method. There is a large literature on Bayesian sample size determination methods. Adcock (1997) gave an early review. More recent contributions include Rahme and Joseph (1998), Rubin and Stern (1998), Katsis and Toman (1999), Simon (1999), Wang and Gelfand (2002), Pham-Gia and Turkkan (2003), Spiegelhalter et al. (2004), Inoue et al. (2005), Clarke and Yuan

(2006), De Santis (2007), M'Lan et al. (2008), Chen et al. (2011), Ibrahim et al. (2012b), Chen et al. (2014a), and Chen et al. (2014b). Much of this literature focuses on simple models including one and two sample normal or binomial models, linear regression models, and generalized linear regression models. Comparatively little has been done with survival models for right-censored data. Notable exceptions are Ibrahim et al. (2012b), Chen et al. (2014a), and Chen et al. (2014b). Bayesian designs which control type type I error in some sense have been recently considered by Chen et al. (2011), Ibrahim et al. (2012b), Chen et al. (2014a), and Chen et al. (2014b). The approach to type I error control considered by these authors is closely related to frequentist type I error control and hence, is quite different from the approach that we propose.

Our design methodology utilizes a stratified proportional hazards regression model with a flexible piecewise constant baseline hazard for each stratum. Information is borrowed from the historical trial by way of the power prior of Ibrahim and Chen (2000). Bayesian analysis of proportional hazards models using the power prior has been considered in Ibrahim and Chen (1998), Ibrahim et al. (2001b), and Chen et al. (2002a). However, our work offers key new insights. We obtain a previously unrecognized closed-form for the posterior distribution of the hazard ratio regression parameters and establish a new connection between this posterior and the weighted Cox partial likelihood (Cox, 1975). By virtue of being able to integrate out the baseline hazard analytically, we are able to develop an efficient simulation-based design methodology that avoids Markov Chain Monte Carlo (MCMC) methods through an accurate normal approximation to the posterior distribution for the hazard ratio regression parameters.

The rest of this chapter is organized as follows: In Section 3.2 we develop a general cumulative hazard model for survival data under a proportional hazards assumption and derive the corresponding likelihood. In Section 3.3 we derive a closed-form for the posterior distribution of the hazard ratio regression parameters (up to normalizing constant) and we discuss a new connection between this posterior and the weighted Cox partial likelihood. In Section 3.4 we formally define our Bayesian versions of type I error and power and discuss the simulation process for determining an appropriate set of controllable trial characteristics for the future trial. Section 3.5 presents a detailed example of our methodology using data from a previously published clinical trial. We close the chapter with some discussion in Section 3.6.

3.2 The Piecewise Constant Hazard Proportional Hazards Model

We consider a flexible proportional hazards model where the baseline hazard is allowed to vary across S levels of a stratification variable. The cumulative hazard for subject i of n is given by

$$\Lambda_i(t) = \Lambda_{[s_i]}(t) \exp\left(\gamma z_i + \boldsymbol{\beta}^T \boldsymbol{x}_i\right) \tag{3.2.1}$$

where s_i is the stratum to which subject i belongs, $\Lambda_{[s]}(t)$ is the baseline cumulative hazard for stratum s, z_i is a binary treatment indicator, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ is a $p \times 1$ vector of baseline covariates, γ is the log hazard ratio for treatment versus control, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is a $p \times 1$ vector of regression coefficients.

Let $\lambda_{[s]}(t)$ denote the piecewise constant baseline hazard for stratum s. We partition the time axis into K_s intervals according to the change points $0 = t_{s,0} < t_{s,1} < .... < t_{s,K_s} = \infty$ and let $\lambda_{sk} > 0$ denote the constant hazard value over interval $I_{s,k} = (t_{s,k-1}, t_{s,k}]$. We denote the set of all baseline hazard parameters by $\boldsymbol{\lambda}$ and all model parameters by $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\beta}, \gamma)$. We refer to this model in its entirety as the piecewise constant baseline hazard proportional hazards model (PWC-PH) model. The PWC-PH model is commonly used in Bayesian survival analysis when the proportional hazards assumption is tenable (Ibrahim et al., 2001b).

Let t_i and c_i be the time-to-event and time-to-censorship for subject i, respectively. We assume c_i is independent of t_i conditional on $(z_i, \boldsymbol{x}_i)$. For subject i, we define $y_i = \min(t_i, c_i)$ to be the observation time, $\nu_i = \mathrm{I}[t_i \leq c_i]$ to be the indicator that an event was observed, ν_{ik} to be the indicator that the event occurred in interval $I_{s_i,k}$, and r_{ik} to be the subject's time at risk in interval $I_{s_i,k}$. We denote the set of indices corresponding to subjects from stratum s by $\mathcal{G}_s$. The likelihood can be written as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{D}) &\propto \prod_{i=1}^{n} \left\{\lambda_{[s_i]}(y_i)\, \phi_i\right\}^{\nu_i} \exp\left[-\Lambda_{[s_i]}(y_i)\, \phi_i\right] \\ &\propto \prod_{s=1}^{S} \prod_{k=1}^{K_s} \lambda_{sk}^{\alpha_{sk}} e^{-\beta_{sk}\lambda_{sk}} \times \prod_{i=1}^{n} \phi_i^{\nu_i} \end{aligned} \tag{3.2.2}$$

where $\alpha_{sk} = \sum_{i \in \mathcal{G}_s} \nu_{ik}$ is the total number of events for subjects from stratum s occurring in

interval $I_{s,k}$, $\phi_i = \exp\left(\gamma z_i + \boldsymbol{\beta}^T \boldsymbol{x}_i\right)$ is the hazard ratio regression function for subject i, $\beta_{sk} = \sum_{i \in \mathcal{G}_s} \phi_i r_{ik}$ is the sum of *scaled* risk times for subjects from stratum s in interval $I_{s,k}$, and $\boldsymbol{D} = \{(y_i, \nu_i, z_i, \boldsymbol{x}_i) : i = 1, ..., n\}$ is the observed data. For each stratum we assume the partition of the time axis is known so that ν_{ik} and r_{ik} are known given y_i and ν_i. Note the factorization of the likelihood in (3.2.2). The left term is a product of $\sum_{s=1}^{S} K_s$ independent gamma kernels (conditional on the hazard ratio regression parameters) and the right term is a function of the hazard ratio regression parameters that does not depend on the baseline hazard. We will exploit this factorization to obtain a closed-form marginal distribution for the hazard ratio regression parameters in Section 3.3.

3.3 The Posterior Distribution under the Basic Power Prior

We consider the case where a single historical trial is available to inform the design and analysis of the future trial. The basic form of the power prior is as follows:

$$\pi_0\left(\boldsymbol{\theta} \middle| \boldsymbol{D}_0, a_0\right) \propto \left[\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{D}_0)\right]^{a_0} \pi_0(\boldsymbol{\theta}) \tag{3.3.1}$$

where $0 \leq a_0 \leq 1$ is a fixed scalar parameter, $\boldsymbol{D}_0$ is the historical data, $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{D}_0)$ is the likelihood for $\boldsymbol{\theta}$ given the historical data, and $\pi_0(\boldsymbol{\theta})$ is an initial prior for $\boldsymbol{\theta}$. When $a_0 = 0$ the historical data is essentially discarded and the power prior reduces to the initial prior. In contrast, when $a_0 = 1$ the power prior corresponds to the posterior distribution from an analysis of the historical data using the initial prior. For intermediate values of a_0 the weight given to the historical data is diminished to some degree leading to a prior for the new trial that is more informative than the initial prior but less informative than using the historical trial posterior as the prior for the new trial.

The power prior is appealing for many reasons, not the least of which is the fact that it provides a semi-automatic mechanism for transforming historical data into a prior for subsequent analyses. One only needs to elicit a value for a_0 for the prior to be fully specified. Often the value of a_0 is determined a priori to ensure the trial design has some desired operating characteristic. In our case, we will maximize a_0 under the constraint that the trial design has Bayesian type I error control. For a complete review of the basic power prior and its generalizations see Ibrahim et al. (2015).

We consider the same PWC-PH model for the historical trial that was given in (3.2.1). For now, we assume that all parameters are shared. We discuss other scenarios in Appendix A.2. A default improper initial prior is obtained by taking

$$\pi_0(\boldsymbol{\theta}) \propto \pi_0(\gamma, \boldsymbol{\beta}) \times \pi_0(\boldsymbol{\lambda})$$

with $\pi_0(\gamma, \boldsymbol{\beta}) \propto 1$ and $\pi_0(\boldsymbol{\lambda}) \propto \prod_{s=1}^{S} \prod_{k=1}^{K_s} \lambda_{sk}^{-1}$. The prior for $\boldsymbol{\lambda}$, when applied to a simple PWC hazard model, is the Jeffreys prior (Jeffreys, 1946). The resulting joint posterior is proper as long as at least one event is observed in every interval for each stratum. A proper alternative is obtained by taking $\pi_0(\gamma, \boldsymbol{\beta}) \propto \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\Sigma}_0 = \sigma_0^2 \cdot \mathbf{I}$ and $\pi_0(\boldsymbol{\lambda}) \propto \prod_{s=1}^{S} \prod_{k=1}^{K_s} \text{Gamma}(\lambda_{sk} \mid \delta_0, \delta_0)$. For the proper priors to be non-informative, one would take σ_0^2 to be large and δ_0 to be small (inverse scale parametrization). In what follows we employ the improper prior, though derivations are similar for the proper prior.

For subject j of n_0 in the historical trial we let s_j be the subject's stratum, $\nu_{j,0}$ be the indicator that an event was observed, $\nu_{jk,0}$ be the indicator that the event occurred in interval $I_{s_j,k}$, and $r_{jk,0}$ be the subject's time at risk in interval $I_{s_j,k}$. We let $\mathcal{G}_{s,0}$ denote the set of indices corresponding to historical trial subjects in stratum s. The power prior can be written as follows:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{D}_0, a_0) \propto \prod_{s=1}^{S} \prod_{k=1}^{K_s} \lambda_{sk}^{\alpha_{sk,0}-1} e^{-\beta_{sk,0}\lambda_{sk}} \times \prod_{j=1}^{n_0} \phi_j^{a_0 \nu_{j,0}} \tag{3.3.2}$$

where $\alpha_{sk,0} = a_0 \sum_{j \in \mathcal{G}_{s,0}} \nu_{jk,0}$, $\beta_{sk,0} = a_0 \sum_{j \in \mathcal{G}_{s,0}} \phi_j r_{jk,0}$, and $\boldsymbol{D}_0$ is the observed data for the historical trial. We refer to $\alpha_{sk,0}$ as the *effective* number of events borrowed from the historical trial. Note that the likelihood factorization that was obtained in (3.2.2) is also exhibited by the power prior.

It is straightforward to show that the posterior distribution has the form

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \quad \propto \quad \pi(\boldsymbol{\lambda} \mid \gamma, \boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{D}_0, a_0) \times \pi(\gamma, \boldsymbol{\beta} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0)$$

where

$$\pi\left(\boldsymbol{\lambda} \mid \gamma, \boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \propto \prod_{s=1}^{S} \prod_{k=1}^{K_s} \text{Gamma}\left(\lambda_{sk} \middle| \alpha_{sk}^*, \beta_{sk}^*\right) \tag{3.3.3}$$

and

$$\pi\left(\gamma, \boldsymbol{\beta} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \propto \prod_{s=1}^{S} \left\{ \frac{\prod_{i \in \mathcal{G}_s} \phi_i^{\nu_i} \prod_{j \in \mathcal{G}_{s,0}} \phi_j^{a_0 \nu_{j,0}}}{\prod_{k=1}^{K_s} \left[\beta_{sk}^*\right]^{\alpha_{sk}^*}} \right\} \tag{3.3.4}$$

with $\alpha_{sk}^* = \alpha_{sk} + \alpha_{sk,0}$ and $\beta_{sk}^* = \beta_{sk} + \beta_{sk,0}$. This result is obtained by combining the gamma kernels from (3.2.2) and (3.3.2), appropriately normalizing each of the resulting conditionally independent gamma kernels, and absorbing the counterbalancing reciprocal terms into the marginal posterior for the hazard ratio regression parameters. The factorization of the joint posterior that we have uncovered is a desirable feature. If one considers $\boldsymbol{\lambda}$ a nuisance parameter then they can work directly with (3.3.4). This is appealing for trial design since the success or failure of the trial depends on inference for γ after integrating out all other parameters.

3.3.1 A Connection with the Weighted Cox Partial Likelihood

We can make an explicit connection between the marginal posterior for the hazard ratio regression parameters under the PWC-PH model and the weighted Cox partial likelihood. Under a particular data-dependent partition of the time axis, the marginal posterior in (3.3.4) is equal to the Breslow approximation of the weighted Cox partial likelihood (Breslow, 1974). For ease of exposition, we take $S = 1$ and omit the notation for the stratification variable. Our arguments trivially extend to the stratified model ($S > 1$).

We partition the time axis into K intervals where K is the number of unique event times from the combined studies and take the upper bound t_k of each interval I_k to be the k^{th} ordered event time. We assume that all subjects not experiencing an event are administratively censored at time t_K and that no other censoring occurs. Thus, all subjects who are at risk in time interval I_k have risk time equal to $t_k - t_{k-1}$. Without loss of generality, we combine the $n + n_0$ subjects from the two studies into a single set indexed by i with weight w_i equal to 1 for new trial subjects and equal to a_0 for historical trial subjects. Let $\mathcal{D}_k$ be the set of indices for the subjects having an event at time t_k and let $\mathcal{R}_k$ be the set of indices for the subjects at risk in time interval I_k. Under these

assumptions, the marginal posterior in (3.3.4) may be reformulated as follows.

$$\begin{aligned}\pi\left(\gamma, \boldsymbol{\beta} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) &\propto \frac{\prod_{i=1}^{n+n_0} \phi_i^{w_i \nu_i}}{\prod_{k=1}^{K}\left(\sum_{i=1}^{n+n_0} w_i \phi_i r_{ik}\right)^{\sum_{i=1}^{n+n_0} w_i \nu_{ik}}} \\ &\propto \prod_{k=1}^{K} \frac{\prod_{i \in \mathcal{D}_k} \phi_i^{w_i}}{\left(\sum_{i \in \mathcal{R}_k} w_i \phi_i\right)^{\sum_{i \in \mathcal{D}_k} w_i}} \end{aligned} \tag{3.3.5}$$

The reader will note that (3.3.5) is the weighted approximate partial likelihood with event time multiplicity handled according to Breslow. If the historical trial is ignored altogether ($a_0 = 0$), the marginal posterior for the hazard ratio parameters is simply the Breslow approximation to the Cox partial likelihood based on the new trial data. Though we may be the first to point out the connection between the marginal posterior for the hazard ratio regression parameters in a PWC-PH model using the power prior to the weighted Cox partial likelihood, the idea that the Cox partial likelihood has a Bayesian justification is not new (Sinha et al., 2003).

In practice, the conditions that we have assumed for the purposes of our derivation will not be met exactly. For example, one might partition the time axis using a small to moderate number of change points rather than one change point per unique event time. In addition, non-administrative censoring may occur. The impact of the time axis partition and non-administrative censoring on the degree of agreement between the marginal posterior for the hazard ratio regression parameters and the weighted Cox partial likelihood is not easy to characterize analytically but our experience and simulations suggest that the agreement will be strong in most practical cases. Thus, Bayesian inference using the PWC-PH model with the power prior will be approximately the same as Bayesian inference using the weighted Cox partial likelihood. We demonstrate this with a simulation study in Appendix A.3.

3.4 Simulation-Based Bayesian Design of a Superiority Study

The null and alternative hypotheses for a superiority trial are as follows:

$$H_0 : \gamma \geq 0 \text{ versus } H_1 : \gamma < 0$$

where γ is the log-hazard ratio for treatment versus control. We accept H_1 if

$$P\left(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) = P\left(H_1 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \tag{3.4.1}$$

is at least as large as some critical value ψ. Both ψ and a_0 are pre-specified constants. During design, we examine various possible values for the number of events required in the new trial (denoted by ν), a_0, and ψ in search of a set of values that yield sufficient Bayesian power while controlling the Bayesian type I error rate at no more than α. We refer to the set of values $\{\nu, a_0, \psi\}$ as the key controllable trial characteristics. We note that increasing of the number of enrolled subjects $n \geq \nu$ will decrease the time required to complete the trial, but will have virtually no impact on the operating characteristics of primary interest.

We will set ψ equal to $1 - \alpha$ and attempt to find the maximum value of a_0 for a given value of ν subject to the Bayesian type I error rate restriction. Our choice of ψ is justified by the fact that the posterior probability $P\left(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0 = 0\right)$ will be approximately uniformly distributed in large samples when $\gamma = 0$ and the model is correct. In other words, the posterior probability of the alternative hypothesis (under no borrowing) has the same asymptotic distribution as a frequentist p-value when the null hypothesis is true. Accordingly, rejecting the null hypothesis when $P\left(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0 = 0\right) \geq \psi = 1 - \alpha$ should provide *frequentist* type I error control (asymptotically). Rigorous exposition on the asymptotic distribution of so-called posterior probabilities of the half-space can be found in Dudley and Haughton (2002). In this light, one can view our design procedure as starting out with a size α frequentist hypothesis test (based on taking $a_0 = 0$) and then modifying the test by borrowing increasing amounts of information from the historical trial until it functions as a size α hypothesis test with respect to the Bayesian type I error rate.

3.4.1 Definition of the Bayesian Type I Error Rate and Power

In order to formally define the Bayesian type I error rate and Bayesian power, we first introduce the concepts of sampling and fitting priors that were formalized by Wang and Gelfand (2002) and extended by Chen et al. (2011) to investigate Bayesian type I error and power. Let $\pi_0^{(s)}\left(\boldsymbol{\theta}\right)$ and $\pi_1^{(s)}\left(\boldsymbol{\theta}\right)$ be the null and alternative sampling priors and let $\pi^{(f)}\left(\boldsymbol{\theta}\right)$ be the fitting prior. A sampling

prior specifies a probability distribution for $\boldsymbol{\theta}$ conditional on a particular hypothesis being true. Hence, the null sampling prior will give zero weight to values of $\boldsymbol{\theta}$ having a negative γ component and the alternative sampling prior will give zero weight to values of $\boldsymbol{\theta}$ having a non-negative γ component. The sampling priors are referred to as such because they are used to sample parameter values in the simulation-based estimation procedure for the Bayesian type I error rate and power. This procedure is detailed in Section 3.4.4. The fitting prior $\pi^{(f)}(\boldsymbol{\theta})$ is simply the prior used to analyze the data. In our case $\pi^{(f)}(\boldsymbol{\theta})$ is the basic power prior given in (3.3.1).

For a fixed value of $\boldsymbol{\theta}$, define the null hypothesis rejection rate as

$$r(\boldsymbol{\theta} \mid \boldsymbol{D}_0, a_0) = \mathrm{E}\left[1\left\{P\left(\gamma < 0 \middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\} \middle| \boldsymbol{\theta}, \boldsymbol{D}_0, a_0\right]$$

where $1\left\{P\left(\gamma < 0 \middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\}$ is an indicator that we accept H_1 based on the posterior probability $P\left(\gamma < 0 \middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right)$ determined by the observed data $\boldsymbol{D}$ (generated from $p(\boldsymbol{D} \mid \boldsymbol{\theta})$), the chosen analysis model (which need not be $p(\boldsymbol{D} \mid \boldsymbol{\theta})$), and the chosen fitting prior. For null values of $\boldsymbol{\theta}$ the quantity $r(\boldsymbol{\theta} \mid \boldsymbol{D}_0, a_0)$ is the type I error rate and for alternative values of $\boldsymbol{\theta}$ it is power. For *chosen* null and alternative sampling priors, the Bayesian type I error rate and Bayesian power, denoted by $\alpha^{(s)}$ and $1 - \beta^{(s)}$, are defined as

$$\alpha^{(s)} = \mathrm{E}_{\pi_0^{(s)}(\boldsymbol{\theta})}\left[r(\boldsymbol{\theta} \mid \boldsymbol{D}_0, a_0)\right] \tag{3.4.2}$$

and

$$1 - \beta^{(s)} = \mathrm{E}_{\pi_1^{(s)}(\boldsymbol{\theta})}\left[r(\boldsymbol{\theta} \mid \boldsymbol{D}_0, a_0)\right]. \tag{3.4.3}$$

The expectation in (3.4.2) is with respect to the null sampling prior distribution for $\boldsymbol{\theta}$ and the expectation in (3.4.3) is with respect to the alternative prior sampling distribution for $\boldsymbol{\theta}$. We note that Chen et al. (cf. equation 5) define the Bayesian the type I error rate and power in terms of the null and alternative prior predictive distribution of the data, $\int p(D \mid \boldsymbol{\theta})\, \pi_0^{(s)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ and $\int p(D \mid \boldsymbol{\theta})\, \pi_1^{(s)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, respectively. Our presentation here simply changes the order of integration ($\boldsymbol{D}$ before $\boldsymbol{\theta}$) to highlight the fact that the Bayesian type I error rate and Bayesian power are weighted averages of the quantities based on fixed values of $\boldsymbol{\theta}$. Our recipe for simulation-based

estimation of the Bayesian type I error rate and Bayesian power follows closely with the presentation in Chen et al.

3.4.2 Default Sampling Priors

It should be clear from Section 3.4.1 that the Bayesian type I error rate and power depend critically on the choice of the null and alternative sampling priors. It is natural that the sampling priors would reflect one's belief about the parameters in light of the information available from the historical trial. Presumably, the historical data will suggest that there is treatment efficacy but the evidence will not be overwhelming (hence the need for and the desire to conduct another clinical trial). We propose a set of default sampling priors that are well suited for this scenario. After collecting the historical data, our posterior belief about the parameters is determined by $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{D}_0\right)$. If one believes the subjects in the historical and future trials are exchangeable, then the most logical choices for the null and alternative sampling priors are $\pi_0^{(s)}\left(\boldsymbol{\theta}\right) = \pi\left(\boldsymbol{\theta} \mid \boldsymbol{D}_0, \gamma \geq 0\right)$ (the historical posterior given that H_0 is true) and $\pi_1^{(s)}\left(\boldsymbol{\theta}\right) = \pi\left(\boldsymbol{\theta} \mid \boldsymbol{D}_0, \gamma < 0\right)$ (the historical posterior given that H_1 true), respectively. Figure 3.1 illustrates the marginal posterior distribution for γ based on our analysis of the historical trial data used in the example application in Section 3.5 along with the corresponding default null and alternative marginal sampling priors for γ. Though we have only plotted the marginal sampling priors for γ, conditioning on the null or alternative hypothesis obviously induces changes in the entire joint distribution for $\boldsymbol{\theta}$. That is to say, the distribution for the baseline hazard parameters and other regression parameters can be quite different for the null and alternative sampling priors. The key point is that by defining the sampling priors as we have done, we preserve the stochastic relationships between the treatment and nuisance parameters that are implied by the historical data under the assumption that a particular hypothesis is true. To make this idea more explicit we contrast it with a commonly used approach that is easier to implement. Consider a point mass alternative sampling prior with all parameters set to their posterior means. For the null sampling prior, the value of γ is simply set to zero with other parameters remaining unchanged. This approach is only sensible when γ is independent of the remaining parameters in $\boldsymbol{\theta}$ given $\boldsymbol{D}_0$. Of course, this is not the case as we will see in Section 3.5.

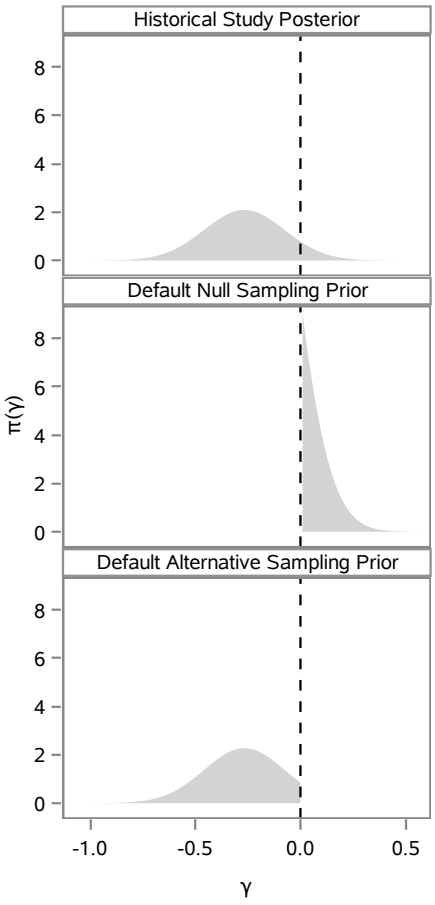

Figure 3.1: $\pi(\gamma \mid \boldsymbol{D}_0)$ and corresponding default marginal sampling priors.

A fundamental challenge in implementing our suggested approach is that the default sampling priors cannot be directly sampled from due to not having closed-forms. However, this issue is easily overcome through the following approximation. First, we fit the model to the historical data using MCMC methods to obtain B samples $\{\boldsymbol{\theta}_b : b = 1, ..., B\}$ from $\pi(\boldsymbol{\theta} \mid \boldsymbol{D}_0)$. We can then approximate sampling from $\pi_0^{(s)}(\boldsymbol{\theta})$ and $\pi_1^{(s)}(\boldsymbol{\theta})$ by sampling with replacement from the sets $\{\boldsymbol{\theta}_b : \gamma_b \geq 0\}$ and $\{\boldsymbol{\theta}_b : \gamma_b < 0\}$, respectively. If the number of sample points in each restricted set is large, the discrete approximations of the sampling priors will be accurate. Obviously the choice of B will depend on how much mass the historical trial posterior puts in the null region. Since this process only needs to be completed once, the computational burden of choosing B to be extremely large (e.g. $5,000,000$) is quite reasonable.

3.4.3 Modifications to the Default Sampling Priors

The default sampling priors previously described are sensible and automatic. However, there will be instances where it is desirable to modify the default priors to incorporate subjective external information. For example, researchers may deem it impossible for the new treatment to cause more than a 10% increased risk relative to the planned control. In addition, researchers may want to compute power over a restricted alternative space that rules out implausibly large or clinically insignificant effect sizes. In this section we introduce intuitive modifications to the default sampling priors that still preserve the stochastic relationships between the treatment and nuisance parameters that are implied by the historical data under the assumption that a particular (restricted) hypothesis is true.

The general null sampling prior is defined as $\pi_0^{(s)}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{D}_0, 0 \leq \gamma \leq \gamma_{0,u})$ and the general alternative sampling prior is defined as $\pi_1^{(s)}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{D}_0, \gamma_{1,l} \leq \gamma < \gamma_{1,u})$. When $\gamma_{0,u} = \infty$, $\gamma_{1,l} = -\infty$, and $\gamma_{1,u} = 0$ the general sampling priors reduce to the default sampling priors. We note that the procedure developed to sample from the default sampling priors applies to the general sampling priors as well. As $\gamma_{0,u} \to 0$, we will see that less and less information can be borrowed from the historical trial if the Bayesian type I error is to be controlled. Note that for $\gamma_{0,u} = 0$, the Bayesian type I error rate is similar to the frequentist type I error rate. Similarly for $\gamma_{1,l} \approx \gamma_{1,u}$ the

Bayesian power is similar to the traditional notion of power. The key difference is that the Bayesian quantities are still averaged with respect to the non-degenerate sampling prior distributions for the nuisance parameters which are defined with respect to the historical trial posterior distribution. Thus, even in the case where $\gamma_{0,u} = 0$, Bayesian type I error control is still not as strict as frequentist type I error control (which requires control of type I error for *any* value of the nuisance parameters).

3.4.4 Estimation of the Bayesian Type I Error Rate and Power

In this section, we describe the simulation process that is used to estimate the Bayesian operating characteristics of interest. Let B be the number of simulation studies to be performed. To estimate the Bayesian type I error rate we proceed as follows:

1. Sample $\boldsymbol{\theta}^{(b)}$ from the null sampling prior $\pi_0^{(s)}(\boldsymbol{\theta})$.
2. Given $\boldsymbol{\theta}^{(b)}$, simulate the new trial data $\boldsymbol{D}^{(b)}$ based on the assumed model for survival times as well as other specified distributions for enrollment, censorship, and covariates.
3. Update the fitting prior $\pi^{(f)}(\boldsymbol{\theta})$ based on the likelihood for the simulated data $\mathcal{L}\left(\boldsymbol{\theta} \mid \boldsymbol{D}^{(b)}\right)$ to obtain the posterior distribution $\pi\left(\gamma, \boldsymbol{\beta} \mid \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)$ and calculate the posterior probability of the alternative hypothesis $P\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)$.
4. Compute the null hypothesis rejection indicator for simulation study b.

$$r^{(b)} = 1\left\{P\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\}$$

5. Estimate the Bayesian type I error rate with the empirical null hypothesis rejection rate.

$$\alpha^{(s)} \approx \frac{1}{B}\sum_{b=1}^{B} r^{(b)}$$

Steps 1-4 are first repeated for $b = 1, ..., B$ to obtain the outcome for each simulation study and then step 5 combines the results to estimate the Bayesian type I error rate. The process for estimating Bayesian power is identical. We simply use the alternative sampling prior in place of the null sampling prior.

3.4.5 Efficient Computation of Posterior Quantities

Simulation-based trial design methods can be computationally demanding due to the need to simulate a large number of datasets and fit models to each one. For Bayesian design problems, the computational burden is substantial since MCMC methods are typically utilized for inference. To alleviate this problem, we employ the Laplace approximation (i.e. multivariate normal approximation) to the marginal posterior for the hazard ratio regression parameters. Using this approximation obviates the need for MCMC altogether thereby providing considerable speed benefits over even the most efficient MCMC approach. The approximation is justified by the fact that the marginal posterior distribution for $(\gamma, \boldsymbol{\beta})$ tends to a multivariate normal as the effective number of events increases. Namely, we have

$$(\gamma, \boldsymbol{\beta} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \dot{\sim} \text{Normal}\left(\begin{bmatrix}\hat{\gamma} \\ \hat{\boldsymbol{\beta}}\end{bmatrix}, \hat{\Sigma} = \begin{bmatrix}\hat{\sigma}_{11}^2 & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12} & \hat{\Sigma}_{22}\end{bmatrix}\right) \tag{3.4.4}$$

where $\left(\hat{\gamma}, \hat{\boldsymbol{\beta}}\right)$ is the posterior mode and $\hat{\Sigma}$ is the inverse negative Hessian matrix for the logarithm of posterior evaluated at $\left(\hat{\gamma}, \hat{\boldsymbol{\beta}}\right)$. Our proposed inference procedure is based on the posterior probability $\text{P}(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0)$. Under (3.4.4), this posterior probability is equal to 1 - $\Phi(\hat{\gamma}/\hat{\sigma}_{11})$ and is readily computed without any sampling once $\hat{\gamma}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\Sigma}$ have been obtained through a Newton-Raphson type procedure. Thus, we replace the computational burden of MCMC with that of optimizing a relatively simple objective function.

To assess the accuracy of the Laplace approximation we performed an array of simulation studies. A discussion of these results can be found in Appendix A.4. The Appendix also contains a description of our customized approach to fitting the model with MCMC, when desired. In summary, our results suggest the difference between the posterior probability computed using the actual marginal posterior for the hazard ratio regression parameters and that computed from the normal approximation is negligible. Accordingly, we can think of no circumstances where one would actually need to utilize MCMC in order to perform inference during *design* with the model we have considered. We note that when performing the *actual* analysis, one will generally use MCMC since it will likely be of interest to characterize the exact posterior distribution for all model parameters

and since the computation burden of MCMC for a single analysis using the PWC-PH model is not great. However, if one is content with the accurate Laplace approximation we have discussed, one can simply draw Monte Carlo samples for $(\gamma, \boldsymbol{\beta})$ from (3.4.4) and then sample $(\boldsymbol{\lambda} \mid \gamma, \boldsymbol{\beta})$ from (3.3.3) to approximate the full posterior. This approach can be combined with rejection sampling to draw approximate samples from the default sampling priors as well.

3.5 Application: Design of a Superiority Study

The E1684 trial was a randomized controlled trial conducted to assess the utility of Interferon Alfa-2b (INF) as an adjuvant therapy following surgery for deep primary or regionally metastatic melanoma. A detailed analysis of the trial was given in Kirkwood et al. (1996). The design and primary analysis were stratified by disease stage according to four groups: (1) deep primary melanomas of Breslow depth more than 4 mm; (2) primary melanomas of any tumor stage in the presence of N1 regional lymph node metastasis detected at elective lymph node dissection with clinically inapparent regional lymph node metastasis; (3) clinically apparent N1 regional lymph node involvement synchronous with primary melanoma of T1-4; and (4) regional lymph node recurrence at any interval after appropriate surgery for primary melanoma of any depth. Subjects treated with the investigational therapy demonstrated a statistically significant prolongation of relapse-free survival compared to those receiving the standard of care (SOC) based on stratified log-rank test ($p = 0.0023$, one-sided). For the purposes of demonstrating the methodology we have presented in this manuscript, we restrict our attention to subjects from the fourth stratum. Based on this group there was evidence of efficacy but the evidence was not overwhelming according to traditional null hypothesis testing criteria.

The number of positive nodes at lymphadenectomy was found to be highly prognostic for relapse-free survival in this disease group and so we considered it as a stratification variable for the future trial (number of positive nodes ≤ 1 (stratum one) versus ≥ 2 (stratum two)). Table 3.1 provides a basic summary of the relapse-free survival data by treatment group and stratum for the targeted risk group. We make the assumption that a model with a treatment effect and stratum-specific baseline hazard holds for these data. No other covariates available appeared to

Table 3.1: Summary survival data for selected E1684 subjects

Treatment	Stratum	N	# Events	Risk Time
SOC	1	37	26	88.4
	2	47	36	105.8
INF	1	43	21	176.3
	2	39	31	81.1

be predictive of relapse-free survival and so we do not include additional covariates in the design model.

We first analyzed the historical trial data using the PWC-PH model with the number of baseline hazard components ranging from one to 12 per stratum and selected the best model according to the deviance information criterion (DIC) (Spiegelhalter et al., 2002). For a given number of components, the baseline hazard change points were chosen to coincide with observed event times such that an approximately equal number of events were observed in each time interval.

Table 3.2 presents the posterior mean, the posterior standard deviation (SD), and 95% highest posterior density (HPD) interval for the treatment effect parameter and for each baseline hazard parameter using the best model according to DIC. The right end points for the chosen time axis partition for each stratum are given in the rightmost column. We note the highest posterior density (HPD) interval for the treatment effect puts mass in both the null region and the alternative region though the evidence clearly favors treatment efficacy. Thus, it is reasonable to assume that if these data were collected in one clinical trial, an additional trial might be conducted where this data could inform the design and analysis. Table 3.3 presents the means and HPDs for all parameters under the default null and alternative sampling priors. Note the changes in the distribution of the baseline hazard parameters when we condition on the null hypothesis being true versus the alternative. This illustrates why it is inappropriate to fix the baseline hazard parameters at, say, their posterior means and simply vary the value of γ during design.

Broadly speaking, we seek a design that controls the Bayesian type I error rate at no more than 2.5% ($\alpha = 0.025$) and that has a Bayesian power of approximately 80%. Our primary purpose in this section is to compare and contrast designs based on different choices of sampling priors. To illustrate how the choice of null sampling prior impacts the amount of historical information that

Table 3.2: Posterior summaries for historical trial

Parm.	Mean	SD	HPD	t_k
γ	-0.267	0.1907	(-0.642,0.106)	n/a
$\lambda_{1,1}$	0.474	0.2172	(0.110,0.906)	0.153
$\lambda_{1,2}$	1.217	0.4756	(0.388,2.161)	0.247
$\lambda_{1,3}$	1.112	0.4342	(0.362,1.978)	0.356
$\lambda_{1,4}$	0.609	0.2562	(0.175,1.122)	0.551
$\lambda_{1,5}$	0.288	0.1323	(0.067,0.551)	0.929
$\lambda_{1,6}$	0.471	0.2164	(0.109,0.900)	1.189
$\lambda_{1,7}$	0.206	0.1056	(0.036,0.415)	1.710
$\lambda_{1,8}$	0.212	0.1097	(0.036,0.429)	2.296
$\lambda_{1,9}$	0.032	0.0167	(0.005,0.065)	∞
$\lambda_{2,1}$	0.874	0.3376	(0.285,1.547)	0.107
$\lambda_{2,2}$	2.529	0.9800	(0.828,4.485)	0.148
$\lambda_{2,3}$	1.544	0.4837	(0.681,2.516)	0.266
$\lambda_{2,4}$	0.914	0.3151	(0.358,1.544)	0.466
$\lambda_{2,5}$	1.140	0.4146	(0.413,1.968)	0.633
$\lambda_{2,6}$	0.418	0.1624	(0.137,0.742)	1.082
$\lambda_{2,7}$	0.262	0.1092	(0.075,0.480)	1.833
$\lambda_{2,8}$	0.234	0.0972	(0.068,0.429)	2.874
$\lambda_{2,9}$	0.088	0.0365	(0.025,0.161)	∞

Table 3.3: Posterior summaries for default sampling priors

	Alternative		Null	
Parm.	Mean	HPD	Mean	HPD
γ	-0.298	(-0.590,0.000)	0.086	(0.000 ,0.236)
$\lambda_{1,1}$	0.481	(0.116,0.919)	0.395	(0.096,0.749)
$\lambda_{1,2}$	1.235	(0.404,2.191)	1.006	(0.331,1.765)
$\lambda_{1,3}$	1.129	(0.363,1.990)	0.918	(0.295,1.603)
$\lambda_{1,4}$	0.618	(0.175,1.130)	0.501	(0.150,0.911)
$\lambda_{1,5}$	0.293	(0.070,0.558)	0.239	(0.059,0.454)
$\lambda_{1,6}$	0.478	(0.113,0.912)	0.386	(0.094,0.733)
$\lambda_{1,7}$	0.210	(0.036,0.420)	0.168	(0.030,0.334)
$\lambda_{1,8}$	0.216	(0.037,0.435)	0.168	(0.030,0.335)
$\lambda_{1,9}$	0.033	(0.005,0.066)	0.025	(0.004,0.050)
$\lambda_{2,1}$	0.885	(0.295,1.565)	0.750	(0.247,1.312)
$\lambda_{2,2}$	2.562	(0.835,4.520)	2.146	(0.732,3.764)
$\lambda_{2,3}$	1.565	(0.681,2.522)	1.307	(0.591,2.098)
$\lambda_{2,4}$	0.926	(0.362,1.554)	0.772	(0.307,1.288)
$\lambda_{2,5}$	1.154	(0.424,1.989)	0.968	(0.351,1.645)
$\lambda_{2,6}$	0.424	(0.139,0.748)	0.354	(0.120,0.624)
$\lambda_{2,7}$	0.265	(0.077,0.484)	0.222	(0.066,0.405)
$\lambda_{2,8}$	0.237	(0.069,0.432)	0.203	(0.061,0.371)
$\lambda_{2,9}$	0.089	(0.026,0.162)	0.078	(0.022,0.140)

can be borrowed, we consider three possibilities: (N1) the null sampling prior obtained by taking $\gamma_{0,u} = 0$, (N2) the null sampling prior obtained by taking $\gamma_{0,u} = \log(1.10)$, and (N3) the default null sampling prior ($\gamma_{0,u} = \infty$). To illustrate how the choice of alternative sampling prior impacts the number of required events to obtain 80% power, we consider two possibilities: (A1) a point mass alternative sampling prior with parameters set to the posterior means in Table 3.2 and (A2) the default alternative sampling prior. The first case corresponds to powering the trial to detect the most likely effect size given the data from the historical trial. Such an approach is commonly used in practice.

For all simulation studies, the generative baseline hazard model for the new trial used the time axis partition from Table 3.2. For a targeted number of events ν, we took $n = 3\nu$ with no censoring other than administrative censoring that occurred when the targeted number of events had been reached. Simulated subjects were allocated to strata in proportions matching the historical trial (i.e. approximately 48% stratum one) and the randomization was 1:1. The PWC-PH model fit to the data did not make use of the true time axis partition. Instead, we took the number of baseline hazard components to be 50 per stratum with change points chosen to coincide with observed event times such that an approximately equal number of events occurred in each time interval within a stratum. In general, the number of baseline hazard components used is fairly inconsequential when interest lies solely in the baseline hazard parameters. We discuss this briefly in Appendix A.3. For all simulation studies, we utilized the normal approximation discussed in Section 3.4.5 to compute the required posterior probability.

For a given null sampling prior, the first step in the design process is to find the value of a_0 that yields a Bayesian type I error rate of 2.5% for each value of ν in the set under consideration. To do this we performed $500,000$ simulations for each ν to estimate the type I error rate over a range of a_0 values, used loess methods to smooth these estimates, and then used linear interpolation to determine the precise value of a_0 that corresponded to the desired error rate.

Figure 3.2 presents loess curves of the Bayesian type I error rate as a function of a_0 for all three null sampling priors for the cases where $\nu = 350$, $\nu = 450$, and $\nu = 550$. It is clear that for these sample sizes, virtually no information can be borrowed from the historical trial under null

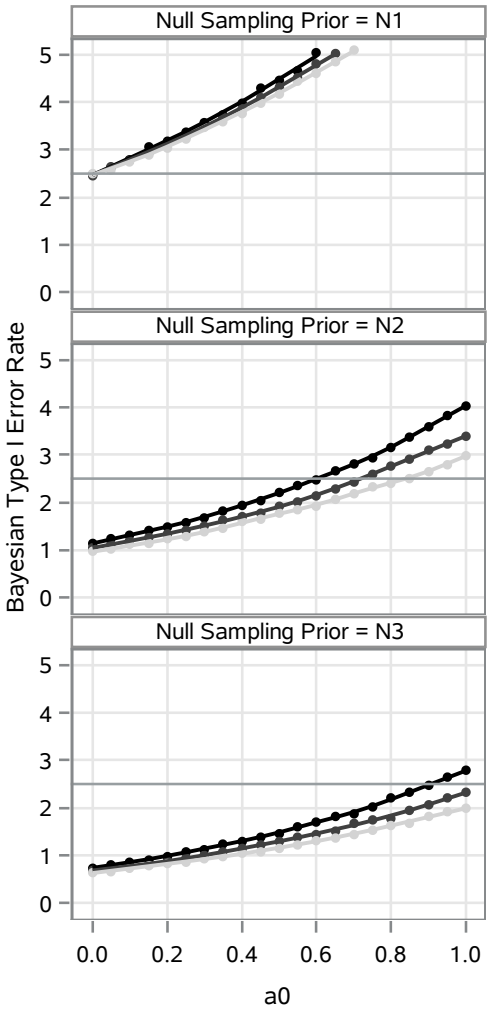

Figure 3.2: Loess curves and point estimates for the Bayesian type I error rate as a function of a_0 for $\nu = 350$ (black), $\nu = 450$ (dark gray), and $\nu = 550$ (light gray). Each curve was estimated from $500,000$ simulation studies performed using values of a_0 from 0 to 1 with a step size of 0.05.

Table 3.4: Bayesian power estimates under various sampling priors

Null Sampling Prior	ν	a_0	Alternative Sampling Prior A1	Alternative Sampling Prior A2
N1	310	0.000	65.0	62.8
	350	0.000	70.2	65.5
	450	0.000	80.7	70.7
	500	0.000	84.7	72.5
	550	0.000	87.8	74.3
	800	0.000	96.5	79.9
N2	310	0.559	75.6	69.1
	350	0.593	80.2	71.5
	450	0.717	89.1	76.4
	500	0.780	91.9	78.4
	550	0.848	94.2	79.9
	800	1.000	98.7	84.7
N3	310	0.849	80.6	72.5
	350	0.906	84.7	74.7
	450	1.000	91.6	78.8
	500	1.000	93.5	80.0
	550	1.000	95.0	81.1
	800	1.000	98.7	84.6

sampling prior N1. In fact, the estimates obtained for a_0 under null sampling prior N1 where never larger than 0.02 and most were approximately zero for every value of ν we considered ($\nu = 250$ to $\nu = 800$). Based on these results, we set $a_0 = 0$ for the purpose of determining power in all design simulations using null sampling prior N1. In contrast, we see that when null sampling prior N2 or N3 is used, we are able to borrow much of and sometimes all of the information in the historical data without surpassing the threshold for the Bayesian type I error rate. Note that a relatively large number of events is required for the future trial to justify borrowing all the information from the historical trial when using the default null sampling prior. Thus, although our Bayesian version of type I error control is less restrictive than frequentist type I error control, there is still significant restriction on the amount of information that can be borrowed. The fundamental difference is that the Bayesian type I error restriction can be overcome by running a future trial of adequate size.

Table 3.4 provides power estimates for each combination of null and alternative sampling priors and for selected values of ν. Each estimate in the table is based on $500,000$ simulation studies. For each value of ν and each null sampling prior, the value of a_0 that was identified in the first stage

of design is also provided in the table. For null sampling priors N2 and N3 we see that the general pattern is that as ν increases, a_0 can also be increased while still controlling the Bayesian type I error rate. If we are permitted to control less restrictive versions of the Bayesian type I error rate (i.e. using null sampling prior N2 or N3 instead of N1), we obtain a Bayesian power of approximately 80% with approximately 250 to 300 fewer events under the default alternative sampling prior (A2) and with approximately 100 to 140 fewer events under the point mass alternative sampling prior (A1).

3.6 Discussion

In this chapter we have developed a Bayesian design methodology for clinical trials with survival endpoints that makes use of data from a historical trial. We have proposed new Bayesian versions of type I error and power that are defined with respect to our belief about the model parameters after observing the historical data. Our results illustrate that if we require frequentist type I error control, it is not possible to borrow meaningful amounts of information from the historical data. However, when we relax the design constraints to require Bayesian type I error control using our default null sampling prior, we are able to borrow substantial amounts of information from the historical trial but full borrowing only comes with a relatively large number of events in the new trial. Thus, while Bayesian type I error control is less restrictive than frequentist type I error control, the former still places significant restrictions on how much information can be borrowed. We have also demonstrated that powering a trial to have adequate Bayesian power under our default alternative sampling prior will generally lead to a much larger number of events in the new trial compared to a point mass alternative sampling prior using the most likely effect size based on the historical data.

The conflict between Bayesian analysis with informative priors and frequentist type I error control is a topic of future research for the authors. In this chapter, our case study focuses on a design application where information is borrowed on the treatment effect as well as nuisance parameters. In this case the apparent conflict between frequentist type I error control and Bayesian analysis with an informative prior is undeniable. However, when one borrows information only on

the nuisance parameters, the conflict is less obvious. This is because the historical trial does not inform the treatment effect parameter and hence, does not overtly suggest with the null hypothesis is false. However, since the sampling priors for the nuisance parameters are defined with respect to the historical data, there is an implicit assumption that the nuisance parameter values in the generative model for the new study are consistent with the historical trial posterior.

Within the proportional hazards framework, there are potential extensions of our methodology that can help protect against unanticipated systematic differences between the historical and new trial subjects beyond what can be addressed by stratification or covariate adjustment (i.e. a lack of exchangeability). We have considered the case where a_0 is determined a priori and fixed to control the Bayesian type I error rate but one can extend the methodology to the case where a_0 is random. When the model includes stratification variables but no covariates, one can implement the normalized power prior (Duan et al., 2006) without approximation. When the baseline risk is allowed to differ between the historical and new studies (i.e. the baseline hazard is not shared), one can also develop a version of the commensurate power prior (Hobbs et al., 2011) through the accurate normal approximation to the marginal posterior for the hazard ratio regression parameters. Both of these extensions attempt to dynamically adjust the amount of information borrowed based on the similarity of the historical and new study data.

CHAPTER 4: BAYESIAN DESIGN OF A SURVIVAL TRIAL WITH A CURED FRACTION USING HISTORICAL DATA

4.1 Introduction

Survival models that accommodate a cured fraction in the studied population, commonly referred to as *cure rate models* (Ibrahim et al., 2001b), have become popular tools for analyzing data from cancer clinical trials. These models have been used for modeling time-to-event data for various types of cancers, including breast cancer (Woods et al., 2009; Tsodikov, 2002), leukemia (Tsodikov et al., 1998), multiple myeloma (Othus et al., 2012), prostate cancer (Zaider et al., 2001), and melanoma (Kirkwood et al., 2000). When a survival curve plateaus in the right tail after an adequate follow-up period, cure rate models can be more advantageous than standard models such as the Cox proportional hazards model or the piecewise exponential model. The horizontal asymptote associated with the aforementioned plateau, often called a cure rate, is an important quantity to model in these settings and using a model that explicitly does so often leads to a better fit compared to standard survival models.

In this paper we develop a Bayesian clinical trial design methodology using the promotion time cure rate model (Yakovlev et al., 1993; Yakovlev, 1996; Chen et al., 1999) in a scenario where a previously completed clinical trial (i.e. a historical trial) is available to inform the design and analysis of the new one. During clinical trial design, one must identify appropriate values of the controllable trial characteristics such as the number of subjects to enroll in the trial. An appropriate statistical methodology must also be proposed for hypothesis testing. These choices are made to ensure the trial design has desirable operating characteristics. For confirmatory trials, the primary operating characteristics of interest are the type I error rate and statistical power. Currently the Food and Drug Administration (FDA) requires that all proposed trial designs demonstrate *reasonable* type I error control. Traditionally, frequentist type I error control has been required.

This is currently the case for the Center for Drug Evaluation and Research but no longer for the Center for Devices and Radiological Health where fully Bayesian designs are becoming common. For a design to exhibit Frequentist type I error control, the type I error rate cannot exceed some pre-specified threshold at the value of the parameter defining the boundary between the null and alternative hypotheses. For unbiased statistical tests, this ensures that the type I error rate is controlled for every possible null value of the parameter. The requirement to have frequentist type I error control is not an issue for objective Bayesian designs (i.e. designs utilizing non-informative priors that are designed to yield good frequentist operating characteristics). In contrast, Bayesian hypothesis testing using informative priors directly conflicts with the frequentist notion of type I error control. Ensuring frequentist type I error control necessitates that all prior information be discarded.

We propose a design methodology that is based on *Bayesian* type I error control. Unlike the traditional frequentist approach, the Bayesian approach controls a weighted average type I error rate with weights determined by the posterior distribution of the parameters given the historical data and conditional on the null hypothesis being true. This design approach is sensible when one has pertinent information about the null and alternative hypotheses (i.e. high-quality data from a previously completed trial) and when one still wants to protect against type I errors in an equitable way. We demonstrate that in a design that exhibits Bayesian type I error control, meaningful amounts of prior information can be incorporated into the design and analysis of the new trial. The Center for Devices and Radiological Health will consider designs that control a Bayesian version of type I error when the historical data are of high quality (Pennello and Thompson, 2007), although the design may still be required to control the frequentist type I error rate at an acceptable level (e.g. twice the nominal Bayesian type I error rate). We also consider designs with this type of added restriction. To complement the Bayesian type I error rate, we introduce a Bayesian version of statistical power, defined as a weighted average statistical power with weights determined by the posterior distribution of the parameters given the historical data and conditional on the alternative hypothesis being true. We demonstrate that designing a trial to have adequate Bayesian power can lead to a much larger sample size than designing a trial to have adequate power based on an optimistic fixed value of the parameters.

The trial design methodology we present in this paper is essentially a Bayesian sample size determination method. There is a large literature on Bayesian sample size determination. Much of it focuses on simple models including one and two sample normal or binomial models, linear regression models, and generalized linear regression models. Comparatively little has been done with survival models for right-censored data. Notable exceptions are Ibrahim et al. (2012b); Chen et al. (2014a,b). Bayesian designs which control type I error in some sense have been recently considered in Chen et al. (2011); Ibrahim et al. (2012b); Chen et al. (2014a,b). The approach to type I error control considered by these authors is closely related to frequentist type I error control and hence is quite different from the approach that we propose here.

For our design methodology, we extend the promotion time cure rate model of Yakovlev et al. (1993) to allow the distribution for the promotion times to vary across levels of a stratification variable. Our results demonstrate that this can be more appropriate than covariate adjustments in the model for the cured fraction which is a more common approach. Information from the historical trial is borrowed by way of the power prior of Ibrahim and Chen (2000). Bayesian analysis of univariate cure rate models has been considered in Chen et al. (1999); Ibrahim et al. (2001a,b); Chen et al. (2002a,b); Tsodikov et al. (2003). Although Bayesian analysis using cure rate models with the power prior has been previously investigated, the work to date has only focused on analysis with no attention being paid to clinical trial design. Frequentist trial design using cure rate models was considered in Bernardo and Ibrahim (2000).

Most Bayesian design methodologies are simulation-based and the model fitting step often requires Markov Chain Monte Carlo (MCMC) methods that can be extremely time consuming. Through a connection between Bayesian analysis with the power prior and weighted maximum likelihood analysis, we develop an asymptotic approximation to the posterior distribution that obviates the need for MCMC in design simulations. By using the approximation, one can simply perform weighted maximum likelihood analysis and use the corresponding asymptotic p-value to approximate the relevant posterior probability for Bayesian hypothesis testing. For reasonably sized trials, this approximation is highly accurate. The positive impact of this very general approximation tool is two-fold. First, since MCMC-based model fitting is replaced by optimization of a low-dimensional objective function, the computational burden of the simulation-based design

procedure is greatly reduced. Second, since standard software is available (e.g. SAS® or R) for weighted maximum likelihood analysis, implementing our method does not require extensive custom programming.

The rest of this article is organized as follows: In Section 4.2 we develop a stratified promotion time cure rate model, discuss some properties, and derive the corresponding likelihood. In Section 4.3 we discuss the basic power prior and the approximation to the posterior that obviates the need for MCMC. In Section 4.4 we formally define Bayesian versions of type I error and power and discuss the simulation process for determining an appropriate set of controllable characteristics for the new trial. In Section 4.5 we present a detailed example design using data from a previously published clinical trial. We close the paper with some discussion in Section 4.6.

4.2 The Promotion Time Cure Rate Model

We consider a flexible promotion time cure rate model where the promotion time distribution is allowed to vary over levels of a stratification variable. The unstratified version was proposed originally by Yakovlev et al. (1993) and a thorough treatment from the Bayesian perspective was first given in Chen et al. (1999). The model is typically motivated through a latent competing risks framework. Using notation similar to Ibrahim et al. (2001b), we let N_i denote the number of *metastasis-competent* tumor cells for subject i that remain after initial treatment. We assume that N_i follows a Poisson distribution with parameter $\theta_i = \exp\left(\gamma z_i + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)$ where z_i is a binary treatment indicator, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ is a $p \times 1$ vector of baseline covariates that includes an intercept, γ is the treatment effect, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is a $p \times 1$ vector of regression coefficients corresponding to the covariates. Further let Z_{ij} denote the random time for the j^{th} metastasis-competent tumor cell to produce detectable disease in subject i. Hence, one can view Z_{ij} as the "promotion time" for the j^{th} metastasis-competent tumor cell. Conditional on N_i, the Z_{ij} are assumed to be independent and identically distributed according to the cumulative distribution function $F\left(z \mid \boldsymbol{\psi}_{s_i}\right) = 1 - S\left(z \mid \boldsymbol{\psi}_{s_i}\right)$ where s_i is the stratum to which subject i belongs and $\boldsymbol{\psi}_s$ represents the promotion time model parameters for stratum s. The time to detectable cancer relapse for subject i is given by $Y_i = \min\left\{Z_{ij}, 0 \leq j \leq N_i\right\}$ where $Z_{i0} = \infty$. Suppressing the

notation for covariates, the marginal probability of survival past time y for subject i is given as follows:

$$\begin{aligned} S_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) &= P(N_i = 0 \mid \theta_i) + P(Y_i > y \mid N_i \geq 1, \theta_i, \boldsymbol{\psi}_{s_i}) \\ &= \exp(-\theta_i) + \sum_{k=1}^{\infty} S\left(y \mid \boldsymbol{\psi}_{s_i}\right)^k \exp(-\theta_i) \frac{\theta_i^k}{k!} \\ &= \exp\left(-\theta_i F\left(y \mid \boldsymbol{\psi}_{s_i}\right)\right) \end{aligned} \tag{4.2.1}$$

The quantity in (4.2.1) is a marginal probability in the sense that it is not conditional on the latent number of metastasis-competent tumor cells or even whether or not the subject is cured. We note that $S_p\left(\infty \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) = P(N_i = 0 \mid \theta_i) = \exp(-\theta_i) > 0$ and hence $S_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right)$ is not a proper survival function. The form in (4.2.1) shows that the time to relapse is influenced by the initial number of metastasis-competent tumor cells as well as their rate of progression. The subdensity corresponding to (4.2.1) is given by

$$f_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) = \theta_i f\left(y \mid \boldsymbol{\psi}_{s_i}\right) \exp\left(-\theta_i F\left(y \mid \boldsymbol{\psi}_{s_i}\right)\right)$$

with corresponding subhazard given by

$$h_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) = \theta_i f\left(y \mid \boldsymbol{\psi}_{s_i}\right). \tag{4.2.2}$$

From (4.2.2), it is apparent that the promotion time formulation of the cure rate model leads to a proportional hazards structure for the subhazard. The probability of survival past time y conditional on subject i being uncured is given by

$$\begin{aligned} S^*\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) &= P(Y_i > y \mid N_i \geq 1, \theta_i, \boldsymbol{\psi}_{s_i}) \\ &= \frac{S_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) - \exp(-\theta_i)}{1 - \exp(-\theta_i)}, \end{aligned} \tag{4.2.3}$$

and the corresponding hazard is

$$h^*\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) = \frac{1}{P(Y_i < \infty \mid Y_i > y, \theta_i, \boldsymbol{\psi}_{s_i})} h\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) \tag{4.2.4}$$

where

$$P(Y_i < \infty \mid Y_i > y, \theta_i, \boldsymbol{\psi}_{s_i}) = \frac{S_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) - \exp(-\theta_i)}{S_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right)}.$$

We note that $S^*\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right)$ is a proper survival function and, accordingly, $h^*\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right)$ is a proper hazard function. Unfortunately (4.2.4) does not have a proportional hazards structure since $P(Y_i < \infty \mid Y_i > y, \theta_i, \boldsymbol{\psi}_{s_i})$ depends on y. It is straight forward to show that $h^*\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right)$ is increasing in θ_i which is desirable. This means that increasingly negative values of the regression parameters are associated with a greater cured fraction and lower hazard in the uncured population.

As pointed out in Chen et al. (1999), the promotion time cure rate model has a connection with the standard mixture cure rate model of Berkson and Gage (1952). It can be readily seen from (4.2.3) that

$$S_p\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right) \quad = \quad \exp(-\theta_i) + (1 - \exp(-\theta_i))\, S^*\left(y \mid \theta_i, \boldsymbol{\psi}_{s_i}\right).$$

Thus, the promotion time cure rate model is a special case of the standard mixture cure rate model. The promotion time cure rate model may be preferred over the standard mixture cure rate model for several reasons as noted in Chen et al. (1999). First, the model has the natural biological motivation described above. Second, the model has a proportional hazards structure leading to convenient interpretation of covariate effects on the subhazard. The standard cure rate model does not have a proportional hazards structure when the cured fraction is modeled as a function of covariates using a logistic regression function. Third, the model can be efficiently sampled with a Gibbs sampler. Lastly, unlike the standard mixture cure rate model, the promotion time cure rate model yields a proper posterior distribution under a wide class of non-informative improper priors for the regression coefficients, including a uniform improper prior.

To complete the specification of the survival model in (4.2.1), we must specify a distribution for the promotion times. Common choices are the Weibull distribution (fully parametric) and piecewise exponential distribution (semi-parametric). Analysis with each of these promotion time models is discussed in detail in Ibrahim et al. (2001b) for the case where a single promotion time model is shared by all subjects. For the design example in Section 5, we utilize a separate Weibull model for each level of the stratification variable. We note that our choice to allow stratification in the

model for the promotion times is uncommon. However, the design model selection results discussed in Section 5 illustrate that this approach can lead to better fit compared to the more standard modeling framework.

Following Ibrahim et al. (2001b), the *complete* data likelihood can be written as follows

$$\mathcal{L}\left(\boldsymbol{\xi}, \boldsymbol{N} \mid \boldsymbol{D}\right) = \prod_{i=1}^{n} S\left(y_i \mid \psi_{s_i}\right)^{N_i - v_i} \left(N_i f\left(y_i \mid \psi_{s_i}\right)\right)^{v_i} \frac{e^{-\theta_i} \theta_i^{N_i}}{N_i!}$$

where $\boldsymbol{\xi} = \{\gamma, \boldsymbol{\beta}, \boldsymbol{\psi}_s : s = 1, ..., S\}$ is the set of all parameters in the model, $\psi_s = \{\lambda_s, \alpha_s\}$ is the set of Weibull promotion time model parameters for stratum s, and $\boldsymbol{D} = \{(y_i, v_i, z_i, \boldsymbol{x}_i, s_i) : i = 1, ..., n\}$ is the observed data. In the literature, Bayesian analysis of the promotion time cure rate model has been exclusively performed using the complete data likelihood with the N_i being treated as missing data and therefore included in the Gibbs sampler with the parameters. This approach was proposed in Chen et al. (1999) and described in full detail in Ibrahim et al. (2001b). The benefit of such an approach is that the full conditionals for all parameters are log-concave (based on the priors discussed in Ibrahim et al. (2001b)) and so the parameters can be easily sampled with rejection sampling or adaptive rejection sampling methods (W. R. Gilks, 1992). The N_i have closed-form full conditionals for direct Poisson sampling. Alternatively, one can analytically sum out the latent N_i variables to obtain the *observed* data likelihood

$$\mathcal{L}\left(\boldsymbol{\xi} \mid \boldsymbol{D}\right) = \prod_{i=1}^{n} \left[\theta_i f\left(y_i \mid \psi_{s_i}\right)\right]^{v_i} \exp\left\{-\theta_i F\left(y_i \mid \psi_{s_i}\right)\right\}. \tag{4.2.5}$$

For MCMC analysis based on the observed data likelihood, the regression parameters still have log-concave full conditionals and they can be sampled efficiently with the same techniques mentioned above. Unfortunately the full conditionals for the parameters in the promotion time model will not necessarily have log-concave full conditionals and so we recommend slice sampling (Neal, 2003) for those parameters. Even though slice sampling does not directly sample from the full conditionals, since the sampling procedure using the marginal likelihood does not condition on the N_i quantities, this approach is likely more efficient than the Gibbs sampler using the complete data likelihood. In our analyses, when MCMC was used, we fit the model using the observed data likelihood as just described.

4.3 The Basic Power Prior and the Posterior Distribution

To simplify our exposition, we focus on a scenario with a single historical trial to be used for design. The form of the basic power prior (Ibrahim and Chen, 2000) using the observed data likelihood as given in (4.2.5) is as follows:

$$\pi_0\left(\boldsymbol{\xi}|\boldsymbol{D}_0, a_0\right) \propto \left[\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{D}_0)\right]^{a_0} \pi_0(\boldsymbol{\xi}) \tag{4.3.1}$$

where $0 \leq a_0 \leq 1$ is a fixed scalar parameter, $\boldsymbol{D}_0 = \{(y_j, v_j, z_j, \boldsymbol{x}_j, s_j) : j = 1, ..., n_0\}$ is the historical data, $\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{D}_0)$ is the likelihood for the historical data, and $\pi_0(\boldsymbol{\xi})$ is an initial non-informative prior. When $a_0 = 0$ the historical data is essentially discarded and the power prior reduces to the initial prior. In contrast, when $a_0 = 1$, the power prior corresponds to the posterior distribution from an analysis of the historical data using the initial prior. For intermediate values of a_0, the weight given to the historical data is diminished to some degree leading to a prior that is more informative than the initial prior but less informative than using the historical trial posterior as the prior for the new trial.

The power prior is appealing for many reasons. We mention the two most relevant properties to our discussion and refer the interested reader to Ibrahim et al. (2015) for a complete review. First, the basic power prior provides a semi-automatic mechanism for transforming historical data into a subjective prior for design and analysis of a future trial. One only needs to specify the initial prior $\pi_0(\boldsymbol{\xi})$ and elicit a value for a_0 for the prior to be fully specified. In our case, the value of a_0 will be chosen so that the design yields desirable Bayesian power while controlling the Bayesian type I error rate. As an alternative to searching for a fixed value of a_0 that leads to control of the Bayesian type I error rate, one might model a_0 as a random variable in an effort to dynamically adjust the amount of information borrowed based on the degree of similarity between the historical data and new trial data at the time of the analysis. Such an approach is a step away from a design-based solution and a step towards meta-analysis. Several extensions of the basic power prior have been proposed for this type of approach and all are discussed in Ibrahim et al. (2015) and the references therein. We only consider the fixed a_0 approach in this paper.

The second appealing characteristic of the basic power prior is that analysis using it with a non-informative initial prior is closely related to weighted maximum likelihood where historical trial subjects are given a weight of a_0 and new trial subjects are given a weight of one. To see this connection, note that the logarithm of the posterior (ignoring the normalizing constant) is given by

$$\begin{aligned}
\log \pi\left(\boldsymbol{\xi} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) &= \log \mathcal{L}(\boldsymbol{\xi}|\boldsymbol{D}) + a_0 \log\left[\mathcal{L}(\boldsymbol{\xi}|\boldsymbol{D}_0)\right] + \log \pi_0(\boldsymbol{\xi}) \\
&= \sum_{i=1}^{n} w_i \left[v_i \left\{\log \theta_i + \log f\left(y_i \mid \lambda_{s_i}, \alpha_{s_i}\right)\right\} - \theta_i F\left(y_i \mid \lambda_{s_i}, \alpha_{s_i}\right)\right] \\
&+ \sum_{j=1}^{n_0} w_{0,j} \left[v_j \left\{\log \theta_j + \log f\left(y_j \mid \lambda_{s_j}, \alpha_{s_j}\right)\right\} - \theta_j F\left(y_j \mid \lambda_{s_j}, \alpha_{s_j}\right)\right] \\
&+ \log \pi_0(\boldsymbol{\xi})
\end{aligned}$$

which is *approximately* equal to the weighted log-likelihood based on the combined studies with $w_i = 1$ for new trial subject i and $w_{0,j} = a_0$ for historical trial subject j. The only difference between the logarithm of the posterior distribution and the weighted log-likelihood is the term $\log \pi_0(\boldsymbol{\xi})$ which has little influence since $\pi_0(\boldsymbol{\xi})$ is non-informative by construction. The Bayesian central limit theorem assures us that, when the effective sample size for the combined trials is reasonably large,

$$\pi\left(\gamma \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \dot{\propto} \text{Normal}\left(\gamma \mid \hat{\gamma}, \sigma^2_{\hat{\gamma}}\right)$$

where $\hat{\gamma}$ is the weighted maximum likelihood estimator (MLE) from a joint analysis of both trials with weights described above and $\sigma^2_{\hat{\gamma}}$ is the relevant diagonal element of the inverse of the observed information matrix for the weighted log-likelihood evaluated at the weighted MLE. Using this connection we can accurately approximate relevant posterior probabilities for our Bayesian analyses using results that are readily obtainable from standard software using weighted maximum likelihood methods. To illustrate this, consider the the null and alternative hypotheses $H_0 : \gamma \geq 0$ and $H_1 : \gamma < 0$, respectively. We can approximate the posterior probability of the alternative hypothesis as follows:

$$P\left(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \approx P\left(Z \leq -\frac{\hat{\gamma}}{\sigma_{\hat{\gamma}}} \middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) = 1 - \Phi\left(\frac{\hat{\gamma}}{\sigma_{\hat{\gamma}}}\right), \tag{4.3.2}$$

where Z is a standard normal variable. We note that the right hand side of (4.3.2) is one minus the one-sided p-value that arises from weighted maximum likelihood analysis of the combined studies.

4.4 Simulation-Based Bayesian Design of a Superiority Trial

The null and alternative hypotheses for a superiority trial are $H_0 : \gamma \geq 0$ and $H_1 : \gamma < 0$, respectively. We will accept H_1 if $P(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0)$ is at least as large as some critical value ψ. During design, one examines various possible values for the number of subjects enrolled in the new trial (n), the duration of the new trial (T), a_0, and ψ in search of a set of values that yield sufficient Bayesian power while controlling the Bayesian type I error rate at no more than α. We refer to the set of values $\{n, T, a_0, \psi\}$ as the key controllable trial characteristics.

In general, T should be at least as large as the duration of time it is expected to take for the survival curves to plateau. Thus, it is natural to fix T for design purposes based on the time when the survival curves approximately leveled off in the historical trial. This is the approach we suggest in practice and the approach taken in our example in Section 4.5. We further restrict the search space for the key controllable trial characteristics by fixing $\psi = 1 - \alpha$. This choice of ψ is justified by the fact that the posterior probability $P(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0 = 0)$ (based on an analysis of the new trial without incorporating historical data) will be asymptotically uniformly distributed when $\gamma = 0$ and the model is correct. In other words, the posterior probability of the alternative hypothesis has the same asymptotic behavior as a frequentist p-value when the null hypothesis is true. Accordingly, rejecting the null hypothesis when $P(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0 = 0) \geq \psi = 1 - \alpha$ should provide *frequentist* type I error control (asymptotically). A heuristic justification for these ideas follows directly from the relationship in (4.3.2) but more rigorous exposition on the connection between so called posterior probabilities of the half-space and frequentist p-values can be found in Dudley and Haughton (2002). In this light, one can view our design procedure as starting out with a size α frequentist hypothesis test (based on taking $a_0 = 0$) and then modifying the test by borrowing increasing amounts of information from the historical trial until it functions as a size α hypothesis test with respect to the Bayesian type I error rate.

4.4.1 Definition of the Bayesian Type I Error Rate and Power

In order to formally define the Bayesian type I error rate and Bayesian power, we first introduce the concepts of sampling and fitting priors that were formalized in Wang and Gelfand (2002) and extended in Chen et al. (2011) to investigate Bayesian type I error and power. Let $\pi_0^{(s)}(\boldsymbol{\xi})$ and $\pi_1^{(s)}(\boldsymbol{\xi})$ be the null and alternative sampling priors and let $\pi^{(f)}(\boldsymbol{\xi})$ be the fitting prior. A sampling prior specifies a probability distribution for the model parameters conditional on a particular hypothesis being true. The null sampling prior will give zero weight to values of $\boldsymbol{\xi}$ having a negative γ component and the alternative sampling prior will give zero weight to values of $\boldsymbol{\xi}$ having a non-negative γ component. The sampling priors are referred to as such because they are used to sample parameter values in the simulation-based estimation procedure for the Bayesian type I error rate and power. This procedure is detailed in Section 4.4.4. The fitting prior $\pi^{(f)}(\boldsymbol{\xi})$ is simply the prior used to analyze the data. In our case $\pi^{(f)}(\boldsymbol{\xi})$ is the basic power prior given in (4.3.1).

For a fixed value of $\boldsymbol{\xi}$, define the null hypothesis rejection rate as

$$r(\boldsymbol{\xi} \mid \boldsymbol{D}_0, a_0) = \mathrm{E}\left[1\left\{P(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \geq \psi\right\} \mid \boldsymbol{\xi}, \boldsymbol{D}_0, a_0\right]$$

where $1\left\{P(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \geq \psi\right\}$ is an indicator that we accept H_1 based on the posterior probability $P(\gamma < 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0)$ determined by the observed data $\boldsymbol{D}$. For null values of $\boldsymbol{\xi}$, the quantity $r(\boldsymbol{\xi} \mid \boldsymbol{D}_0, a_0)$ is the type I error rate and, for alternative values of $\boldsymbol{\xi}$, it is power. For *chosen* null and alternative sampling priors, the Bayesian type I error rate $\alpha^{(s)}$ and Bayesian power $1 - \beta^{(s)}$ are defined as

$$\alpha^{(s)} = \mathrm{E}_{\pi_0^{(s)}(\boldsymbol{\xi})}\left[r(\boldsymbol{\xi} \mid \boldsymbol{D}_0, a_0)\right] \tag{4.4.1}$$

and

$$1 - \beta^{(s)} = \mathrm{E}_{\pi_1^{(s)}(\boldsymbol{\xi})}\left[r(\boldsymbol{\xi} \mid \boldsymbol{D}_0, a_0)\right]. \tag{4.4.2}$$

The expectation in (4.4.1) is with respect to the null sampling prior distribution for $\boldsymbol{\xi}$ and the expectation in (4.4.2) is with respect to the alternative prior sampling distribution for $\boldsymbol{\xi}$. We note that in Chen et al. (2011) (cf. equation 5) the authors define the Bayesian type I error rate and power in terms of the null and alternative prior predictive distribution of the data, $\int p(D \mid \boldsymbol{\xi})\, \pi_0^{(s)}(\boldsymbol{\xi})\, d\boldsymbol{\xi}$ and

$\int p(D \mid \boldsymbol{\xi}) \pi_1^{(s)}(\boldsymbol{\xi}) d\boldsymbol{\xi}$, respectively. Our presentation here simply changes the order of integration to highlight the fact that the Bayesian type I error rate and Bayesian power are weighted averages of the quantities based on fixed values of $\boldsymbol{\xi}$. Our recipe for simulation-based estimation of the Bayesian type I error rate and Bayesian power follows closely the presentation in Chen et al. (2011).

4.4.2 Default Sampling Priors

It should be clear from Section 4.4.1 that the Bayesian type I error rate and power depend critically on the choice of the null and alternative sampling priors. It is natural that the sampling priors would reflect one's belief about the parameters in light of the information available from the historical trial. Presumably, the historical trial data will suggest that there is treatment efficacy but the evidence will not be overwhelming (hence the desire to conduct another trial). We propose a set of default sampling priors that are well suited for this scenario. After collecting the historical data, one's belief about the parameters is determined by $\pi(\boldsymbol{\xi} \mid \boldsymbol{D}_0)$. If one believes the historical and future trial patients are exchangeable (within levels of the covariates), then reasonable choices for the null and alternative sampling priors are $\pi_0^{(s)}(\boldsymbol{\xi}) = \pi(\boldsymbol{\xi} \mid \boldsymbol{D}_0, \gamma \geq 0)$ (the historical posterior given that H_0 is true) and $\pi_1^{(s)}(\boldsymbol{\xi}) = \pi(\boldsymbol{\xi} \mid \boldsymbol{D}_0, \gamma < 0)$ (the historical posterior given that H_1 is true).

Figure 4.1 illustrates the marginal posterior distribution for γ based on our analysis of the historical trial data used in the example application in Section 4.5 along with the corresponding default null and alternative marginal sampling priors for γ. Though we have only plotted the marginal sampling priors for γ, conditioning on the null or alternative hypothesis obviously induces changes in the entire joint distribution for $\boldsymbol{\xi}$. The distribution for the promotion time model parameters and other regression parameters can be quite different for the null and alternative sampling priors. The key point is that by defining the sampling priors as we have, we preserve the stochastic relationships between the treatment and nuisance parameters that are implied by the historical data under the assumption that a particular hypothesis is true. To make this idea more explicit, we contrast it with a commonly used approach that is easier to implement. Consider a point mass alternative sampling prior with all parameters set to their posterior means. For the null

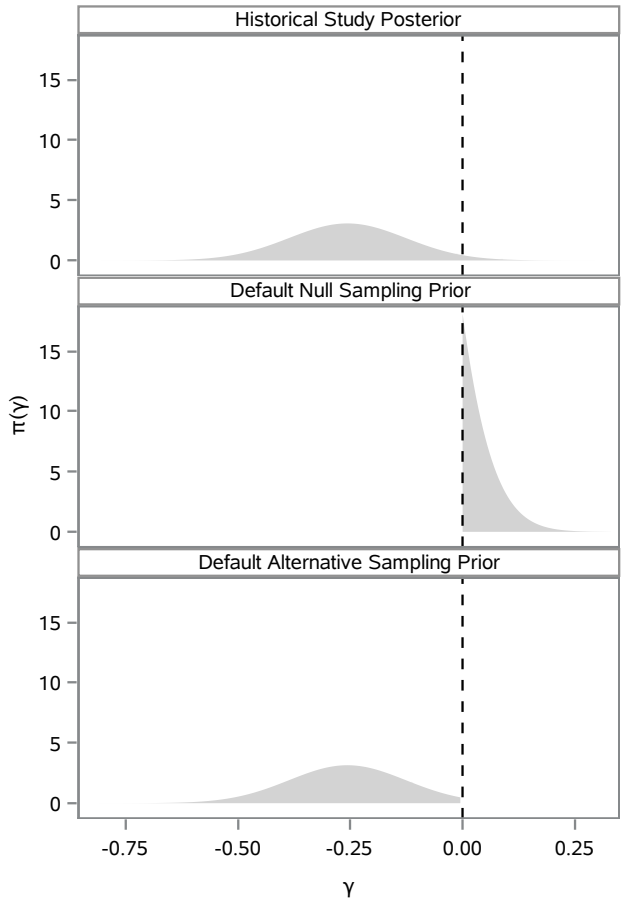

Figure 4.1: $\pi(\gamma \mid \boldsymbol{D}_0)$ and corresponding default marginal sampling priors.

sampling prior, the value of γ is simply set to zero with other parameters remaining unchanged. This approach is only sensible when γ is independent of the remaining parameters in $\boldsymbol{\xi}$ given $\boldsymbol{D}_0$. Of course, this is never the case.

A fundamental challenge in using the proposed sampling priors is that they cannot be directly sampled from since they do not have closed-forms. However, this issue is easily overcome through the following approximation. First, we fit the model to the historical trial data using MCMC methods to obtain B samples $\{\boldsymbol{\xi}_b : b = 1, ..., B\}$ from $\pi(\boldsymbol{\xi} \mid \boldsymbol{D}_0)$. We can then approximate sampling from $\pi_0^{(s)}(\boldsymbol{\xi})$ and $\pi_1^{(s)}(\boldsymbol{\xi})$ by sampling with replacement from the sets $\{\boldsymbol{\xi}_b : \gamma_b \geq 0\}$ and $\{\boldsymbol{\xi}_b : \gamma_b < 0\}$, respectively. If the number of sample points in each restricted set is large, the discrete approximations of the sampling priors will be accurate. Obviously the choice of B will depend on how much mass the historical trial posterior puts in the null region. Since this process only needs to be completed once, the computational burden of choosing B to be extremely large (e.g. $5,000,000$) is quite feasible.

4.4.3 Alternatives to the Default Sampling Priors

The default sampling priors previously described are sensible and automatic. However, there will be instances where it is desirable to modify the default priors. For example, researchers may deem it impossible for the new treatment to *decrease* the cure fraction by more than a certain amount relative to the planned control. In addition, researchers may want to compute power over a restricted alternative space that rules out implausibly large or clinically insignificant effect sizes. In this section, we introduce intuitive modifications to the default sampling priors that still preserve the stochastic relationships between the treatment and nuisance parameters that are implied by the historical data under the assumption that a particular hypothesis is true. We note that the procedure discussed in Section 4.4.2 for sampling from the default sampling priors applies to the modified versions discussed in this section.

The general null sampling prior is defined as $\pi_0^{(s)}(\boldsymbol{\xi}) = \pi(\boldsymbol{\xi} \mid \boldsymbol{D}_0, 0 \leq \gamma \leq \gamma_{0,u})$. As $\gamma_{0,u} \to 0$, less and less information can be borrowed from the historical data if the Bayesian type I error rate is to be controlled. For $\gamma_{0,u} = 0$, Bayesian type I error control is similar to frequentist type I

error control. The difference is that Bayesian type I error control implicitly assumes the nuisance parameters in the sampling model for the new trial are consistent with the posterior distribution for the historical trial (based on the reduced model without a treatment effect). In order to have Frequentist type I error control, the design methodology would have to be robust to differences between the nuisance parameters in the generative model for the new trial and the historical trial posterior. In the example application in Section 4.5, we consider designs based on the default null sampling prior and the most restrictive general null sampling prior defined by taking $\gamma_{0,u} = 0$. We will see that no information can be borrowed from the historical trial in the latter case. By extension, this implies that information borrowing is impossible if the design is required to have frequentist type I error control.

The general alternative sampling prior is defined as $\pi_1^{(s)}(\boldsymbol{\xi}) = \pi(\boldsymbol{\xi} \mid \boldsymbol{D}_0, \gamma_{1,l} \leq \gamma < \gamma_{1,u})$. Researchers will often be content with a trial design that does not lead to detection of small effect sizes with high probability. Thus, we expect that $\gamma_{1,u}$ will be chosen to be less than zero in most applications. Another option to be considered in lieu of the default alternative sampling prior is the point mass alternative sampling prior

$$\pi_1^{(s)}(\boldsymbol{\xi}) = 1\{\boldsymbol{\xi} = \mathrm{E}[\boldsymbol{\xi} \mid \boldsymbol{D}_0, \gamma < 0]\} \tag{4.4.3}$$

which sets the parameter values equal to their posterior means given the historical data and conditional on the alternative hypothesis being true. Figure 4.2 illustrates the marginal posterior distribution for γ based on the default alternative sampling prior and the restricted alternative sampling prior obtained by setting $\gamma_{1,l}$ and $\gamma_{1,u}$ equal to the limits of a 50% HPD interval for the treatment effect parameter based on the analysis of the historical trial used in the example application in Section 4.5. In that section, we contrast the power of designs based on the default alternative sampling prior, the restricted alternative sampling prior defined by choosing $\gamma_{1,l}$ and $\gamma_{1,u}$ as just described, and the point-mass sampling prior in (4.4.3).

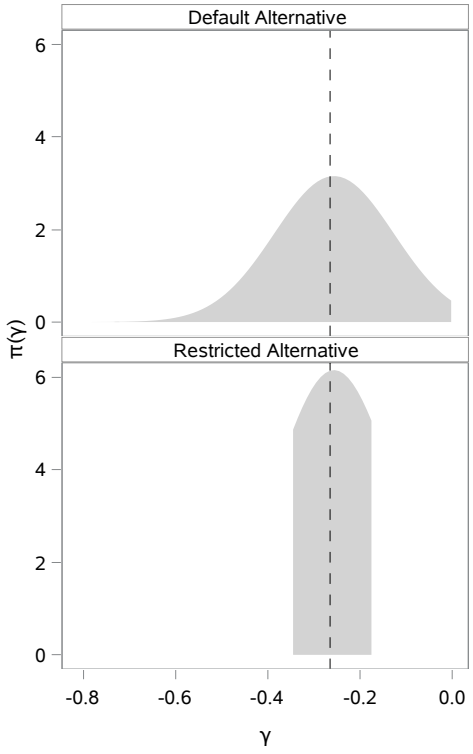

Figure 4.2: Alternative marginal sampling priors for γ.

4.4.4 Estimation of the Bayesian Type I Error Rate and Power

In this section, we describe the simulation process that is used to estimate the Bayesian type I error rate and power. Let B be the number of simulation studies to be performed. To estimate the Bayesian type I error rate, we proceed as follows:

1. Sample $\boldsymbol{\xi}^{(b)}$ from the null sampling prior $\pi_0^{(s)}(\boldsymbol{\xi})$.

2. Given $\boldsymbol{\xi}^{(b)}$, simulate the new trial data $\boldsymbol{D}^{(b)}$. This can be done using the following steps (for each subject):

   i. Simulate $\boldsymbol{x}_i$, z_i, and s_i based on the chosen randomization fraction, distribution for the covariates, and distribution for the stratification variable.

   ii. Calculate $\theta_i = \exp\left(\gamma z_i + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)$ and simulate $N_i \sim \text{Poisson}(\theta_i)$.

   iii. Simulate $Z_{ij} \sim F(z \mid \boldsymbol{\psi}_{s_i})$ independently for $j = 1, ..., N_i$ and calculate $z_i = \min(Z_{ij} : j = 0, ..., N_i)$ with $Z_{i0} = T$.

iv. Simulate the time-to-censorship, denoted as c_i, according to the chosen distribution. If only administrative censoring is entertained, then set $c_i = T$.

iv. If $z_i < c_i$ then set $y_i = z_i$ and $v_i = 1$, otherwise set $y_i = c_i$ and $v_i = 0$.

3. Update the fitting prior $\pi^{(f)}(\boldsymbol{\xi})$ based on the likelihood for the simulated data $\mathcal{L}\left(\boldsymbol{\xi} \mid \boldsymbol{D}^{(b)}\right)$ to obtain the posterior distribution $\pi\left(\boldsymbol{\xi} \mid \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)$ and calculate the posterior probability of the alternative hypothesis $P\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)$.

4. Compute the null hypothesis rejection indicator for simulated trial b:

$$r^{(b)} = 1\left\{P\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\}$$

5. Approximate the Bayesian type I error rate with the empirical null hypothesis rejection rate:

$$\alpha^{(s)} \approx \frac{1}{B}\sum_{b=1}^{B} r^{(b)}$$

Steps 1-4 are first repeated for $b = 1, ..., B$ to obtain the outcome for each simulated trial and then step 5 combines the results to estimate the Bayesian type I error rate. The process for estimating Bayesian power is identical. One simply needs to use the alternative sampling prior in place of the null sampling prior in the algorithm above.

4.5 Bayesian Design of a Superiority Trial in High-Risk Melanoma

The E1690 trial was conducted to assess the utility of Interferon Alfa-2b (INF) as an adjuvant therapy following surgery for deep primary or regionally metastatic melanoma. A detailed report on the trial was given in Kirkwood et al. (2000). Briefly, E1690 was a prospective, randomized, three-arm clinical trial designed to evaluate the efficacy of high-dose IFN for one year and low-dose IFN for two years relative to observation (OBS) in high-risk melanoma patients using relapse-free survival (RFS) and overall survival endpoints. We restrict our attention to the high-dose IFN regimen and consider the design of a follow-up RFS trial to confirm the efficacy of IFN.

Patients enrolled in the E1690 trial had histologically proven American Joint Committee on

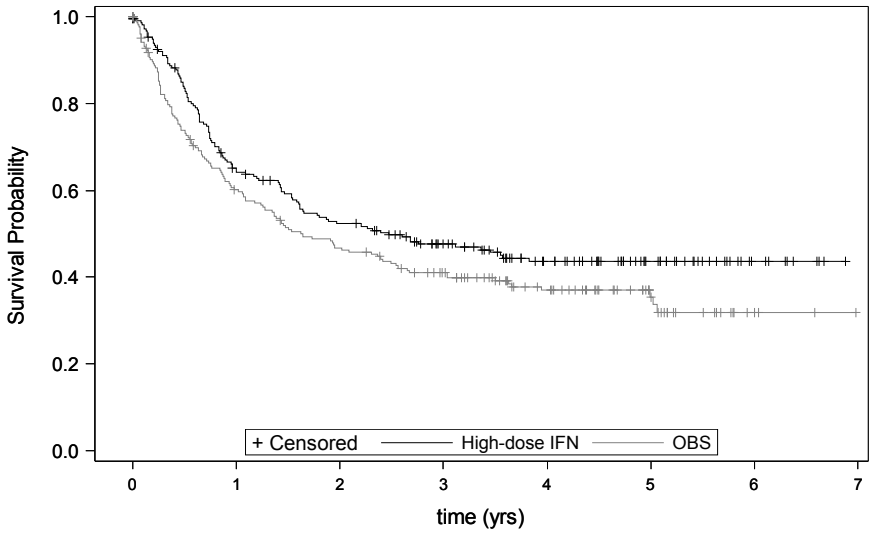

Figure 4.3: Kaplan-Meier curves for the high-dose INF and OBS groups.

Cancer (AJCC) stage IIB or stage III primary or recurrent regional nodal involvement from cutaneous melanoma without evidence of systemic metastatic disease (disease stages 1: T4cN0, 2: T1-4pN1cN0, 3: T1-4cN1, and 4: T1-4N1 recurrent). The randomization and primary analysis were stratified by disease stage and the number of positive nodes at lymphadenectomy. The primary analysis was based on a stratified log-rank test and the two-sided p-value was 0.054. There were 215 subjects and 114 relapses observed in the high-dose IFN group and 211 subjects and 126 relapses observed in the OBS group. Among the set of subjects who did not experience relapse, the median observation time was over four years. Figure 4.3 presents the Kaplan-Meier estimator for the survival curves for the high-dose IFN and OBS groups. Note the clear plateau appearing at approximately 4 years. This suggests a cure rate model is reasonable for these data. Overall the data from E1690 suggests a treatment benefit for RFS but the evidence is not overwhelming by traditional statistical criteria.

In the E1690 trial, disease stage and the number of positive nodes at lymphadenectomy were highly prognostic for RFS and so we consider these characteristics for inclusion in the design model to help ensure exchangeability of subjects across the two studies. To formally choose the design model, we compared a variety of promotion time cure rate models that adjusted for these covariates

Table 4.1: DIC for six best candidate design models

Stratification Variables	Cured Fraction Model Covariates	Weibull DIC	Exponential DIC
Stages 1-2 and 3-4	Treatment, 2-3 nodes, ≥ 4 nodes	1011.317	1011.557
Stages 1,2,3 and 4	Treatment, 2-3 nodes, ≥ 4 nodes	1014.234	1015.368
Stages 1-2 and 3-4	Treatment, ≥ 2 nodes	1017.845	1017.530

Table 4.2: Posterior summaries for the historical trial and default sampling priors

	— Posterior —		— Default Alternative —		— Default Null —	
Parm	Mean (SD)	HPD	Mean (SD)	HPD	Mean (SD)	HPD
γ	-0.26 (0.130)	(-0.51,0.00)	-0.26 (0.121)	(-0.49,-0.03)	0.05 (0.044)	(0.00, 0.14)
β_1	-0.10 (0.122)	(-0.34,0.14)	-0.09 (0.120)	(-0.32, 0.15)	-0.25 (0.107)	(-0.46,-0.04)
β_2	0.23 (0.172)	(-0.11,0.57)	0.23 (0.172)	(-0.11, 0.56)	0.21 (0.172)	(-0.12, 0.55)
β_3	0.77 (0.158)	(0.46,1.08)	0.77 (0.157)	(0.46, 1.08)	0.77 (0.156)	(0.46, 1.08)
λ_1	0.48 (0.088)	(0.31,0.65)	0.48 (0.088)	(0.31, 0.65)	0.49 (0.086)	(0.32, 0.66)
α_1	1.20 (0.134)	(0.94,1.47)	1.20 (0.135)	(0.94, 1.47)	1.21 (0.133)	(0.95, 1.47)
λ_2	0.67 (0.077)	(0.51,0.82)	0.67 (0.077)	(0.51, 0.82)	0.67 (0.078)	(0.51, 0.82)
α_2	1.06 (0.072)	(0.92,1.21)	1.06 (0.073)	(0.92, 1.21)	1.06 (0.073)	(0.92, 1.21)

in the model for the cured fraction and/or stratified by them in the model for the promotion times. Table 4.1 lists the six best fitting models according to the deviance information criterion (DIC) (Spiegelhalter et al., 2002). The DIC values were computed based on $50,000$ MCMC samples from an analysis using a uniform improper prior on the regression parameters and a Gamma $(0.001, 0.001)$ prior on each of the parameters in the promotion time model. It is shown in Ibrahim et al. (2001b) (cf. Theorem 5.2.1) that the resulting posterior is proper under mild conditions which are met by the E1690 data. In addition to the six models shown in Table 4.1, a variety of other models were considered including models that adjusted for disease stage in the model for the cured fraction and models that stratified by treatment and/or the number of positive nodes at lymphadenectomy in the promotion time model. We selected the model having the best fit according to DIC for design. Thus, the design model had separate Weibull promotion time distributions for disease stages 1-2 and for disease stages 3-4. The model for the cured fraction included an intercept, a treatment indicator, an indicator for having 2-3 positive nodes, and an indicator for having ≥ 4 positive nodes.

Table 4.2 presents the posterior mean, the posterior standard deviation (SD), and 95% highest posterior density (HPD) interval for all parameters based on an analysis of the historical E1690 data as well as for the default sampling priors. We note the highest posterior density (HPD) interval for the treatment effect puts a non-negligible amount of mass in both the null region and

the alternative region though the evidence clearly favors treatment efficacy (about 97.5% of the mass is in the alternative region). Note the dramatic change in the posterior mean for β_1 (the intercept) when we condition on the null hypothesis being true versus the alternative. If the null is true, the data imply the cured fraction in the untreated group is much higher than if the alternative is true (approximately 46% compared to 40%, for subjects with stage 1 or 2 cancer having ≤ 1 positive node at lymphadenectomy). This illustrates why it is inappropriate to fix the nuisance parameters at, say, their posterior means and simply vary the value of γ during design (unless one is purposefully considering sampling priors that are inconsistent with the historical data).

Our primary purpose in this section is to compare and contrast designs based on different choices of sampling priors. To illustrate how the choice of null sampling prior impacts the amount of information that can be borrowed from the historical trial, we consider two possibilities: the frequentist-like null sampling prior obtained by taking $\gamma_{0,u} = 0$ and the default null sampling prior obtained by taking ($\gamma_{0,u} = \infty$). To illustrate how the choice of alternative sampling prior impacts power, we consider three possibilities: the point mass alternative sampling prior with parameters set to the posterior means in Table 4.2 (for the default alternative sampling prior), the restricted alternative sampling prior with $\gamma_{1,l} = -0.343$ and $\gamma_{1,u} = -0.169$ which are the limits of the 50% HPD interval for γ based on the historical trial posterior, and the default alternative sampling prior. The point mass alternative sampling prior corresponds to a difference in the cured fraction of approximately 9.3% in subjects with stage 1 or 2 cancer having ≤ 1 positive node at lymphadenectomy. For design simulations, we assumed uniform enrollment over a period of 4 years and a trial duration of $T = 6.5$ years. The only censoring was administrative at trial completion. Moreover, the distribution for cancer stage and the number of positive nodes at lymphadenectomy was taken to match the observed distribution from E1690 and the randomization was 1:1.

For a chosen null sampling prior, the first step in the design process is to find the largest value of a_0 that leads to control of the Bayesian type I error rate for each sample size being considered. To do this, we performed $200,000$ simulations studies for $n = 550$ to $n = 800$ with a step size of 25 over a range of values for a_0 and estimated the Bayesian type I error rate for each combination of n and a_0. Next, we used degree-two polynomial regression to smooth estimates of the Bayesian type I error rate for each sample size and interpolated the correct choice of a_0. For these regression curve

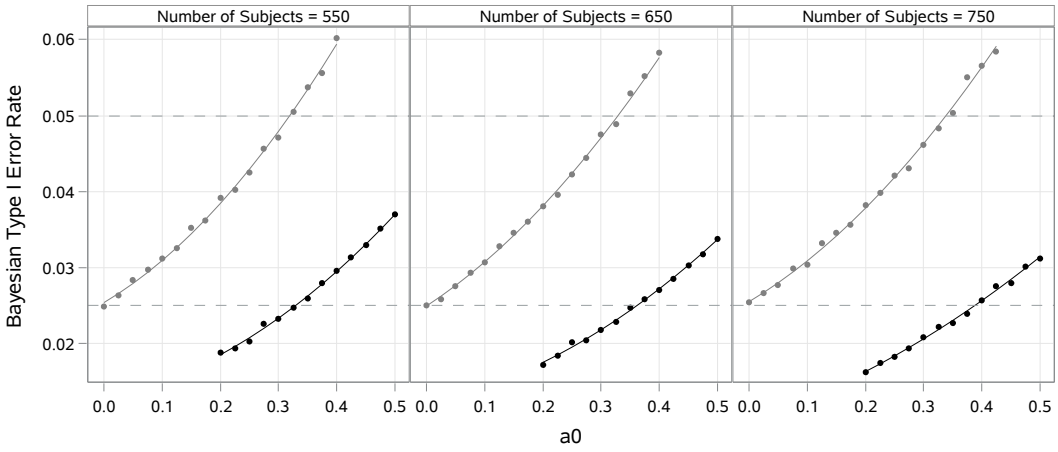

Figure 4.4: Regression curves and point estimates of the Bayesian type I error rate for $n = 550$, $n = 650$, and $n = 750$ based on the Frequentist-like null sampling prior (light gray) and the default null sampling prior (black). Each point estimate was based on $200,000$ simulation studies.

fits, the smallest R^2 value was 0.994 indicating a near perfect fit. Figure 4.4 presents estimates of the Bayesian type I error rate as a function of a_0 for both null sampling priors for sample sizes $n = 550$, $n = 650$, and $n = 750$. Lastly, we smoothed the chosen values of a_0 based on a simple linear regression model using sample size as the regressor. For the linear regression curve fits, the smallest R^2 value was 0.974 indicating near perfect fit. Figure 4.5 presents the estimated curves for a_0 as a function of sample size that lead to control of the Bayesian type I error rate at 2.5% for the default null sampling prior and at 5.0% for the frequentist-like null sampling prior.

It is clear from Figure 4.4 that virtually no information can be borrowed from the historical trial when the frequentist-like null sampling prior is used and when the Bayesian type I error rate is to be controlled at 2.5%. In fact, the estimates obtained for a_0 when using that null sampling prior were less than 0.005 for every value of n we considered. Based on these results, we set $a_0 = 0$ for the purpose of determining power in all design simulations using the frequentist-like null sampling prior (when the Bayesian type I error rate was to be controlled at 2.5%). In contrast, we see that when using the default null sampling prior, we are able to borrow a meaningful amount of information from the historical trial without surpassing the 2.5% Bayesian type I error threshold. As illustrated in Figure 4.5, we can borrow approximately 33% of the historical trial information when $n = 550$ and approximately 40% when $n = 800$. Note that due to the informativeness of the historical trial, it is impossible to borrow all of the historical trial information unless the future

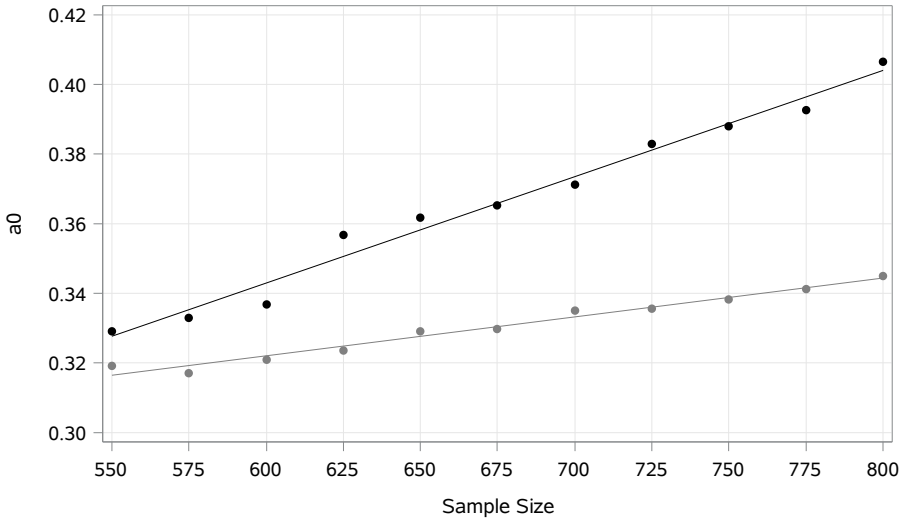

Figure 4.5: Regression curves and point estimates for a_0 as a function of sample size for designs that control of the Bayesian Type I error rate at 2.5% based on the default null sampling prior (black) and at level $2 \times 2.5 = 5.0\%$ based on the frequentist-like null sampling prior (light gray).

trial is unrealistically large. Thus, although the Bayesian version of type I error control (based on the default null sampling prior) is less restrictive than the frequentist-like type I error control, there is still significant restriction on the amount of information that can be borrowed. The fundamental difference is that when using the default null sampling prior, the restriction can be overcome by increasing the size of the future trial. It is also apparent from Figure 4.5 that controlling the Bayesian type I error rate at 2.5% based on the default null sampling prior does not meet the added restriction of controlling the Bayesian type I error rate at 5.0% based on the frequentist-like null sampling prior.

We estimated power using each of the three alternative sampling priors discussed in Section 4.4.3 for sample sizes from $n = 550$ to $n = 800$ using a step size of 25 and 200,000 simulations in each case. For each sample size and alternative sampling prior, we considered values of a_0 satisfying three possible criteria for type I error control: values that control the Bayesian type I error rate at 2.5% when using the default null sampling prior ($\text{DN}_{2.5}$), values that control the Bayesian type I error rate at 2.5% when using the Frequentist-like null sampling prior ($\text{FN}_{2.5}$), and values that control the Bayesian type I error rate at no more than 2.5% when using the default null sampling prior *and* at no more than 5.0% when using the Frequentist-like null sampling prior ($\text{FN}_{5.0}$). The latter approach may be appealing to regulatory bodies since it ensures a reasonable upper bound

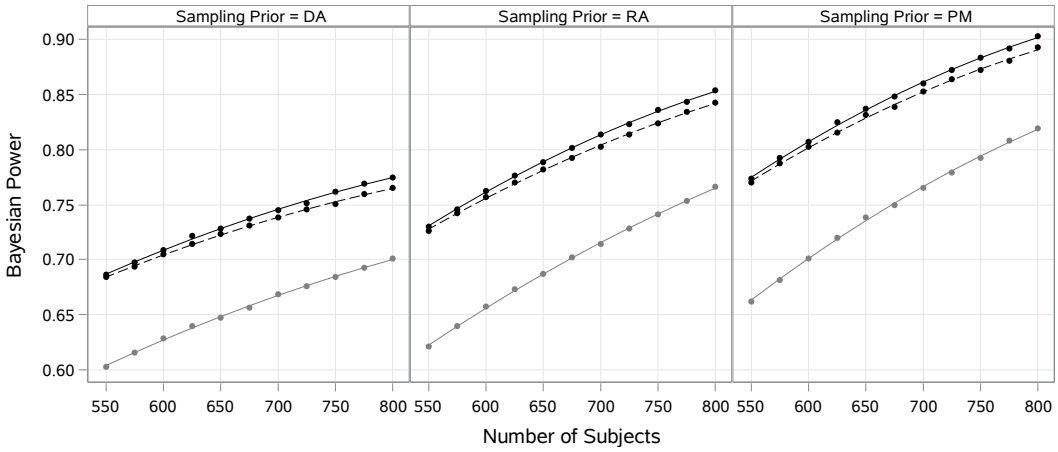

Figure 4.6: Quadratic regression curves and point estimates for Bayesian power as a function of sample size when the value of a_0 is chosen to control the Bayesian type I error rate at 2.5% when using the default null sampling prior (solid black line), to control the Bayesian type I error rate at 2.5% when using the Frequentist-like null sampling prior (solid gray line), and to control the Bayesian type I error rate at no more than 2.5% when using the default null sampling prior *and* at no more than 5.0% when using the Frequentist-like null sampling prior (dashed black line).

on the frequentist-like type I error rate for the design. Using degree-two polynomial regression model, we interpolated the power for intermediate samples sizes. For these regression curve fits, the smallest R^2 value was 0.998, indicating a near perfect fit. Figure 4.6 presents estimates of Bayesian power as a function of sample size for each criteria for choosing a_0. Table 4.3 includes the power estimates and associated values of a_0 for select sample sizes to faciliate more direct comparison. Using the point mass alternative sampling prior (PM), we obtain 80% power with sample sizes of approximately 590, 600, and 760 for Bayesian type I error criteria $\text{DN}_{2.5}$, $\text{FN}_{2.5}$, and $\text{FN}_{5.0}$, respectively. Using the more conservative restricted alternative sampling prior (RA), we obtain 80% power with sample sizes of approximately 670 and 690 for Bayesian type I error criteria $\text{DN}_{2.5}$ and $\text{FN}_{5.0}$, respectively. At the maximum sample size considered, the power only reaches approximately 76.5% when the frequentist-like type I error rate is controlled at 2.5%. Using the default alternative sampling prior (DA), which is the most conservative with respect to power, we do not reach 80% power even when the sample size is 800 for any of the three Bayesian type I error rate criteria. We expect very few trials will be powered using the default alternative sampling prior. Nonetheless, this sampling prior can provide a realistic estimate of power for the trial design based on what is deemed plausible given the historical data.

Table 4.3: Power estimates for select sample sizes

	$DN_{2.5}$				$FN_{5.0}$				$FN_{2.5}$		
n	a_0	DA	RA	PM	a_0	DA	RA	PM	DA	RA	PM
550	0.328	0.687	0.731	0.775	0.317	0.684	0.728	0.771	0.604	0.623	0.663
560	0.331	0.692	0.737	0.782	0.318	0.689	0.734	0.778	0.609	0.630	0.671
570	0.334	0.696	0.743	0.788	0.319	0.693	0.739	0.784	0.614	0.636	0.679
580	0.337	0.700	0.750	0.795	0.320	0.697	0.745	0.790	0.618	0.643	0.686
590	0.340	0.705	0.756	0.801	0.321	0.701	0.750	0.796	0.623	0.650	0.694
600	0.343	0.709	0.761	0.807	0.322	0.704	0.756	0.802	0.627	0.656	0.701
610	0.346	0.713	0.767	0.813	0.323	0.708	0.761	0.807	0.632	0.663	0.708
620	0.349	0.717	0.773	0.819	0.324	0.712	0.766	0.813	0.636	0.669	0.715
630	0.352	0.721	0.778	0.825	0.325	0.715	0.771	0.818	0.640	0.675	0.722
640	0.355	0.725	0.784	0.830	0.327	0.719	0.776	0.823	0.644	0.681	0.729
650	0.358	0.728	0.789	0.836	0.328	0.722	0.781	0.829	0.648	0.687	0.735
660	0.361	0.732	0.794	0.841	0.329	0.726	0.786	0.834	0.652	0.693	0.742
670	0.364	0.736	0.799	0.846	0.330	0.729	0.791	0.838	0.656	0.699	0.748
680	0.367	0.739	0.804	0.851	0.331	0.732	0.795	0.843	0.660	0.705	0.754
690	0.370	0.742	0.809	0.856	0.332	0.735	0.800	0.848	0.664	0.710	0.761
700	0.374	0.746	0.813	0.861	0.333	0.739	0.804	0.852	0.668	0.716	0.766
710	0.377	0.749	0.818	0.866	0.334	0.741	0.808	0.857	0.671	0.721	0.772
720	0.380	0.752	0.822	0.870	0.335	0.744	0.812	0.861	0.675	0.726	0.778
730	0.383	0.755	0.827	0.875	0.337	0.747	0.816	0.865	0.678	0.731	0.783
740	0.386	0.758	0.831	0.879	0.338	0.750	0.820	0.869	0.682	0.737	0.789
750	0.389	0.761	0.835	0.883	0.339	0.753	0.824	0.873	0.685	0.742	0.794
760	0.392	0.764	0.839	0.887	0.340	0.755	0.828	0.877	0.688	0.746	0.799
770	0.395	0.767	0.842	0.891	0.341	0.758	0.832	0.880	0.691	0.751	0.804
780	0.398	0.770	0.846	0.895	0.342	0.760	0.835	0.884	0.694	0.756	0.809
790	0.401	0.772	0.849	0.898	0.343	0.762	0.839	0.887	0.697	0.761	0.814
800	0.404	0.775	0.853	0.902	0.344	0.765	0.842	0.890	0.700	0.765	0.818

4.6 Discussion

In this paper, we have developed a Bayesian design methodology for time-to-event clinical trials with a cured fraction that makes use of data from a previously completed clinical trial. We have proposed Bayesian versions of type I error and power that are defined with respect to one's belief about the model parameters after observing the historical study data. Our results illustrate that if one requires the design to exhibit frequentist type I error control, it is not possible to borrow any information from the historical data. However, when one relaxes the design constraints to require Bayesian type I error control using our default null sampling prior, we are able to borrow substantial amounts of information from the historical study. Depending on the informativeness of the historical data, one may only be able to borrow a fraction of the available information. Thus, while Bayesian type I error control is less restrictive than frequentist type I error control, the former still places significant restrictions on how much information can be borrowed. We have also demonstrated that powering a study to have adequate Bayesian power under a non-degenerate alternative sampling prior will generally lead to a much larger study compared to a point mass alternative sampling prior using the most likely effect size based on the historical study posterior.

The conflict between Bayesian analysis with informative priors and frequentist type I error control is a topic of future research for the authors. In this paper, our case study focuses on a design application where information is borrowed on the treatment effect as well as nuisance parameters. In this case, the apparent conflict between frequentist type I error control and Bayesian analysis with an informative prior is undeniable. However, when one borrows information only on the nuisance parameters, the conflict is less obvious. This is because the historical study does not inform the treatment effect parameter and hence, does not overtly suggest the null hypothesis is false. However, since the sampling priors for the nuisance parameters are defined with respect to the historical study data, there is an implicit assumption that the nuisance parameter values in the generative model for the new study are consistent with the historical study posterior. In future work, we plan to consider designs that borrow information on the control group only, giving practical advice on defining and using default null and alternative sampling priors in this setting.

All design computations were performed using the weighted maximum likelihood approximation

to the posterior distribution as described in Section 4.3 and using the KillDevil computing cluster at the University of North Carolina at Chapel Hill. Weighted maximum likelihood analysis was performed using SAS/STAT® software, specifically the NLP procedure. The accuracy of the weighted maximum likelihood approximation is demonstrated via simulation in Appendix B.1. Data simulations and MCMC-based model fitting (to obtain discrete approximations of the sampling priors) was performed using custom C++ programming that was written by the authors.

CHAPTER 5: BAYESIAN DESIGN OF A CARDIOVASCULAR OUTCOMES TRIAL

5.1 Introduction

In December of 2008, the US Food and Drug Administration (FDA) issued a guidance for industry effectively establishing a two-stage framework for the assessment of cardiovascular risk in all new therapeutic agents intended for the treatment of Type 2 Diabetes Mellitus (T2DM) (Food and Drug Administration, 2008). The guidance specifies that the stage one data should rule out an 80% increase in cardiovascular risk for treated subjects compared to controls. If the stage one data do not also rule out a 30% increase in cardiovascular risk, and the overall risk-benefit analysis supports approval, a randomized controlled trial will generally be required to ultimately rule out a 30% increase in stage two. Thus, the stage one objective is to rule out a 80% increase in cardiovascular risk and the stage two objective is to further rule out a 30% increase in cardiovascular risk. In practice, the hazard ratio estimated from a Cox proportional hazards model is used as the basis for cardiovascular risk assessment and so stages one and two might equivalently be characterized as having to rule out hazard ratios of 1.8 and 1.3, respectively. The values 1.8 and 1.3 are commonly referred to as the stage one and stage two risk margins.

Various design strategies have been proposed and utilized for the evaluation of cardiovascular risk to satisfy the two-stage framework described above. These approaches and their relative merits are discussed in Geiger et al. (2015) and Marchenko et al. (2015). Strategies for ruling out the stage one risk margin include performing a meta-analysis of phase II and phase III trials with or without interim data from an ongoing cardiovascular outcome trial (CVOT) and conducting a CVOT designed and powered to ultimately rule out the stage two risk margin. At this time, it is expected that the stage two risk margin will be ruled out using data from a CVOT. The required number of subjects that must be recruited to enable timely completion of a CVOT designed to

rule out the stage two risk margin is substantial. This is because, even in the high risk T2DM population, cardiovascular events are relatively rare. In several published CVOTs the observed annualized event rate for major adverse cardiac events (MACE), the primary endpoint used to evaluate cardiovascular risk in these trials, was as low as 2%-3% (The ADVANCE Collaborative Group, 2008; The ACCORD Study Group, 2011; Scirica et al., 2013). Not surprisingly, there is considerable interest in reducing the size of these large trials. As more and more CVOTs complete, it is becoming increasingly apparent that the pool of completed CVOTs could be a valuable source of information to be used in the design and analysis of future CVOTs. Koch (2015) noted that Bayesian methods might be used to borrow information for the placebo control group leading to a new CVOT that might complete more quickly.

In this chapter, we develop two Bayesian adaptive design strategies based on assumed availability of subject-level placebo control data from a previously completed CVOT (i.e. a historical CVOT). Our focus is ruling out the stage two risk margin of 1.3 using information from the CVOT being planned as well as control information from a selected historical CVOT which is to be incorporated through an informative Bayesian prior. Ibrahim et al. (2012b) and Chen et al. (2014a) have previously developed Bayesian meta-design methods for CVOTs using trial-level covariate data with a goal of controlling the type I error rate in the traditional frequentist sense. Their results show that little to no information was borrowable under that constraint. The constraint that a design must control the type I error rate in the traditional sense forces one to discard all prior information (Berry et al. (2010), cf Chapter 5).

We take an approach that differs from Ibrahim et al. (2012b) in two key respects. First, we will not attempt to control the type I error rate in the traditional sense for the reasons described above. Instead, we propose adaptive designs that ensure the maximum type I error rate has a reasonable upper bound (e.g. twice the targeted level) and the minimum statistical power has a reasonable lower bound (i.e. 10% less than the targeted amount) over a range of possible parameter values for the generative model in the planned CVOT. Second, rather than relying on trial-level covariates (which may be extracted from publications), we propose methodology that requires subject-level data since covariate adjustment at that level is, in our opinion, critically important for justifiable information borrowing in a single arm of a multi-arm trial.

We propose and evaluate two novel strategies one might take for borrowing information from a historical CVOT. The first approach, which we call the all-or-nothing adaptive design, uses the basic power prior (Ibrahim and Chen, 2000). The basic power prior is a traditional prior in the sense that it allows a fixed amount of information to be borrowed from the historical CVOT. The amount of information to be borrowed would be negotiated with the pertinent regulatory body during the planning stages of the new CVOT. For example, one might target borrowing 25% of the events needed to have a desired power for the new CVOT. At a pre-planned interim analysis, which should occur after a reasonable number of events have been accrued in the new CVOT, one assesses the degree of conflict between the prior information and the new CVOT data using a likelihood ratio statistic where large values indicate a lack of agreement between the two data sources and hence a potential lack of exchangeability of the enrolled subjects (beyond what is addressed by covariate adjustment). If the likelihood ratio statistic is sufficiently large, the study continues to the final analysis where the data are analyzed with an objective Bayesian prior (i.e. no borrowing). Otherwise, the study is stopped at the interim analysis and the analysis is performed using the basic power prior constructed to yield the desired amount of borrowing. One determines what qualifies as a "sufficiently large" likelihood ratio statistic through a simulation-based procedure that evaluates the type I error rate and statistical power of the design over a range of possible parameter values for the generative model in the new CVOT.

For the second approach, which we call the dynamic borrowing adaptive design, one allows the amount of information borrowed from the historical CVOT to be dynamically adjusted using a novel extension of the joint power prior (Ibrahim and Chen, 2000) which we call the *restricted maximal borrowing power prior*. In this framework, the fraction of information borrowed from the historical CVOT is treated as a random variable (denoted by a_0). The restricted maximal borrowing power prior is constructed so that one obtains a desirable posterior distribution for a_0 when the observed data in the new CVOT and historical CVOT are perfectly homogeneous. Thus, when using this formulation of the power prior, one essentially specifies an upper bound posterior distribution governing the amount of information that can be borrowed in a best case scenario. At a pre-planned interim analysis, which should occur after a reasonable number of events have been accrued in the new CVOT, the posterior mean of a_0 (or an alternative posterior functional) is

examined to determine whether or not sufficient information is being borrowed from the historical CVOT. If it is determined that too little information is being borrowed relative to the desired amount, the study continues to the final analysis. Otherwise, the study is stopped at the interim analysis. In either case, the formal analysis utilizes the restricted maximal borrowing power prior to borrow the amount of information that is supported by the level of agreement between the data from the new and historical CVOTs at the time of the analysis. One determines the minimal amount of information borrowing that must take place in order to stop the CVOT at the interim analysis using a simulation-based design procedure that evaluates the type I error rate and statistical power of the design over a range of possible parameter values for the generative model in the new CVOT.

The rest of this chapter is organized as follows: In Section 5.2 we discuss historical CVOT selection, practical considerations for specifying the sampling model, and practical design considerations for the new CVOT so as to justify borrowing information from a selected historical CVOT. In section 5.3 we develop the general proportional hazards model that we use as the sampling model for both adaptive designs. In Section 5.4.1 we formalize the all-or-nothing adaptive design framework. In Section 5.4.2 we formalize the dynamic borrowing adaptive design framework. In Section 5.5 we discuss choosing the number of events at which to perform the interim and final analyses, possible choices for the randomization ratio, and the simulation-based procedure for identifying a reasonable criteria for stopping the trial at the pre-planned interim analysis. In Section 5.6 we present and compare detailed example designs using both adaptive design methods using the SAVOR trial (Scirica et al., 2011, 2013) as the historical CVOT. In Section 5.7 we close the chapter with some discussion.

5.2 Practical Design Considerations

The set of CVOTs that have been initiated to date reflect a high degree of variability in virtually all aspects of the targeted patient populations. There are differences in basic demographic characteristics (e.g. age ranges), differences in the required level of glycemic control (i.e. ranges for hemoglobin A1c), and differences in definitions of qualifying cardiovascular disease history. To illustrate this, we compare select characteristics of the target patient populations for three completed

CVOTs: the SAVOR, EXAMINE, and TECOS trials. The SAVOR trial serves as the historical CVOT in our example design applications to follow.

This SAVOR trial (Scirica et al., 2011, 2013) was a multi-center, randomized, double-blind, placebo-controlled CVOT designed to evaluate the effect of saxagliptin on the incidence of major adverse cardiovascular events (MACE), which is the required primary endpoint for ruling out the stage two risk margin in all CVOTs. In common with all other CVOTs, the target population for the SAVOR trial was enriched to include subjects that were at comparatively high risk for cardiovascular events. Enrolled subjects were required to be at least 40 years of age and to have a hemoglobin A1c (HbA1c) value of at least 6.5% but also less than 12.0%. In addition to age and glycemic control criteria, subjects enrolled in the SAVOR trial were required to have a history of cardiovascular disease or to present with multiple risk factors that included renal failure. The set of qualifying events defining a history of cardiovascular disease were ischemic heart disease, peripheral arterial disease (PAD), and/or ischemic stroke. Acceptable risk factors for cardiovascular disease included dyslipidemia, hypertension, and being a smoker at enrollment. Subjects were excluded from the SAVOR trial if they had an acute cardiovascular event in the two month period before enrollment, if they were severely obese (BMI>50), had severe dyslipidemia, or severe hypertension.

The EXAMINE trial (White et al., 2011, 2013) was a multi-center, randomized, double-blind, placebo-controlled CVOT designed to evaluate the effect of alogliptin on MACE events. Subjects were required to be at least 18 years of age. Those subjects not being treated with insulin at enrollment were required to have a HbA1c value of at least 6.5% but also less than 11.0%. Subject being treated with insulin at enrollment were required to have a HbA1c value of at least 7.0% but also less than 11.0%. Unlike the SAVOR trial, the EXAMINE trial enrolled only those subjects that presented with acute coronary syndrome (ACS) (acute myocardial infarction or unstable angina requiring hospitalization) between 15 and 90 days prior to enrollment. Subjects were excluded from the EXAMINE trial if they had one of several hemodynamically unstable cardiovascular disorders (NYHA class 4 heart failure, refractory angina, uncontrolled arrhythmia, critical valvular heart disease, or severe hypertension). Due to the requirement that enrolled subjects have ACS, the target population in the EXAMINE trial had fundamentally higher cardiovascular risk than the population targeted by SAVOR (at least in the period immediately following enrollment). Since an

inclusion criteria in the EXAMINE trial was essentially an exclusion criteria for the SAVOR trial, it should be clear that no amount of covariate adjustment could justify designing a future CVOT in the image of EXAMINE while targeting a SAVOR-like trial for information borrowing.

The TECOS trial (Green et al., 2013, 2015) was a multi-center, randomized, double-blind, placebo-controlled CVOT designed to evaluate the effect of sitagliptin on MACE events. The TECOS trial did not specifically target subjects who had acute coronary events (although it appears these subjects were not specifically excluded either). Like the SAVOR trial, the TECOS trial enrolled subjects with chronic conditions that are associated with increased cardiovascular risk. TECOS subjects were required to have a history of cardiovascular disease defined as having coronary artery disease, ischemic cerebrovascular disease (e.g. ischemic stroke), or peripheral arterial disease. These criteria would suggest the target population for TECOS was at least qualitatively similar to SAVOR (with respect to cardiovascular disease). However, unlike in the SAVOR trial, TECOS subjects were required to be at least 50 years old and to have a HbA1c value of at least 6.5% and but also less than 8.0% resulting in a target population that was somewhat older but had better glycemic control compared to the SAVOR and EXAMINE trials. Moreover, the TECOS trial excluded subjects in renal failure.

At this point we would like to make two important observations. First, it is quite obvious that no two CVOTs are the same in terms of their target populations (at least not yet). The fact is most CVOTs appear to be systematically different in this respect. Thus, the assumption of exchangeability of trials that is often made to justify the use of hierarchical models for information borrowing across multiple historical trials seems untenable. For this reason, we suggest borrowing information from a single historical CVOT that was selected and used, as much as possible, as a blue print for the design of the new CVOT. By this we mean that, on paper the historical CVOT and the new CVOT should be as similar as possible. Basic inclusion and exclusion criteria related to age, level of glycemic control, cardiovascular disease history, and other known prognostic factors for MACE events should be nearly identical.

The second point is that no matter what one does from a design perspective, it is not reasonable to assume that all subjects in the historical and new CVOTs are exchangeable so as to justify

performing the primary analysis using a very simple survival model as is commonly done in a single randomized controlled trial. It will be necessary to adjust for a reasonable set of prognostic factors to help better ensure exchangeability of subjects across trials. In our experience there are a multitude of characteristics (i.e. potential covariates) that are associated with increased risk for MACE events but only a few that are captured in clinical databases with the level of clarity and consistency that makes them useful across CVOTs. For example, history of percutaneous coronary intervention (PCI) is commonly collected in medical history but the information is often incomplete. Knowing whether the procedure was elective or performed to resolve an acute life-threatening event would be valuable as would knowing the date that the procedure was performed. Presumably, procedures from the distant past may not be relevant. The true relationship between a subject's history of PCI and their underlying cardiovascular risk is likely too complicated to be useful in a statistical analysis. In contrast, characteristics like age, duration of diabetes, baseline HbA1c, and estimated glomerular filtration rate (eGFR) provide clear and concrete information about the overall health of the subject and about cardiovascular risk (even if these characteristics represent surrogates for true underlying risk factors that are not clearly captured in the available data). We suggest one focus on adjusting for the latter type of characteristics in the statistical model and hope that the similarity of the study designs does an adequate job of addressing the former. This approach is admittedly imperfect and it speaks to the need for one to evaluate the assumption of exchangeability of subjects (conditional on covariates) to determine if information borrowing from the historical CVOT seems reasonable once some data is collected in the new CVOT.

5.3 The Piecewise Constant Hazard Proportional Hazards Model

In this section we present the sampling model used for analysis. We consider a flexible proportional hazards model where the baseline hazard is allowed to vary across S levels of a stratification variable. The cumulative hazard for subject i in the new CVOT is given by

$$\Lambda_i(t) = \Lambda_{[s_i]}(t) \exp\left(\gamma z_i + \boldsymbol{\beta}^T \boldsymbol{x}_i\right) \tag{5.3.1}$$

where s_i is the stratum to which subject i belongs, $\Lambda_{[s]}(t)$ is the baseline cumulative hazard for stratum s, z_i is a binary treatment indicator, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ is a $p \times 1$ vector of baseline covariates, γ is the log hazard ratio for treatment versus control, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is a $p \times 1$ vector of regression coefficients corresponding to the covariates.

Let $\lambda_{[s]}(t)$ denote the piecewise constant baseline hazard for stratum s. We partition the time axis into K_s intervals according to the change points $0 = t_{s,0} < t_{s,1} < .... < t_{s,K_s} = \infty$ and let $\lambda_{sk} > 0$ denote the constant hazard value over interval $I_{s,k} = (t_{s,k-1}, t_{s,k}]$. We denote the set of all baseline hazard parameters by $\boldsymbol{\lambda}$ and all model parameters by $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}, \boldsymbol{\lambda})$. We refer to this model in its entirety as the piecewise constant baseline hazard proportional hazards model (PWC-PH) model. The PWC-PH model is commonly used in Bayesian survival analysis when the proportional hazards assumption is tenable (Ibrahim et al., 2001b).

5.4 Adaptive Design Strategies

We assume the new CVOT will be an event-driven trial with $J = 2$ analyses and that analysis j will be performed when a pre-planned number of events ν_j have been accrued. We refer to the first analysis as the interim analysis and the second analysis as the final analysis. Throughout our discussion, we represent the observed data for the new CVOT at the time of analysis j by $\boldsymbol{D}_j$ and we represent the historical CVOT data by $\boldsymbol{D}_0$. Our goal is to develop a design that exhibits reasonable type I error control and that

(1) stops early at the interim analysis, resulting in an adequately powered analysis by virtue of borrowing control events from the historical CVOT, or

(2) stops at the final analysis, resulting in a adequately powered analysis by accruing a sufficient number of events in the new CVOT regardless of whether or not any events are borrowed from the historical CVOT.

In Section 5.4.1 we provide an overview of the all-or-nothing adaptive design approach and in Section 5.4.2 we provide an overview of the dynamic borrowing adaptive design approach. As noted previously, we consider information borrowing through historical CVOT control subjects

only. The data from the treated subjects in the historical CVOT would be discarded. Information from the historical CVOT is borrowed through the nuisance parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ which are common to the sampling models for the two CVOTs.

5.4.1 All-or-Nothing Adaptive Design

The all-or-nothing adaptive design is based on the basic power prior (Ibrahim and Chen, 2000). In this setting, the basic power prior can be written as follows:

$$\pi_0\left(\boldsymbol{\theta}\middle|\boldsymbol{D}_0, a_0\right) \propto \left[\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{a_0} \times \pi_0(\gamma, \boldsymbol{\beta}, \boldsymbol{\lambda}) \tag{5.4.1}$$

where $0 \leq a_0 \leq 1$ is a fixed scalar parameter, $\boldsymbol{D}_0$ is the historical data, $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)$ is the likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ given the historical data, and $\pi_0(\gamma, \boldsymbol{\beta}, \boldsymbol{\lambda})$ is an initial objective prior. We consider the initial prior

$$\pi_0(\gamma, \boldsymbol{\beta}, \boldsymbol{\lambda}) \propto \pi_0\left(\gamma, \boldsymbol{\beta}\right) \times \pi_0\left(\boldsymbol{\lambda}\right) \propto 1 \times \prod_{s=1}^{S} \prod_{k=1}^{K_s} \lambda_{sk}^{-1}.$$

When $a_0 = 0$ the historical CVOT data is essentially discarded and the power prior reduces to the initial prior. In contrast, when $a_0 = 1$ the power prior corresponds to the posterior distribution from an analysis of the historical CVOT using the initial prior. For intermediate values of a_0 the weight given to the historical CVOT is diminished to some degree leading to a prior that is more informative than the initial prior but less informative than using the historical CVOT posterior as the prior for the new CVOT.

Definition of the Interim Stoppage Criteria

For the all-or-nothing adaptive design, one fixes a_0 at some target value representing the desired fraction of the ν_0 control events that are to be borrowed from the historical CVOT. Generally, a_0 would be chosen to ensure that the *effective number of control events* at the interim analysis is sufficient to have adequate power under the assumption of no treatment effect and that the historical CVOT provides unbiased information about the generative model parameters in the new CVOT. The effective number of control events at the interim analysis is simply the number of

control events in the new CVOT plus the fraction borrowed from the historical CVOT, $a_0\nu_0$. At the interim analysis one evaluates the log-likelihood for the new CVOT data at the maximum a posteriori (MAP) estimates $\hat{\boldsymbol{\theta}}_0$ based on no borrowing and at $\hat{\boldsymbol{\theta}}_{a_0}$ based on the pre-planned amount of borrowing. If the log-likelihood ratio statistic

$$\log \mathcal{L}(\hat{\boldsymbol{\theta}}_0|\boldsymbol{D}_1) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_{a_0}|\boldsymbol{D}_1) \tag{5.4.2}$$

is large, this suggests that the underlying generative models for the historical and new CVOTs may be different (i.e. subjects are not exchangeable) and that information borrowing may not be justified. If it happens that

$$0 < \log \mathcal{L}(\hat{\boldsymbol{\theta}}_0|\boldsymbol{D}_1) - \log \mathcal{L}(\hat{\boldsymbol{\theta}}_{a_0}|\boldsymbol{D}_1) \leq \Delta\left(a_0\right)$$

for chosen $\Delta\left(a_0\right)$, then the study is stopped and the data are analyzed using the pre-planned amount of borrowing. If the inequality does not hold, the study continues to the final analysis and the data are analyzed using the initial prior (i.e. $a_0 = 0$).

The value of $\Delta\left(a_0\right)$ must be determined through simulation to ensure that the type I error rate and power for the design are reasonable under various generative models for the new CVOT. In particular, we propose choosing $\Delta\left(a_0\right)$ so that the statistical power does not drop by more than 10% when the baseline hazard in the generative model for the new CVOT is moderately less than the MAP estimate from historical CVOT (in every stratum). In addition, we require that the choice of $\Delta\left(a_0\right)$ results in a type I error rate that is less than double the targeted rate when the baseline hazard in the generative model for the new CVOT is moderately more than the MAP estimate from historical CVOT (in every stratum). We discuss how to formally identify an acceptable value for $\Delta\left(a_0\right)$ in Section 5.5.1.

We note that there are an infinite number of ways in which the historical study MAP estimates could differ from the generative model parameters in the new CVOT. Clearly, it is impossible to investigate all possibilities. We choose to focus on equal multiplicative perturbations in all baseline hazard parameters since this type of modification has the intuitive interpretation of systematically shifting the baseline cardiovascular risk in the target population of the new CVOT up or down

relative to the historical CVOT. In practice, if there is concern about a particular covariate effect, it would make sense to consider perturbations of the corresponding hazard ratio regression parameter as well.

5.4.2 Dynamic Borrowing Adaptive Design

The joint power prior (Ibrahim and Chen, 2000) generalizes the basic power prior by treating the a_0 parameter as a random variable and giving it a prior distribution. For control group only borrowing using model (5.3.1), the form of the joint power prior is as follows:

$$\pi_0\left(\boldsymbol{\theta}, a_0 \middle| \boldsymbol{D}_0\right) \propto [\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)]^{a_0} \times \pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda}) \times \pi_0(\gamma) \times \pi_0(a_0) \tag{5.4.3}$$

where $\pi_0(a_0)$ is an initial prior for a_0 and other quantities are the same as in (5.4.1). The joint power prior allows the level of agreement between the observed data in the historical and new trial to influence the amount of information that is borrowed.

Several authors have made arguments in favor of alternative generalizations of the basic power prior. In particular, Duan et al. (2006) advocate the normalized power prior which can be viewed a special case of the joint power prior that is obtained by dividing (5.4.3) by

$$\int \{\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)^{a_0} \times \pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})\}\, d\,(\boldsymbol{\beta}, \boldsymbol{\lambda})$$

and thus changing the initial prior on a_0. The primary argument in favor of the normalized power prior is that it obeys the likelihood principle whereas the joint power prior does not. That is, in the normalized power prior one can replace $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)$ with $c \cdot \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)$ for $c > 0$ and the resulting inference will not change. In the case of the normalized power prior, $\pi_0(a_0)$ represents the marginal prior distribution for a_0 whereas in (5.4.3) it does not. These properties of the normalized power prior would be compelling if the initial prior for $\pi_0(a_0)$ represented ones belief about the parameter a_0. However, in practice, the a_0 parameter is simply modeled as random to facilitate data-driven information borrowing. Accordingly, we view $\pi_0(a_0)$ as nothing more than a characteristic of the statistical model that one can manipulate to control the operational characteristics of the design.

From this point of view, it should be obvious that the choice to use the normalized power prior or joint power prior is arbitrary.

Another generalization of the basic power prior is the commensurate power prior (Hobbs et al., 2011, 2012). When the data are not normally distributed, the commensurate power prior cannot be applied without likelihood approximation (e.g. Laplace approximation) which will be potentially inaccurate for baseline hazard parameters. Moreover, the baseline hazard can be integrated out when using the joint power prior and this greatly simplifies MCMC sampling during design (since estimating the baseline hazard is irrelevant in design simulations). The same integration could not be done analytically when a commensurate prior is placed on the baseline hazard and, as a result, the computational burden of model fitting would be much greater. For these reasons we did not consider a version of the commensurate power prior.

Choosing the Initial Prior $\pi_0(a_0)$

As previously discussed, we view the initial prior $\pi_0(a_0)$ as a characteristic of the statistical model that can and should be manipulated to ensure the design has reasonable operating characteristics. The form of the initial prior that yields good operating characteristics will depend on a variety of things including the size of the historical study (e.g. number of events) and the model complexity. As a general rule, as the ratio of the number of parameters to total information (e.g. number of events) increases, so to will the amount of information borrowed from the historical CVOT even if the most likely parameter values determined by the two likelihoods are quite different. This is because, as the number of parameters increases, the likelihoods will be increasingly flat and, hence, relatively consistent with one another. Failing to acknowledge this property can result in poor type I error control.

In light of these observations it would seem ideal that we specify the initial prior for a_0 in a way that takes into account the complexity of the chosen sampling model and the total number of events in the historical CVOT and new CVOT at the time of the interim analysis. To that end, we propose the restricted maximal borrowing power prior which has the same form as (5.4.3) with

initial prior

$$\pi_0(a_0) \propto \frac{\pi^*(a_0)}{\int \{\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}_0)^{a_0+\delta}\pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda})\}\, d\,(\boldsymbol{\beta}, \boldsymbol{\lambda})} \tag{5.4.4}$$

where δ is the ratio of events in the new CVOT at the interim analysis to the number of events in the historical CVOT controls. This formulation of the joint power prior has the property that, in the idealized scenario where the new CVOT data and historical CVOT data are perfectly homogeneous, the *posterior* distribution $\pi(a_0 \mid \boldsymbol{D}_1, \boldsymbol{D}_0)$ is equal to $\pi^*(a_0)$. By perfectly homogeneous we mean that the likelihood for the new CVOT is identical to the likelihood for the historical CVOT apart from being scaled by the power δ which accounts for differing numbers of events in the two trials. Details on the calculation of the denominator in (5.4.4) can be found in Appendix C.1.

By specifying the prior in this way, one essentially places an upper bound on the amount of information that can be borrowed from the historical CVOT. Not only does this approach ensure that model complexity does not unduly influence the amount of information borrowing, it also provides a very intuitive frame of reference for how much information might be borrowed in the analysis. In practice we propose a beta distribution for $\pi^*(a_0)$. When specifying the mean parameter and dispersion parameter, it is important to note that obtaining the amount of borrowing implied by $\pi^*(a_0)$ is essentially impossible. Thus, the mean parameter should be larger than the fixed value of a_0 that is used in the all-or-nothing approach. Regardless of what distribution is chosen for $\pi^*(a_0)$, several choices for the parameters governing the mean and spread of the distribution will likely need to be explored in order to ensure that the amount of borrowing actually obtained (on average) is reasonable when the historical CVOT provides unbiased information about the parameters in the generative model for the new CVOT.

Definition of the Interim Stoppage Criteria

At the interim analysis, one computes the effective number of events being borrowed

$$\mathrm{E}\left[a_0 \mid \boldsymbol{D}_1, \boldsymbol{D}_0\right] \times \nu_0$$

and compares that value to a pre-identified minimum acceptable number of events δ_0. If it happens that

$$\mathrm{E}\left[a_0 \mid \boldsymbol{D}_1, \boldsymbol{D}_0\right] \times \nu_0 < \delta_0$$

then the study will continue to the final analysis. Otherwise, the study will be stopped at the interim analysis. In either case, the formal analysis will incorporate whatever amount of information is supported by the historical and new CVOT data at the time of the formal analysis.

5.5 The Simulation-Based Design Strategy

The null and alternative hypotheses for ruling out the stage two risk margin are as follows:

$$H_0 : e^{\gamma} \geq 1.3 \text{ versus } H_1 : e^{\gamma} < 1.3$$

where γ is the log-hazard ratio for treatment versus control. We accept H_1 if the posterior probability of the alternative hypothesis is at least as large as some critical value ψ. To decrease the search space for the number of study characteristics that are manipulated in design simulations, we recommend fixing $\psi = 1 - \alpha$ where α is the targeted type I error rate for the design. One needs to select the number of events ν_1 and ν_2 at which the interim and final analyses will take place. Traditionally, the final analysis for a CVOT is performed when approximately 612 events have accrued. This will result in approximately 90% power to rule out the stage two risk margin under the assumption of no treatment effect (i.e $\gamma = 0$) and 1:1 randomization. Since the worst case scenario for both the all-or-nothing adaptive design and the dynamic borrowing adaptive design is no borrowing, it makes sense to take $\nu_2 = 612$ for both approaches. The choice of ν_1 is more challenging. If the interim analysis is very early, a large fraction of the total events will need come from the historical CVOT and there may not be enough events in the new CVOT to fully evaluate whether information borrowing is reasonable. For example, one would want to ensure a reasonable amount of observation time in each time interval defined by the baseline hazard model. A very late interim analysis will require less borrowing from the historical CVOT but will also result in less efficiency gain over the traditional CVOT (i.e. little decrease in the average trial duration compared

to the traditional CVOT). For our discussion, we plan the interim analysis at $\nu_1 = 612 \times 0.75 = 459$ events which should error on the side of having accrued sufficient events in the new CVOT to allow legitimate evaluation of the comparability of the data from the two trials.

The number of events that must be borrowed from the historical CVOT will depend on the randomization strategy that is implemented. There are two reasonable choices for allocating subjects to the treatment arms in the new CVOT. First, one could implement unbalanced randomization (i.e. 2:1 allocation) which should result in an approximately equal effective number of events in the two arms at the interim analysis in the situations where the CVOT actually stops at the interim. In this case, the unbalanced randomization will result in the formal analysis having optimal power (under the assumption of no treatment effect). In contrast, balanced randomization (i.e. 1:1 allocation) would be preferred if the study proceeds to the final analysis (since there will be little or no borrowing in this case). Our simulation studies suggest that the performance of balanced randomization is superior to unbalanced randomization and so we recommend using that allocation strategy regardless of the adaptive design strategy taken. This means that slightly more than 153 events will need to be borrowed from the historical CVOT in order to have 90% power at the interim analysis.

Having decided on the number of events at which to perform the interim and final analyses as well as the treatment allocation ratio, what remains is to determine a reasonable criteria for when the trial should be stopped at the interim analysis. This requires a large scale simulation study using a range of possible parameter values for the new CVOT generative model. The parameter values should range from being in complete agreement with the historical CVOT MAP estimates to being in strong disagreement with them.

Let $\boldsymbol{\theta}_{0,1}$,...,$\boldsymbol{\theta}_{0,M}$ represent the collection of null parameter values considered and $\boldsymbol{\theta}_{1,1}$,...,$\boldsymbol{\theta}_{1,M}$ represent the corresponding alternative parameter values. We assume $\boldsymbol{\theta}_{0,m} = \boldsymbol{\theta}_{1,m}$ apart from the value of γ which will be zero for alternative cases and $\log(1.3)$ for null cases. For hypothesis $h = 0, 1$ and parameter value $m = 1, ..., M$, one must simulate the data at the time of the final analysis for B hypothetical new CVOTs, denoted by $\boldsymbol{D}^{(1)}_{hm,2}, ..., \boldsymbol{D}^{(B)}_{hm,2}$, and then back calculate the interim analysis data, denoted by $\boldsymbol{D}^{(1)}_{hm,1}, ..., \boldsymbol{D}^{(B)}_{hm,1}$. Next one computes the posterior probability

of the alternative hypothesis and determines whether or not to reject the null hypothesis for each simulated CVOT at each analysis. Let $r_{hm,j}^{(b)}$ be the indicator that one rejects the null hypothesis based on the observed data at analysis j for simulation study b using generative model parameter value $\boldsymbol{\theta}_{h,m}$. Using the set of rejection indicators $\left\{r_{hm,j}^{(b)} : h = 0, 1; m = 1, ..., M; b = 1, ..., B\right\}$ along with corresponding the interim analysis log-likelihood ratio statistic from (5.4.2) (for the all-or-nothing adaptive design) or posterior mean for a_0 (for the dynamic borrowing adaptive design), one can determine an acceptable criteria for when the study should be stopped at the interim analysis.

5.5.1 Stoppage Criteria for All-or-Nothing Adaptive Designs

Let $\ell_{hm}^{(b)}$ denote the interim analysis log-likelihood ratio statistic for simulation study b based on generative model parameter value $\boldsymbol{\theta}_{h,m}$. We identify a reasonable value of $\Delta(a_0)$ by examining an array of possible values beginning at zero and ending at a value just larger than the maximum likelihood log-likelihood ratio statistic observed over all simulation studies. If $\Delta(a_0) = 0$ then the CVOT will always continue to the final analysis resulting in a design that mirrors the traditional CVOT approach. If $\Delta(a_0)$ is larger than the maximum log-likelihood ratio statistic observed in the simulation studies, the CVOT will always stop at the interim analysis and the pre-planned amount of information will always be borrowed. To determine the subset of choices for $\Delta(a_0)$ that yield acceptable designs, we proceed in the following way. Let Δ represent a potential cutoff value for the log-likelihood ratio statistic. One computes the empirical null hypothesis rejection rate $\hat{r}_{h,m}(\Delta)$ defined as

$$\hat{r}_{h,m}(\Delta) = \frac{1}{B}\sum_{b=1}^{B}\left[r_{hm,1}^{(b)} \cdot 1\left(\ell_{hm}^{(b)} \leq \Delta\right) + r_{hm,2}^{(b)} \cdot 1\left(\ell_{hm}^{(b)} > \Delta\right)\right] \tag{5.5.1}$$

for each generative model parameter value $\boldsymbol{\theta}_{h,m}$. For $h = 0$ the empirical null hypothesis rejection is an estimate of the type I error rate and for $h = 1$ it is an estimate of power. If the maximum estimated type I error rate over all M null parameter values is less than an acceptable upper bound (e.g. 5.0%, twice the targeted type I error rate) and the minimum power is more than an acceptable lower bound (e.g. 80%, 10% less than the targeted power), then the value of Δ yields an acceptable design for the chosen value of a_0. Among the set of values that yield an acceptable design, one

would set $\Delta(a_0)$ equal to the value that results in the the maximum estimated probability of early stoppage.

5.5.2 Stoppage Criteria for Dynamic Borrowing Adaptive Designs

The approach for choosing when the effective number of events being borrowed from the historical study is too low to justify stopping the CVOT at the interim analysis is similar to the above procedure for the all-or-nothing design. We consider an array of possible cutoff values for the effective number of events being borrowed ranging from 0 the mean parameter for $\pi^*(a_0)$. For each potential cutoff δ, one computes the empirical null hypothesis rejection rate $\hat{r}_{h,m}(\delta)$ defined as

$$\hat{r}_{h,m}(\delta) = \frac{1}{B}\sum_{b=1}^{B}\left[r_{hm,1}^{(b)} \cdot 1\left(\mathrm{E}\left[a_0 \mid \boldsymbol{D}_1, \boldsymbol{D}_0\right] \times \nu_0 \geq \delta\right) + r_{hm,2}^{(b)} \cdot 1\left(\mathrm{E}\left[a_0 \mid \boldsymbol{D}_1, \boldsymbol{D}_0\right] \times \nu_0 < \delta\right)\right] \quad (5.5.2)$$

for each generative model parameter value $\boldsymbol{\theta}_{h,m}$. If the maximum estimated type I error rate over all M null parameter values is less than an acceptable upper bound (e.g. 5.0%, twice the targeted type I error rate) and the minimum power is more than an acceptable lower bound (e.g. 80%, 10% less than the targeted power), then the value of δ yields an acceptable design for the given choice of $\pi^*(a_0)$. Among the set of values that yield an acceptable design over all choices for $\pi^*(a_0)$, one would set δ_0 equal to the value that results in the the maximum estimated probability of early stoppage.

5.6 Designing a CVOT to Borrow from the SAVOR Trial

In this section we compare the all-or-nothing adaptive design and the dynamic borrowing adaptive design. For this exercise, we utilize the SAVOR trial as the historical CVOT. As noted in Section 5.2, the SAVOR trial was a multi-center, randomized, double-blind, placebo-controlled CVOT designed to evaluate the effect of saxagliptin on the incidence of major adverse cardiovascular events (MACE). The rationale for the SAVOR trial was discussed in Scirica et al. (2011) and the study results are discussed in Scirica et al. (2013).

We included the following characteristics in the sampling model: age, duration of diabetes, baseline HBA1c, baseline eGFR, gender, history of MI, and history of stroke. Each of these characteristics had strong association with the MACE endpoint in the SAVOR trial and it was felt that these characteristics could be consistently measured across trials. Binary data for many other medical history characteristics (e.g. history of PCI, history of PAD, history of hypertension) were available and could be evaluated for potential inclusion in the sampling model. Most medical history provides imprecise information about underlying cardiovascular risk. For example, the date of a previous MI and the underlying reason for a PCI procedure are clearly important, but both were essentially unavailable. Our decision to include a particular characteristic in the sampling model reflected a balance between the apparent statistical significance of the characteristic in the SAVOR trial and the clarity of the information provided by the characteristic as judged by subject matter experts. Building a "best" model for borrowing from the SAVOR trial is beyond the scope of this work. Our goal here is to simply illustrate the two approaches we have developed using a reasonable sample model.

Due to the computational complexity of the dynamic borrowing adaptive design, we were forced to dichotomize duration of diabetes (> 10 years), baseline HbA1c ($> 8\%$), and baseline eGFR (< 60 mL/min). For the PWC-PH model, the data can be reduced to one observation per unique value of the covariates. When the covariates are all discrete, this leads to a tremendous speed up in model fitting via MCMC. For the dynamic borrowing adaptive design, MCMC cannot be avoided. However, this is not the case for the all-or-nothing adaptive design. With virtually no approximation error, the all-or-nothing adaptive design can be performed by exploiting a connection between the PWC-PH model and a weighted Poisson regression model. The asymptotic p-value from a maximum likelihood analysis of the weighted Poisson regression model is essentially equal to the required posterior probability for Bayesian inference using the PWC-PH model with the basic power prior. Thus, one can get what is needed for the simulation-based design without fitting the model with MCMC. However, to facilitate comparison between the two design approaches, we used the dichotomized covariates for both.

The sampling model was stratified by age using three stratum (≤ 60 years, > 60 years and ≤ 70 years, and > 70 years). A constant hazard was assumed for the first two strata and a three

Table 5.1: Posterior summaries for SAVOR trial

Parameter	Characteristic	MAP	Mean	SD	HPD
β_1	Male	0.3666	0.3682	0.0930	(0.1887,0.5539)
β_2	History of Stroke	0.6844	0.8606	0.1038	(0.4731,0.8800)
β_3	History of MI	0.4864	0.4869	0.0829	(0.3233,0.6477)
β_4	Diabetic $>$ 10 years	0.2782	0.2796	0.0873	(0.1069,0.4469)
β_5	HBA1c $>$ 8.0%	0.4049	0.4043	0.0841	(0.2338,0.5627)
β_6	eGFR $<$ 60 mL/min	0.6057	0.6052	0.0861	(0.4353,0.7710)
$\lambda_{1,1}$	Age $\leq$ 60	0.0104	0.0104	0.0014	(0.0078,0.0132)
$\lambda_{2,1}$	60 $<$ Age $\leq$ 70	0.0118	0.0118	0.0014	(0.0091,0.0146)
$\lambda_{3,1}$	Age $>$ 70	0.0121	0.0121	0.0020	(0.0084,0.0159)
$\lambda_{3,2}$		0.0149	0.0149	0.0025	(0.0104,0.0200)
$\lambda_{3,3}$		0.0208	0.0208	0.0033	(0.0146,0.0274)

component model was assumed for the age $>$ 70 years stratum. The change points were fixed at 0.85 years and 1.50 years which lead to an even number of events in each time interval. The choice to stratify by age was made due to a potential violation of the proportional hazards assumption.

There were 8145 SAVOR control subjects with complete data for each of the characteristics that were included in the sampling model. Of those 8145 subjects, 605 experienced a MACE event. Table 5.1 presents posterior summaries from an analysis of the SAVOR data using the objective prior described in Section 5.4.1. Posterior summaries are based on 25,000 MCMC samples. For all parameter estimates, the MAP estimates were essentially equal to the posterior mean.

When simulating data for the new CVOT, we used the MAP estimates from Table 5.1 for the hazard ratio regression parameters. For the baseline hazard parameters, we considered generative model parameters that differed from the MAP estimates in Table 5.1 by 0%, 10%, 20%, and 30%. We considered increases and decreases to the baseline hazard parameters. The covariates and stratification variable were simulated to match that observed marginal distribution for each based on the SAVOR trial. Enrollment was simulated to be uniform over a two year period and no dropout was assumed. For each possible value of the generative model parameters, we simulated 100,000 CVOTs to evaluate the all-or-nothing adaptive design and 40,000 CVOTs to evaluate the dynamic borrowing adaptive design. The difference in the number of simulation studies was a result of the computational burden of MCMC model fitting for the dynamic borrowing adaptive design and the need to investigate multiple possible choices for $\pi^*(a_0)$. For the all-or-nothing

borrowing approach, we took $a_0 = 0.263$ resulting in a target of 159 events being borrowed from the SAVOR trial ($\nu_2 - \nu_1 = 153$). For the dynamic borrowing approach, we consider nine possible beta distributions for $\pi^*(a_0)$ corresponding to mean parameters 0.45, 0.50, and 0.55 and strength parameters 1.5, 2.0, and 2.5. For the all-or-nothing adaptive design and for each choice of $\pi^*(a_0)$ in the dynamic borrowing adaptive design, we determined a reasonable interim stoppage criteria using the procedures described in Sections 5.5.1 and 5.5.2, respectively.

Table 5.2 presents characteristics of the optimal all-or-nothing adaptive design and the optimal dynamic borrowing adaptive design. The optimal all-or-nothing adaptive design is based on stopping at the interim analysis if the log-likelihood ratio statistic in (5.4.2) is no more than log(1.8). The optimal dynamic borrowing adaptive design is based on stopping at the interim analysis if the effective number of events being borrowed is at least 121 and based on a beta distribution for $\pi^*(a_0)$ with mean parameter 0.55 and strength parameter 2. For this choice, the average posterior mean for a_0 at the interim analysis across all simulations was approximately 0.28 in the situation where the historical CVOT MAP estimates were unbiased with respect to the generative model parameters in the new CVOT. The mean decrease in CVOT duration and percent change in CVOT duration are computed with respect to a CVOT that always proceeds to the final analysis.

The two optimal designs are quite similar in terms of the probabilities for early stoppage and reduction in the duration of the average CVOT. The designs are also similar in terms of power and type I error control (by construction). One key property of both designs is that when the baseline hazard in the generative model for the new CVOT is at least 30% different from the historical CVOT MAP estimates, the new CVOT nearly always proceeds to the final analysis leading to adequate type I error control and high power. This highly desirable property is what allows us to avoid considering more extreme differences between the baseline hazard in the generative model for the new CVOT and the historical CVOT MAP estimates. If we did not observe this behavior with a 30% difference, we would need to consider larger differences.

It is apparent that the type I error rate is most inflated when the baseline hazard in the generative model for the new CVOT is modestly lower than the historical CVOT MAP estimates. Similarly, the power reduction is greatest when the baseline hazard in the generative model for

Table 5.2: Optimal adaptive designs

True Hazard Ratio	New CVOT BL. Hazard Scale Factor	Early Stoppage Probability	Estimated Power/Type I Error Rate	Mean Duration (years)	Mean Decrease in Duration	Percent Change in Duration
			Optimal All-or-Nothing Adaptive Design			
1.0	0.7	0.001	0.900	5.69	0.00	0.01
	0.8	0.070	0.898	5.06	0.07	1.42
	0.9	0.513	0.911	4.22	0.48	10.18
	1.0	0.796	0.900	3.68	0.67	15.34
	1.1	0.541	0.821	3.65	0.41	10.15
	1.2	0.153	0.842	3.72	0.11	2.80
	1.3	0.018	0.892	3.62	0.01	0.32
1.3	0.7	0.001	0.025	5.11	0.00	0.02
	0.8	0.077	0.051	4.56	0.07	1.52
	0.9	0.514	0.050	3.83	0.42	9.82
	1.0	0.790	0.020	3.37	0.58	14.66
	1.1	0.537	0.024	3.34	0.36	9.72
	1.2	0.164	0.025	3.39	0.10	2.90
	1.3	0.022	0.024	3.30	0.01	0.38
			Optimal Dynamic Borrowing Adaptive Design			
1.0	0.7	0.001	0.901	5.68	0.00	0.03
	0.8	0.118	0.900	5.00	0.12	2.40
	0.9	0.595	0.906	4.14	0.55	11.80
	1.0	0.811	0.886	3.67	0.68	15.63
	1.1	0.646	0.814	3.57	0.49	12.13
	1.2	0.284	0.801	3.63	0.20	5.21
	1.3	0.061	0.874	3.59	0.04	1.09
1.3	0.7	0.002	0.027	5.11	0.00	0.04
	0.8	0.149	0.049	4.49	0.14	2.93
	0.9	0.608	0.044	3.75	0.49	11.63
	1.0	0.804	0.021	3.35	0.59	14.93
	1.1	0.654	0.021	3.26	0.44	11.84
	1.2	0.315	0.026	3.30	0.19	5.57
	1.3	0.081	0.025	3.27	0.05	1.41

the new CVOT is modestly higher than the historical CVOT MAP estimates. When the historical CVOT MAP estimates are unbiased with respect to the generative model parameters in the new CVOT, we are able stop the study early approximately 80% of the time leading to a decrease in the mean duration of the new CVOT by more than half a year (for either design method). Even when the historical CVOT MAP estimates are off by as much as 10% relative to the baseline hazard in the generative model for the new CVOT, we are still able to stop the new CVOT over 50% of the time using the all-or-nothing design and even more often using the dynamic borrowing design.

They key differences in the two optimal designs are that the dynamic borrowing design tended to stop the new CVOT earlier leading to a slightly greater decrease in the average CVOT duration. However, the fixed borrowing design had higher minimum power when the bias of the historical CVOT MAP estimates was in the direction that compromised power. When discussing differences between the two designs we must point out the major caveat that the dynamic borrowing design characteristics are estimated with significantly lower precision than the all-or-nothing design characteristics. Moreover, the dynamic borrowing adaptive design is only optimal over the choices for $\pi^*(a_0)$ that we considered. Other choices for $\pi^*(a_0)$ may perform better.

5.7 Discussion

In this chapter we have developed two Bayesian adaptive design methodologies that allow one to borrow control events from a previously completed CVOT as a means to increase the efficiency of future CVOTs. We developed the all-or-nothing adaptive design which seeks to borrow a prespecified amount of events from the historical CVOT and that uses an interim analysis to evaluate whether borrowing is reasonable. As an alternative, we developed the dynamic borrowing adaptive design which allows the observed data to determine the amount of events that are borrowed from the historical CVOT and that uses the interim analysis to assess if enough events are being borrowed to justify stopping the study. For both designs, our simulation studies demonstrated that, when information borrowing is justifiable, the adaptive designs lead to future CVOTs that can be significantly shorter in duration. Our methods require subject-level data and we devoted significant discussion to the practical reasons why that is the case. We hope that our work can foster

discussion on subject-level data sharing, standardization of data collection, and (where reasonable) standardization of the disease populations that are targeted for recruitment into CVOTs.

CHAPTER 6: DNA METHYLATION DECONVOLUTION USING BISULFITE SEQUENCING DATA

6.1 Introduction

DNA methylation is the most widely studied epigenetic marker in mammals and a growing number of human diseases, including cancer, are associated with aberrant DNA methylation patterns (Robertson, 2005; Kulis and Esteller, 2010). In addition to its implications in human disease, DNA methylation has an important role in mammalian development (Jones and Takai, 2001) and has been shown to uniquely characterize functionally different cell and tissue types with high fidelity (Baron et al., 2006; Reinius et al., 2012; Varley et al., 2013).

Methylation data are usually collected from heterogeneous tissue samples, such as blood, which can be harvested in a cost effective manner. The tissue samples are heterogeneous in the sense that they are comprised of several (possibly many) functionally distinct cell types with each having a potentially distinct DNA methylation signature. For example, blood contains many different types of white blood cells, including neutrophils, basophils, eosinophils, lymphocytes, and monocytes. In addition to white blood cells, organs such as the liver can contribute significantly to the circulating pool of DNA in blood (Lo et al., 1998). Presumably, each cell type represented in a heterogeneous tissue sample contributes DNA for methylation analysis in accordance with its relative abundance in the tissue sample.

Methylation data collected from heterogeneous tissues is used in predominantly two ways: (1) to test for differential methylation or (2) to estimate the composition of the heterogeneous tissues using supplemental methylation data from reference tissue samples. Reference tissue samples are comprised predominately of one cell type though they are generally not perfectly homogeneous (Reinius et al. (2012), Supplemental Table 2).

When testing for differential methylation, if the cell types comprising the set of heterogeneous tissue samples have functionally distinct methylation signatures, as they do in blood (Reinius et al., 2012; Houseman et al., 2012), it is important to control for tissue composition in analyses performed at the tissue level. It has been argued that many purported associations between age and DNA methylation can be attributed to tissue composition bias that was poorly addressed in the analysis (Jaffe and Irizarry, 2014). One could argue that the far more relevant biological question is whether differential methylation exists in one or more of the constituent cell types in the heterogeneous tissue samples. However, no existing methods attempt to estimate cell type-specific methylation signatures for the purposes of associating phenotypes of interest with cell type-specific differential methylation. The method we develop herein is a first step towards filling this need.

Since our goal is to estimate cell type-specific methylation levels using reference samples, it is quite natural to simultaneously estimate tissue composition. Quadratic programming approaches are commonly applied to estimate tissue composition using data from high-throughput sequencing experiments. This technique has been recently applied by Sun et al. (2015) and Ziller et al. (2013). Quadratic programming techniques treat the reference tissue methylation levels as the known average methylation levels for the constituent cell types. Houseman et al. (2012) proposed an alternative method designed for the Infinium ®HumanMethylation27 BeadChip that acknowledges that reference tissues only provide an estimate of average methylation levels for the constituent cell types but this method requires reference tissue replication, making it impractical for high-throughput sequencing data. None of these methods produced updated estimates of cell type-specific methylation levels.

In this chapter we develop a reference-based statistical deconvolution approach for DNA methylation that is ideally suited for data from high-throughput bisfulfite sequencing experiments. By positing a joint model for the heterogeneous tissue sample data as well as the reference tissue sample data, our approach allows simultaneous estimation of both tissue composition and cell-type specific methylation levels. The rest of this chapter is organized as follows: In Section 6.2, we present a simple deconvolution model and discuss model fitting. We also discuss extensions of the simple deconvolution model to accommodate overdispersion in the cell type-specific methylation levels and discuss why using the more complicated model may be unnecessary in practice. In Sec-

tion 6.3, we use real data from DNA mixture experiments to demonstrate that estimates of tissue composition based on the proposed deconvolution model are superior to those from the typical quadratic programming approach and from an overdispersed deconvolution model. In Section 6.4 we use simulation studies to demonstrate that estimates of cell type-specific methylation levels can be significantly improved through deconvolution. We close the chapter with some discussion in Section 6.5.

6.2 A Simple Deconvolution Model

In this section we develop a reference-based deconvolution model that facilitates simultaneous estimation of heterogeneous tissue composition and average cell type-specific methylation. We note that the term "cell type" may be replaced with "homogenous tissue type" without any loss of generality (e.g. liver cell versus liver tissue). In many cases the latter term may be more appropriate. We assume that G loci are available for analysis and that a subset of K cell types comprise each heterogeneous tissue. We use locus as a generic term that may refer to a single CpG dinucleotide or a specified genomic region. We propose a deconvolution model for the individual DNA molecules on which methylation is measured. The model is comprised of two components: (1) a sample-specific latent multinomial allocation distribution that identifies the cell type of origin for each DNA molecule and (2) a locus and cell type-specific binomial methylation distribution that identifies whether or not each DNA molecule is methylated conditional on its cell type of origin.

Let k index cell type, g index locus, n index the heterogeneous tissue sample, and d index the DNA molecule. The latent multinomial cell type allocation model is given by

$$p\left(z_{ngd,1}, ..., z_{ngd,K} \mid \boldsymbol{\theta}_n\right) \propto \theta_{n,1}^{z_{ngd,1}} \cdots \theta_{n,K}^{z_{ngd,K}}$$

where $z_{ngd,k}$ is an indicator that the d^{th} DNA molecule covering locus g from heterogeneous tissue sample n originated from a type k cell and $\boldsymbol{\theta}_n = [\theta_{n,1}, \ldots, \theta_{n,K}]$ is a $K \times 1$ vector of tissue composition fractions that are non-negative and sum to one. The cell type allocation model arises naturally from the assumption that the cell type of origin for each DNA molecule is determined by the relative abundance of the cell type in the heterogeneous tissue. The binomial methylation model is given

by

$$p\left(y_{ngd} \mid z_{ngd,1}, ..., z_{ngd,K}\right) \propto \prod_{k=1}^{K} \left[\pi_{g,k}^{y_{ngd}} \left(1-\pi_{g,k}\right)^{1-y_{ngd}}\right]^{z_{ngd,k}}$$

where y_{ngd} is an indicator that the DNA molecule is methylated and $\pi_{g,k}$ is the methylation probability for type k cells at locus g. The complete data log-likelihood for the heterogeneous tissue samples is given by

$$\begin{aligned} &\ell\left(\boldsymbol{\pi}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N, \mathbf{z}_1, ..., \mathbf{z}_N \mid \mathbf{y}_1, ..., \mathbf{y}_N\right) \\ &= \sum_{n=1}^{N} \sum_{g=1}^{G} \sum_{d=1}^{D_{ng}} \sum_{k=1}^{K} \left(z_{ngd,k} \left[\log\left(\theta_{n,k}\right) + y_{ngd} \log\left(\pi_{g,k}\right) + \left(1-y_{ngd}\right) \log\left(1-\pi_{g,k}\right)\right]\right) \end{aligned} \tag{6.2.1}$$

where N is the total number of heterogeneous tissue samples, D_{ng} is the total number of DNA molecules covering locus g in heterogeneous tissue sample n, $\mathbf{y}_n = \{y_{ngd} : \forall(g,d)\}$, $\mathbf{z}_n = \{z_{ngd,k} : \forall(g,d,k)\}$, and $\boldsymbol{\pi} = \{\pi_{g,k} : \forall(g,k)\}$.

For the reference tissue samples, the cell type of origin for each DNA molecule is known (by assumption) and so only the binomial methylation model is relevant for these samples. In many cases the sequencing depth in the reference tissue samples will be much greater than in the heterogeneous tissue samples. We have found it to be useful to down weight the reference samples so they do not dominate estimation of the cell type-specific methylation probabilities. The weighted log-likelihood for the reference tissue samples is given by

$$\ell\left(\boldsymbol{\pi} \middle| \left\{\mathbf{y}_n^k : \forall(n,k)\right\}\right) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \sum_{g=1}^{G} \omega_{ng}^k \left(\sum_{d=1}^{D_{ng}^k} y_{ngd}^k \log\left(\pi_{g,k}\right) + \left(1-y_{ngd}^k\right) \log\left(1-\pi_{g,k}\right)\right) \tag{6.2.2}$$

where N_k is the total number of reference tissue samples comprised of type k cells, D_{ng}^k is the total number of DNA molecules covering locus g in cell type k reference tissue sample n, ω_{ng}^k is the weight given to the set of DNA molecules covering locus g, y_{ngd}^k is an indicator that DNA molecule d is methylated, and $\mathbf{y}_n^k = \left\{y_{ngd}^k : \forall(g,d)\right\}$.

One combines (6.2.1) and (6.2.2) to form the full complete data log-likelihood. Unfortunately, since the variables $\mathbf{z}_1, ..., \mathbf{z}_N$ are not observed, the likelihood cannot be directly maximized. However, one can derive a simple expectation-maximization (EM) algorithm (Dempster et al., 1977)

to optimize the associated observed data likelihood, which would theoretically be obtained by summing out $\mathbf{z}_1, ..., \mathbf{z}_N$.

6.2.1 Model Fitting via the EM Algorithm

The EM algorithm is comprised of two steps that are completed in an iterative loop: the E-step and the M-step. To complete the E-step, one needs to compute the expectation of (6.2.1) with respect to the conditional distribution of $\mathbf{z}_n$ given $\mathbf{y}_n$ using the estimates of the parameters at the current iteration of the algorithm. This is equivalent to replacing $z_{ngd,k}$ in (6.2.1) with its conditional expectation given y_{ngd}. It is not difficult to show that

$$\mathrm{E}\left[z_{ngd,k}\middle|y_{ngd}=1,\boldsymbol{\theta}_n^{(t)},\boldsymbol{\pi}_g^{(t)}\right]=\psi_{ngk,1}^{(t)}=\frac{\theta_{n,k}^{(t)}\pi_{g,k}^{(t)}}{\sum_{k'=1}^{K}\theta_{n,k'}^{(t)}\pi_{g,k'}^{(t)}} \tag{6.2.3}$$

and

$$\mathrm{E}\left[z_{ngd,k}\middle|y_{ngd}=0,\boldsymbol{\theta}_n^{(t)},\boldsymbol{\pi}_g^{(t)}\right]=\psi_{ngk,0}^{(t)}=\frac{\theta_{n,k}^{(t)}\left(1-\pi_{g,k}^{(t)}\right)}{\sum_{k'=1}^{K}\theta_{n,k'}^{(t)}\left(1-\pi_{g,k'}^{(t)}\right)} \tag{6.2.4}$$

where $\boldsymbol{\theta}_n^{(t)}$ is the current estimate of the tissue composition fractions for heterogeneous tissue sample n and $\boldsymbol{\pi}_g^{(t)}$ represents the current estimates for the methylation probabilities for the K cell types at locus g.

The M-step maximizes the expected full complete data likelihood. One can obtain closed-form updates for both sets of parameters. The update for the cell type k tissue composition fraction in heterogeneous tissue sample n is given by

$$\theta_{n,k}^{(t+1)}=\frac{\sum_{g=1}^{G}D_{ng}\times\left(\psi_{ngk,1}^{(t)}+\psi_{ngk,0}^{(t)}\right)}{\sum_{k'=1}^{K}\sum_{g=1}^{G}D_{ng}\times\left(\psi_{ngk',1}^{(t)}+\psi_{ngk',0}^{(t)}\right)}$$

and the update for the methylation probability for type k cells at locus g is given by

$$\pi_{g,k}^{(t+1)}=\frac{\sum_{n=1}^{N}\left(\psi_{ngk,1}^{(t)}\sum_{d=1}^{D_{ng}}y_{ngd}\right)+\sum_{n=1}^{N_k}\left(\omega_{ng}^{k}\sum_{d=1}^{D_{ng}}y_{ngd}^{k}\right)}{\sum_{n=1}^{N}\left(\psi_{ngk,1}^{(t)}\sum_{d=1}^{D_{ng}}y_{ngd}+\psi_{ngk,0}^{(t)}\left[D_{ng}-\sum_{d=1}^{D_{ng}}y_{ngd}\right]\right)+\sum_{n=1}^{N_k}\omega_{ng}^{k}D_{ng}^{k}}.$$

The EM algorithm is guaranteed to find a locally optimal solution but can be sensitive to the initial values for the parameters when the observed data likelihood is multimodal. In practice, we initialize the methylation probability for type k cells at locus g to the value

$$\pi_{g,k}^{(0)} = \frac{\sum_{n=1}^{N_k} \omega_{ng}^k \sum_{d=1}^{D_{ng}^k} y_{ngd}^k}{\sum_{n=1}^{N_k} \omega_{ng}^k D_{ng}^k}$$

and all tissue composition fractions to $1/K$. We have found the algorithm to be quite robust to the initial values used for the tissue composition fractions.

6.2.2 On Overdispersion in the Cell Type-Specific Methylation Model

Modeling overdispersion is a staple in the analysis of methylation data from bisulfite sequencing experiments in situations where tissue composition is ignored (Hebestreit et al., 2013; Dolzhenko and Smith, 2014; Feng et al., 2014; Sun et al., 2014). It is reasonable to assume that the presence of overdispersion in these cases is at least partially attributable to ignoring tissue composition but it is also highly plausible that overdispersion exists for other reasons. Initially we sought to accommodate overdispersion in our deconvolution model. This was done by using a beta-binomial model for the cell type-specific methylation levels instead of a binomial model.

Fitting the overdispersed deconvolution model cannot be done with a simple EM algorithm. To fit the overdispersed deconvolution model, we implemented a variational EM algorithm similar to that proposed by Blei et al. (2003) for estimation of the latent Dirichlet allocation model. The variational EM algorithm does not maximize the observed data likelihood but instead maximizes a tractable lower bound on the observed data likelihood. In simulation studies we observed that, unless the number of replicates for the reference samples were large (which will never be the case), it was impossible to estimate the overdispersion parameters with any reliability. The dispersion parameter estimates obtained from the variation EM algorithm were consistent with a binomial assumption even when the data were simulated from highly overdispersed beta-binomial distributions. Moreover, we observed that the estimates of the cell type-specific methylation probabilities were no better than those produced by fitting the simple deconvolution model described above. In light of these findings, we elected to pursue the simplified deconvolution model. The computational

complexity of fitting the overdispersed deconvolution model using a variational EM algorithm is orders of magnitude greater than fitting the simple deconvolution model using an EM algorithm. The results we present in Sections 6.3 and 6.4 provide evidence that fitting the overdispersed deconvolution model with a variational EM algorithm leads to inferior estimates compared to those from the simple deconvolution model we have proposed (even when the overdispersed deconvolution model is correct).

6.3 Estimation of Tissue Composition in DNA Mixture Experiments

In this section we present results from an analysis of a set of heterogeneous tissues that were created by Sun et al. (2015) by mixing known quantities of DNA from several cell types. Sun et al. (2015) constructed 18 DNA mixture tissues using buffy coat DNA (i.e. white blood cells), placental DNA, and liver DNA. The buffy coat DNA was obtained from a 40 year old, non-pregnant woman. The placental DNA was obtained immediately following the delivery of a healthy female baby having a gestational age of 38 weeks. The liver DNA was obtained from the non-neoplastic liver tissue of a 57 year old female subject during resection of a hepatocellular carcinoma.

We obtained unprocessed bisulfite sequencing data for the DNA mixture tissues from Sun et al. (2015), unprocessed bisulfite sequencing data for a reference placental tissue from Lun et al. (2013), unprocessed bisulfite sequencing data for a reference liver tissue from the NCBI Gene Expression Omnibus (GSM1058027), and unprocessed bisulfite sequencing data for a reference B cell tissue and a reference neutrophil tissue from Hodges et al. (2011). We obtained processed methylation calls for CD4+T cells and CD8+T cells from Ziller et al. (2013). Unfortunately, unprocessed bisulfite sequencing data are not publicly available for the T cell data. All unprocessed bisulfite sequencing data were then uniformly processed using Bismark (Krueger and Andrews, 2011) version 0.14.4 and aligned to the hg19/GRCh37 reference assembly, which matched the reference assembly used by Ziller et al. (2013).

We restricted our attention to methylation calls for 5,820 genomic loci that were identified by Sun et al. (2015) as being informative for one or more of 14 reference cell types (which include the cell types used to construct the DNA mixture tissues). Each locus corresponded to a genomic

region that was 500bp in length. Complete details on the genomic location of each region can be found in Supplemental Table 1 from Sun et al. (2015). We combined the number of methylated calls and the number of unmethylated calls over all CpGs in a region to form the observed data. The average sequencing depth was 67 for the 18 heterogeneous tissues, 184 for the B cell reference tissue, 330 for the neutrophil reference tissue, 2028 for the CD4+T cell reference tissue, 2118 for the CD8+T cell reference tissue, 1039 for the placenta reference tissue, and 1096 for the liver reference tissue. We excluded 103 loci that were not covered in all 18 DNA mixture tissues and/or that had extremely high or low sequencing depth compared to the average depth in the 18 heterogeneous tissues. This resulted in 5717 loci that were used for analysis.

For the weighted EM algorithm (used to fit the simple deconvolution model), the weight given to a particular reference sample at a given locus was equal to the ratio of the average sequencing depth from the 18 DNA mixture tissues to the sequencing depth in the reference tissue. We also considered an unweighted approach but observed superior performance when using weights (data not known). The EM algorithm was terminated when the maximum change in all parameters was less than 10^{-5}. We implemented the quadratic programming approach using the SAS System (v9.4, OPTMODEL Procedure) using the most stringent convergence criteria. Following Sun et al. (2015), the proportion of all calls in a region that were methylated was used as the response in the linear system of equations. We also fit the overdispersed deconvolution model using a variational EM algorithm. The overdispersed deconvolution model assumed a separate overdispersion parameter for each cell type and locus. The algorithm was terminated after 15,000 iterations and all parameter estimates were stable except for the dispersion parameters. Estimates of the dispersion parameters were all very large (large values imply no overdispersion).

Figure 6.1 presents a plot of the estimated versus the true cell type composition fractions and corresponding linear regression curves for white blood cells, liver, and placenta in the 18 DNA mixture tissues. Each point on the figure corresponds to the estimate from one of the 18 samples. The true composition fractions were assumed to equal the DNA fraction used in the mixtures as reported by Sun et al. (2015). The quadratic programming approach resulted in consistent over estimation of the white blood cell fraction in cases where the true fraction was small. Correspondingly, there was consistent under estimation of the liver fraction when the true fraction

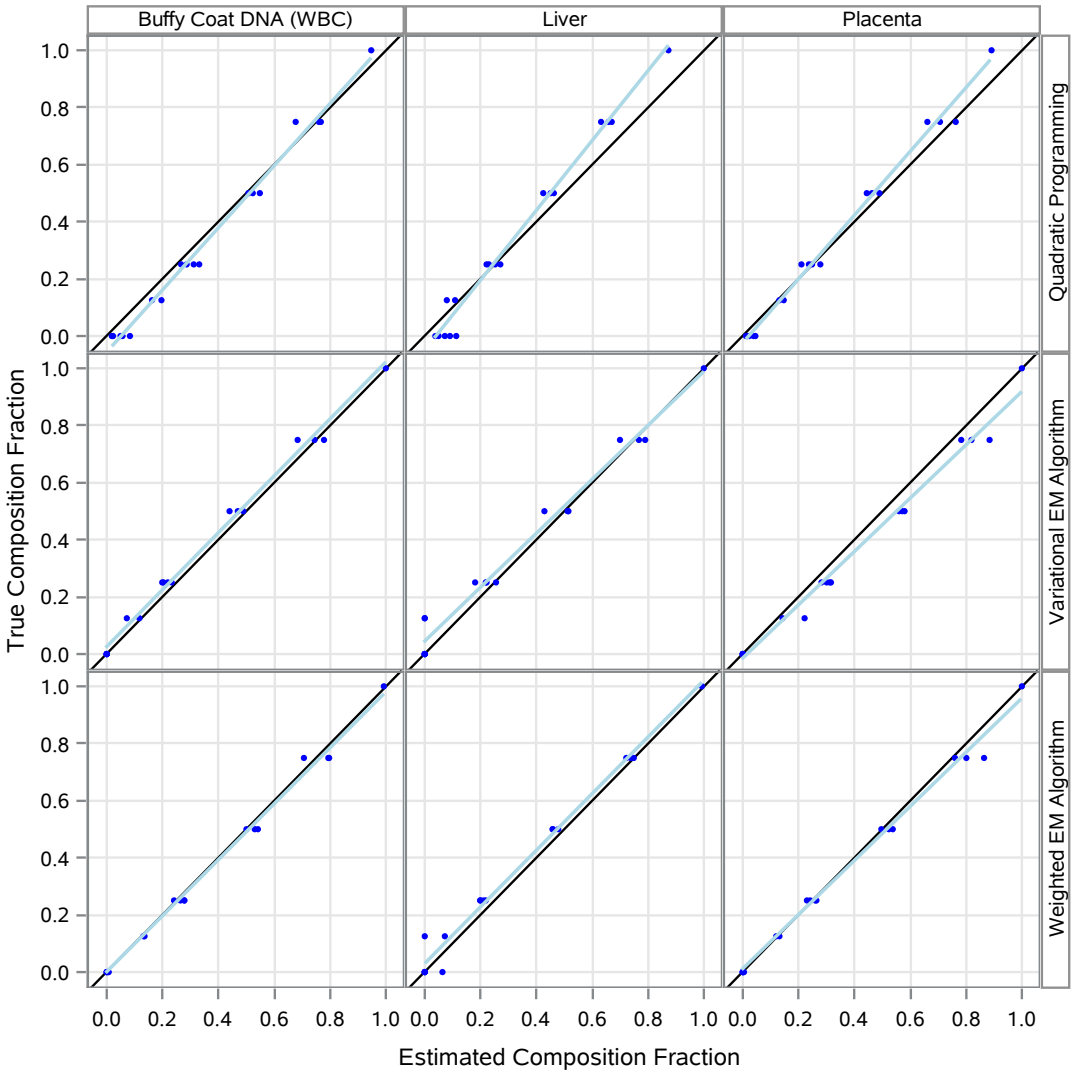

Figure 6.1: Estimated cell type composition in 18 DNA mixture tissues

Table 6.1: DNA Mixture tissue composition fraction estimate quality

	Quadratic Programming		Overdispersed Deconvolution Model		Simple Deconvolution Model
Cell Type	MAD	Ratio	MAD	Ratio	MAD
White Blood Cells	0.042	2.341	0.023	1.272	0.018
Liver	0.060	1.864	0.033	1.025	0.032
Placenta	0.033	1.914	0.043	2.452	0.017

was large. Results from the simple deconvolution model and the overdispersed deconvolution model (to a lesser extent), corrected this bias.

To provide a more quantitative assessment of the relative quality of the estimates provided by the simple deconvolution model, we computed the mean absolute deviation (MAD) for the estimated composition fraction for each cell type relative to the true composition fraction using each of the three approaches discussed. Table 6.1 provides the mean absolute deviations as well as the ratio of the mean absolute deviations for the other two methods to the mean absolute deviation from the simple deconvolution model. We see that estimates based on the simple deconvolution model are more accurate on average than those from the other two approaches. The mean absolute deviation from the quadratic programming approach is approximately twice that from the simple deconvolution model for each cell type. The mean absolute deviation for the overdispersed deconvolution model is nearly the same as that from the the simple deconvolution model for liver cells but is worse than both other methods for placenta cells.

6.4 Estimation of Tissue Composition and Cell Type-Specific Methylation in Simulation Studies

In the DNA mixture example from Section 6.3, we were not able to assess the relative quality of the simple deconvolution model estimates for the cell type-specific methylation probabilities. In this section we use simulation studies to demonstrate that the estimates for this set of parameters (and composition fractions) can improve significantly when one incorporates information from heterogeneous tissues through deconvolution. To accomplish this, we compare the quality of estimates based on the simple deconvolution model to those obtained from the overdispersed deconvolution

model and from quadratic programming. Note that all three methods allow estimation of the tissue composition fractions but the quadratic programming approach does not produce estimates of the cell type-specific methylation probabilities. Thus, for these parameters, we compare estimates from the two deconvolution methods to the proportion of methylated calls in the reference samples.

We simulated bisulfite sequencing data for 25 independent studies, each with $G = 1000$ loci and $N = 30$ heterogeneous tissue samples. The heterogeneous tissues were simulated to have neutrophil, B cell, CD4+T cell, CD8+T cell, and placenta fractions. The composition of the heterogeneous tissues was simulated so that the tissues were 40% neutrophils, 30% placenta, 10% B cells, 10% CD4+T cells, and 10% CD8+T cells, on average. This composition is roughly consistent with blood samples taken from pregnant females. The 1,000 loci were randomly selected from the set of 5,280 loci identified by Sun et al. (2015). Using the reference tissue samples discussed in Section 6.3, we set the cell-type specific methylation probabilities at each locus equal to the proportion of methylated calls at the locus in the reference tissue sample. For each simulated heterogeneous tissue sample, the sequencing depth at each locus was randomly drawn from a Poisson distribution having mean parameter equal to the average sequencing depth at the locus from the DNA mixture tissues discussed in Section 6.3. The sequencing depth in each simulated reference tissue sample was set to the sequencing depth in the actual reference tissue sample discussed in Section 6.3. A single reference sample was simulated for each cell type (independently for each simulation study). The methylation calls for the set of DNA molecules covering a locus (from a given cell type) were simulated using a beta-binomial model having an overdispersion parameter value of 30 (larger values mean less variability). The chosen value of the dispersion parameter corresponds to a relatively large but realistic amount of overdispersion (Feng et al., 2014).

For the weighted EM algorithm, we weighted each reference sample as described in Section 6.3. The quadratic programming approach was implemented as described in Section 6.3 (to estimate composition fractions). In these simulations, data were generated from the overdispersed deconvolution model. Thus, if there is a benefit to fitting that model using a variational EM algorithm, we would expect to observe better estimates compared to the simple deconvolution model. We will see that is not the case.

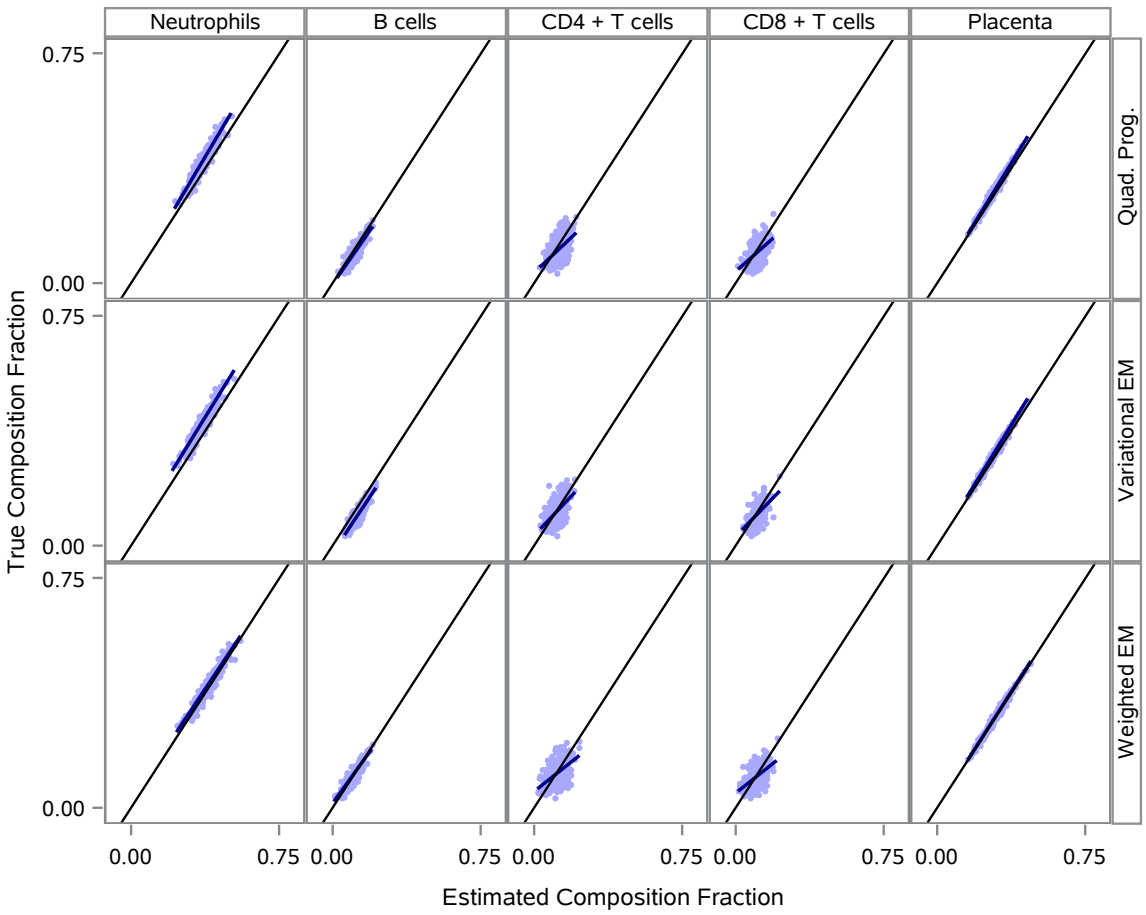

Figure 6.2: Estimated cell type composition in simulated heterogeneous tissues

Figure 6.2 presents a plot of the estimated versus true cell type composition fractions and corresponding linear regression curves for a randomly selected subset of 250 (out of 750) simulated heterogeneous tissues. Each point on the figure corresponds to the estimated composition fraction for a particular cell type for one of the selected samples. While estimates from all three methods appear qualitatively similar for every cell type, one can clearly see that the estimates from the simple deconvolution model have less bias than the estimates from the other two methods for the neutrophil fraction. The other obvious observation is that none of the three methods do well at estimating the fraction of CD4+T cells and CD8+T cells. This is not surprising since the methylation signatures for these two cell types are highly similar (Reinius et al., 2012). This suggests that including reference samples for cell types that do not have unique methylation signatures may not adversely

Table 6.2: Tissue composition fraction estimate quality

	Quadratic Programming		Overdispersed Deconvolution Model		Simple Deconvolution Model
Cell Type	MAD	Ratio	MAD	Ratio	MAD
Neutrophils	0.031	2.083	0.039	2.595	0.015
B cells	0.017	1.607	0.030	2.804	0.011
CD4+T cells	0.031	1.061	0.027	0.928	0.029
CD8+T cells	0.025	1.006	0.023	0.943	0.025
Placenta	0.011	1.957	0.010	1.697	0.006

Table 6.3: Cell type-specific methylation estimate quality

	Reference Only		Overdispersed Deconvolution Model		Simple Deconvolution Model
Cell Type	MAD	Ratio	MAD	Ratio	MAD
Neutrophils	0.0017	1.434	0.0014	1.197	0.0012
B cells	0.0020	1.045	0.0020	1.034	0.0019
CD4+T cells	0.0020	1.048	0.0020	1.070	0.0019
CD8+T cells	0.0021	1.035	0.0022	1.106	0.0020
Placenta	0.0023	1.433	0.0016	1.159	0.0016

effect parameter estimates for cell types that do. In cases where it is of interest to study methylation from a solid tissue (e.g. placenta or liver) using blood samples, this property is quite appealing.

Table 6.2 provides the average mean absolute deviation of the composition fraction estimates for each cell type and deconvolution method. The average was computed over the 25 simulation studies using all simulated heterogeneous tissue samples. For the quadratic programming and overdispersed deconvolution model approaches, the ratio of the average mean absolute deviation to the corresponding average mean absolute deviation from the simple deconvolution model is provided as well. For neutrophils, B cells, and placenta, the composition fraction estimates based on the simple deconvolution model are appreciably better (on average) than those from the other two methods. For CD4+T cells and CD8+T cells, the estimates from all three deconvolution methods are comparable.

Table 6.3 provides similar information for the cell type-specific methylation probability estimates for each deconvolution method. For neutrophils and placenta, on average the estimates based on the simple deconvolution model are appreciably better than those based entirely on the reference samples and modestly better than those from the overdispersed deconvolution model. The fact

that the simple deconvolution model improves estimates for neutrophils and placenta methylation probabilities more than for the other cell types is expected. Since, B cells, CD4+T cells, and CD8+T cells each only represent 10% of the heterogeneous tissue samples, one should not expect the same level of improvement as seen in neutrophils and placenta, which represented 40% and 30% of the heterogeneous tissue samples, respectively. There is simply less information in the heterogeneous tissue samples about methylation levels in B cells, CD4+T cells, and CD8+T cells.

6.5 Discussion

In this chapter we developed a reference-based deconvolution model for DNA methylation data and proposed an easy-to-implement EM algorithm that can be used to fit the model. Using real and simulated data, we compared the quality of the tissue composition fraction estimates based on the proposed deconvolution model to those based on a quadratic programming approach and based on an overdispersed deconvolution model. Using simulated data, we also compared the quality of the cell-type specific methylation probability estimates based on the proposed deconvolution model to those based on the overdispersed deconvolution model and to those based only on reference samples. We observed that the estimates produced by our deconvolution model were seemingly always better than those produced by the other approaches considered. These results suggest that the proposed deconvolution model could be a valuable alternative to quadratic programming techniques for tissue composition estimation.

In our results from Section 6.4 we noted that all methods considered were unable to estimate the composition of CD4+T cells and CD8+T cells and we posited that this is because of the similarity of the methylation signatures for these cell types. It is clear that incorporation of both those cell types in the deconvolution model did not adversely effect composition estimation (or methylation probability estimation) for neutrophils or placenta which both are known to have unique methylation signatures and which both represented larger fractions of the heterogeneous tissues. This suggests that one might choose a conservative set of blood cell references (e.g. B cells, T cells, neutrophils, monocytes, eosinophils, etc.) to include in the deconvolution model when the goal is to estimate the blood composition fraction or methylation levels of a tissue like placenta

or liver whose DNA can be found in blood in reasonable amounts. For example, if the goal of a hypothetical study is to associate a fetal disease phenotype with differential placental methylation, it would seem possible to do this with the proposed deconvolution model using a set of blood cell reference tissues, one reference placenta tissue for a case, one reference placenta tissue for a control, and then a larger number of heterogeneous blood tissue samples from pregnant mothers. A hypothesis testing framework for this type of problem is a topic of future work.

CHAPTER 7: FUTURE WORK

7.1 Bayesian Clinical Trial Design

There are a variety of ways that the methods we have developed for Bayesian clinical trial design can be extended. First, there is the obvious extension of the methodology developed in Chapters 3 and 4 to other, more advanced types of historical data. Our work thus far has focused on basic time-to-event data but there is significant interest in recurrent events data with a terminating death event (Rondeau, 2010) and data with a survival and non-survival component (Ibrahim et al., 2010). Each of these types of data is associated with a more complicated statistical model than what we considered in Chapters 3 and 4. Due to this added complexity, customized approaches are needed to make Bayesian design computationally feasible.

The adaptive design methodology that we developed in Chapter 5 can also be extended to accommodate more advanced data types (at least in the case of the all-or-nothing adaptive design). In addition to extending the proposed adaptive design methods to other data types, one might also consider alternative criteria for when the study should be stopped. One promising idea is to use the prior predictive distribution for the data to assess prior-data conflict (Evans et al., 2006). This approach would seem more Bayesian than what we have proposed but its application has not been considered outside of very simple models.

The all-or-nothing adaptive design would seem widely applicable in situations where a single historical study is available. While the notion of using an informative prior and checking the exchangeability assumption at an interim analysis is intuitive, it has not been widely adopted. Our adaptive design results demonstrate that such an approach may be as effective as more complicated procedures that attempt dynamic information borrowing. In our opinion there is much that can be done within the framework of the all-or-nothing adaptive design due to relative simplicity of that

approach.

7.2 DNA Methylation Deconvolution

Our work to date on the DNA methylation deconvolution problem serves as a proof-of-concept for the idea that one can learn about cell type-specific methylation levels through the process of deconvolution beyond what is known from reference samples. Our long-term goal is to develop a hypothesis testing framework that can be used to identify differential cell type-specific methylation for one or more conditions under study using heterogeneous tissue samples. Exploring the extent to which this type of approach requires reference samples for each condition under study is an important intermediate goal. If reference samples are required for each condition, the practicality of such a method may be limited. However, if under reasonable assumptions it is sufficient to have reference tissue samples for only the healthy condition, a method of this type could prove valuable.

APPENDIX A: CHAPTER 3 SUPPLEMENTAL MATERIALS

A.1 Type I Error Control with Informative Priors

For this example we suppose data $\boldsymbol{D} = \{x_i, i = 1, ..., N_x\}$ from N_x subjects are available for the new study and that $X_i \sim \text{Normal}\left(\mu, \sigma^2\right)$ for $i = 1, ..., N_x$ with σ^2 known. We consider testing the interval null hypothesis as $H_0 : \mu \leq 0$ versus alternative $H_1 : \mu > 0$. Further suppose historical data $\boldsymbol{D}_0 = \{y_i, i = 1, ..., N_y\}$ are available form subjects assumed to be exchangeable with those from the new study. Based on assuming the objective initial prior $\pi_0(\mu) \propto 1$, the basic power prior has the form

$$\pi_0\left(\mu \middle| \boldsymbol{D}_0, a_0\right) \sim \text{Normal}\left(\bar{Y}, \frac{\sigma^2}{N_y a_0}\right) \tag{A.1.1}$$

where $\bar{Y}$ is the sample mean based on the historical study subjects. Thus, the value of a_0 has the interpretation as the fraction of information (i.e. Fisher Information) that is borrowed from the historical study. The corresponding posterior distribution is straight forward to derive and is given as follows:

$$\pi\left(\mu \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \quad \propto \quad \text{Normal}\left(\mu_*, \sigma_*^2\right)$$

with

$$\sigma_*^2 \quad = \quad \frac{1}{\frac{N_x}{\sigma^2} + \frac{N_y a_0}{\sigma^2}}$$

and

$$\mu_*^2 \quad = \quad \sigma_*^2 \left(\frac{N_x}{\sigma^2} \cdot \bar{X} + \frac{N_y a_0}{\sigma^2} \cdot \bar{Y}\right)$$

where $\bar{X}$ is the sample mean from new study. We will accept the alternative hypothesis when

$$P\left(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \phi_{a_0}$$

where ϕ_{a_0} is the posterior probability critical value associated with the chosen value of a_0. We want to compare the statistical power associated with different values of a_0 when the posterior probability critical values are chosen to provide a test procedure that controls the frequentist type I error rate at level α. First, we consider the case where no information is borrow so that we are effectively performing an objective Bayesian analysis of the new data. We show that the Bayesian test is equivalent to the frequentist uniformly most powerful (UMP) test.

Lemma A.1.1. *Inference based on the Bayesian rejection rule*

$$P\left(\mu > 0 \,|\, \boldsymbol{D}, \boldsymbol{D}_0, a_0 = 0\right) \equiv P\left(\mu > 0 \,|\, \boldsymbol{D}\right) \geq \phi_0$$

is identical to the frequentist one-sided, level α UMP test when ϕ_0 is $1 - \alpha$.

Proof. In this setting we view the posterior probability as a test statistic in the sense that it is simply a random function of the data. The type I error rate of the test is

$$\mathrm{E}\left[1\left\{P\left(\mu > 0 \,|\, \boldsymbol{D}\right) \geq \phi_0\right\} \,|\, \mu_{\text{true}} = 0\right] \tag{A.1.2}$$

where the expectation is with respect to the data $\boldsymbol{D}$. We want to find ϕ_0 so that (A.1.2) is equal to α. To do this we manipulate the quantity inside the indicator to show that it is equivalent to the frequentist UMP rejection criteria. Note that in the posterior probability, the data are fixed and not random. We have that

$$\begin{aligned}
P\left(\mu > 0 \,|\, \boldsymbol{D}\right) \geq \phi_0 \quad &\Longleftrightarrow \quad P\left(\frac{\mu - \bar{X}}{\frac{\sigma}{\sqrt{N_x}}} > \frac{-\bar{X}}{\frac{\sigma}{\sqrt{N_x}}} \middle| \boldsymbol{D}\right) \geq \phi_0 \\
&\Longleftrightarrow \quad P\left(Z_1 > \frac{-\bar{X}}{\frac{\sigma}{\sqrt{N_x}}} \middle| \boldsymbol{D}\right) \geq \phi_0 \\
&\Longleftrightarrow \quad P\left(Z_2 < \frac{\bar{X}}{\frac{\sigma}{\sqrt{N_x}}} \middle| \boldsymbol{D}\right) \geq \phi_0 \\
&\Longleftrightarrow \quad \Phi\left(\frac{\bar{X}}{\frac{\sigma}{\sqrt{N_x}}}\right) \geq \phi_0 \\
&\Longleftrightarrow \quad \frac{\bar{X} - 0}{\frac{\sigma}{\sqrt{N_x}}} \geq \Phi^{-1}\left(\phi_0\right)
\end{aligned} \tag{A.1.3}$$

where Z_1 and Z_2 are standard normal random variables. The expression in (A.1.3) is precisely the rejection rule for the one-sided normal UMP test. Thus, the optimal choice of ϕ_0 satisfies $\Phi^{-1}(\phi_0) = Z_{1-\alpha}$ which implies $\phi_0 = 1 - \alpha$. □

We have shown that for $a_0 = 0$ the most powerful (unbiased) Bayesian test procedure requires $\phi_0 = 1 - \alpha$. It should be intuitive that for $a_0 > 0$ we must then have $\phi_{a_0} > 1 - \alpha$ in order to maintain control type I error (when $\bar{Y} > 0$). We show this now.

Lemma A.1.2. *In order for inference based on the Bayesian rejection rule*

$$P(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \geq \phi_{a_0}$$

to control type I error at level α when the true mean is zero, we must have

$$\phi_{a_0} = \Phi\left[\left(\frac{\Phi^{-1}(1-\alpha)}{\frac{\sigma}{\sqrt{N_x}}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}}\right)\sigma^*\right] > 1 - \alpha. \tag{A.1.4}$$

Proof. We must solve the equation

$$\mathrm{E}\left[1\left\{P(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \geq \phi_{a_0}\right\} \mid \mu_{\text{true}} = 0, \boldsymbol{D}_0, a_0\right] = \alpha$$

for ϕ_{a_0}. Just as before, we manipulate the posterior probability. We have the following:

$$\begin{aligned}
P(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \geq \phi_{a_0} \quad &\iff \quad P\left(\frac{\mu - \mu^*}{\sigma^*} > \frac{-\mu^*}{\sigma^*}\middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \phi_{a_0} \\
&\iff \quad P\left(Z_1 > \frac{-\mu^*}{\sigma^*}\middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \phi_{a_0} \\
&\iff \quad P\left(Z_2 < \frac{\mu^*}{\sigma^*}\middle| \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \phi_{a_0} \\
&\iff \quad \Phi\left(\frac{\mu^*}{\sigma^*}\right) \geq \phi_{a_0} \\
&\iff \quad \frac{\bar{X}}{\frac{\sigma^2}{N_x}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}} \geq \frac{\Phi^{-1}(\phi_{a_0})}{\sigma^*} \\
&\iff \quad \frac{\bar{X} - 0}{\frac{\sigma}{\sqrt{N_x}}} \geq \frac{\sigma}{\sqrt{N_x}}\left(\frac{\Phi^{-1}(\phi_{a_0})}{\sigma^*} - \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}}\right)
\end{aligned} \tag{A.1.5}$$

The expression in (A.1.5) is precisely the rejection rule for the one-sided frequentist UMP test. Thus, for the Bayesian decision rule to be optimal, ϕ_{a_0} must satisfy the following equation.

$$Z_{1-\alpha} = \Phi^{-1}(1-\alpha) = \frac{\sigma}{\sqrt{N_x}} \left(\frac{\Phi^{-1}(\phi_{a_0})}{\sigma^*} - \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}} \right) \tag{A.1.6}$$

To obtain the equality in (A.1.4), one must simply solve (A.1.6) for ϕ_{a_0}. To show that $\phi_{a_0} > 1-\alpha$, we reformulate the expression given in (A.1.4) as follows.

$$\begin{aligned}
\phi_{a_0} &= \Phi\left[\left(\frac{\Phi^{-1}(1-\alpha)}{\frac{\sigma}{\sqrt{N_x}}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}} \right) \sigma^* \right] \\
&= \Phi\left[\Phi^{-1}(1-\alpha) \left(\frac{1}{\frac{\sigma}{\sqrt{N_x}}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0} \cdot \Phi^{-1}(1-\alpha)} \right) \sqrt{\frac{1}{\frac{N_x}{\sigma^2} + \frac{N_y \cdot a_0}{\sigma^2}}} \right] \\
&= \Phi\left[\Phi^{-1}(1-\alpha) \cdot h(a_0) \right]
\end{aligned}$$

It is enough to show that $h(a_0)$ is increasing in a_0 and that $\lim_{a_0 \to 0} h(a_0) = 1$ for fixed $\bar{Y} > 0$. Doing so is straightforward and is omitted for brevity. □

Thus far we have shown that incorporating any amount of prior information that favors the alternative hypothesis will inflate the type I error of the test procedure unless we increase the posterior probability critical value to the quantity defined by ϕ_{a_0}. Now we show that by using the correct posterior probability critical value to control type I error we effectively discard all prior information resulting in a power function identical to that obtained by the Bayesian analysis with $a_0 = 0$.

Theorem 0.1. *The power functions for the set of Bayesian hypothesis tests based on the rejection rules* $\{P(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0) \geq \phi_{a_0} : a_0 \in [0,1]\}$ *are identical. Hence, all prior information is effectively discarded when the rejection rule based on* $a_0 > 0$ *is calibrated to ensure frequentist I error control.*

Proof. The power function associated with a fixed value of a_0 and corresponding posterior proba-

bility critical value ϕ_{a_0} is given as follows:

$$\mathrm{E}\left[1\left\{P\left(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \phi_{a_0}\right\} \mid \boldsymbol{D}_0, a_0, \mu_{\text{true}}\right] \tag{A.1.7}$$

where the expectation is taken with respect to the distribution for the new data $\boldsymbol{D}$. To show equality of the power functions for all a_0, it is enough to show that the power function for an arbitrary a_0 is the same as the power function for $a_0 = 0$. Further, it is enough to show equality of the corresponding events in the indicator functions since the expectations are with respect to the same distribution for the data.

$$\begin{aligned}
P\left(\mu > 0 \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \geq \phi_{a_0} \quad &\Longleftrightarrow \quad \Phi\left(\frac{\mu^*}{\sigma^*}\right) \geq \Phi\left[\left(\frac{\Phi^{-1}(1-\alpha)}{\frac{\sigma}{\sqrt{N_x}}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}}\right)\sigma^*\right] \\
&\Longleftrightarrow \quad \frac{\mu^*}{\sigma^*} \geq \left(\frac{\Phi^{-1}(1-\alpha)}{\frac{\sigma}{\sqrt{N_x}}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}}\right)\sigma^* \\
&\Longleftrightarrow \quad \frac{\bar{X}}{\frac{\sigma^2}{N_x}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}} \geq \frac{\Phi^{-1}(1-\alpha)}{\frac{\sigma}{\sqrt{N_x}}} + \frac{\bar{Y}}{\frac{\sigma^2}{N_y \cdot a_0}} \\
&\Longleftrightarrow \quad \frac{\bar{X} - 0}{\frac{\sigma}{\sqrt{N_x}}} \geq \Phi^{-1}(1-\alpha) \\
&\Longleftrightarrow \quad P\left(\mu > 0 \mid \boldsymbol{D}\right) \geq \phi_0
\end{aligned}$$

□

We have shown in a simple case that when an informative prior is used in a Bayesian analysis where the decision rule is calibrated to control type I error in the frequentist sense, all prior information is effectively discarded resulting in a test procedure that is identical to Bayesian analysis with a objective prior, or in this case, the frequentist UMP test. Although the results derived are for a very simple model, it is likely the behavior holds in all cases for which a level α hypothesis test is available.

A.2 Bayesian Design with Unshared Parameters

In this appendix we derived the marginal posterior for the hazard ratio regression parameters for the scenario where the historical study model and the new study model have an identical set of parameters. We refer to this scenario as the identical model scenario (IM). This will not always be the case. In this section, we derive the marginal posterior for the hazard ratio regression parameters in two additional scenarios that may be of interest to practitioners.

In the first scenario, we assume that the baseline hazard differs between the historical study and the new study so that only the hazard ratio regression parameters are shared between the models. We refer to this scenario as the shared hazard ratio scenario (SHR). This scenario might arise when one wishes to provide a mechanism for the historical and new studies to differ systematically in terms of baseline risk yet still allow information borrowing through the hazard ratio regression parameters where it is reasonable to assume a common effect across studies. For example, one may choose to stratify by geographic region in an effort to account for temporal changes in baseline risk and adjust for more concrete biomarkers in the hazard ratio regression model. In the second scenario we assume that only information about the baseline hazard (i.e. control group) is to be borrowed. We refer to this scenario as the shared baseline hazard scenario (SBH). This scenario might arise when the historical study targeted the same disease population that is to be targeted in the new study and when the investigational therapies are not the same. In both the SHR and SBH scenarios we will see that the marginal posterior for the hazard ratio regression parameters is of the same form as that obtained in the IM scenario. Thus, from a software implementation standpoint, a tool that can be used to perform Bayesian design for the IM scenario can address all other design scenarios one might want to consider.

The SHR Scenario

In this scenario we assume the baseline hazard for the historical study is allowed to vary across S_0 levels of a stratification variable. We partition the time axis into $K_{s,0}$ intervals for stratum s. Denote the baseline hazard parameters for the historical study model by $\boldsymbol{\lambda}_0$ and let $\boldsymbol{\theta}_0 = (\boldsymbol{\lambda}_0, \gamma, \boldsymbol{\beta})$ be the complete set of parameters in that model. The power prior is simply the marginal posterior

for the hazard ratio regression parameters based on the exponentiated historical study likelihood multiplied by the initial prior for the new study baseline hazard.

$$\begin{aligned}\pi_0\left(\boldsymbol{\lambda},\gamma,\boldsymbol{\beta}\middle|\boldsymbol{D}_0,\alpha_0\right) &\propto \pi_0\left(\gamma,\boldsymbol{\beta}\middle|\boldsymbol{D}_0,\alpha_0\right)\times\pi_0\left(\boldsymbol{\lambda}\right)\\ &\propto \prod_{s=1}^{S_0}\left\{\frac{\prod_{j\in\mathcal{G}_{s,0}}\phi_j^{\alpha_0\nu_{j,0}}}{\prod_{k=1}^{K_{s,0}}\left[\beta_{sk,0}\right]^{\alpha_{sk,0}}}\right\}\times\pi_0\left(\boldsymbol{\lambda}\right)\end{aligned}$$

It follows that the marginal posterior distribution for the hazard ratio regression parameters after observing observing new data is as follows:

$$\begin{aligned}\pi\left(\gamma,\boldsymbol{\beta}\mid\boldsymbol{D},\boldsymbol{D}_0,\alpha_0\right) &\propto \prod_{s=1}^{S}\left\{\frac{\prod_{i\in\mathcal{G}_s}\phi_i^{\nu_{i,0}}}{\prod_{k=1}^{K_s}\left[\beta_{sk}\right]^{\alpha_{sk}}}\right\}\\ &\times \prod_{s=1}^{S_0}\left\{\frac{\prod_{j\in\mathcal{G}_{s,0}}\phi_j^{\alpha_0\nu_{j,0}}}{\prod_{k=1}^{K_{s,0}}\left[\beta_{sk,0}\right]^{\alpha_{sk,0}}}\right\}.\end{aligned}$$

This result is identical in form to that obtained for the IM scenario when we imagine there being a single stratification variable having $S+S_0$ levels. Thus, we might equivalently consider this an IM scenario with parameter vector $\boldsymbol{\theta}=(\boldsymbol{\lambda},\boldsymbol{\lambda}_0,\gamma,\boldsymbol{\beta})$ where not every parameter is informed by both studies.

The SBH Scenario

In this case we assume that the hazard ratio regression function ϕ only consists of a treatment effect (i.e. $\phi_i=e^{\gamma}$ or $\phi_i=1$ for all i). The historical study dataset can be reduced to the set of subjects that belong to the control group. The marginal posterior for γ is as follows:

$$\pi\left(\gamma\mid\boldsymbol{D},\boldsymbol{D}_0,\alpha_0\right) \propto \prod_{s=1}^{S}\left\{\frac{\prod_{i\in\mathcal{G}_s}\phi_i^{\nu_i}}{\prod_{k=1}^{K_s}\left[\beta_{sk}^*\right]^{\alpha_{sk}^*}}\right\}$$

where $\alpha_{sk}^*=\alpha_{sk}+\alpha_0\sum_{j\in\mathcal{G}_{s,0}}\nu_{jk,0}$ and $\beta_{sk}^*=\beta_{sk}+\alpha_0\sum_{j\in\mathcal{G}_{s,0}}r_{jk,0}$ with $\mathcal{G}_{s,0}$ now representing the set of indices for the historical control subjects from stratum s. As with the SHR scenario, the form of the posterior distribution in the SBH scenario is of the same family as the IM scenario.

One important consideration for the SBH scenario is that by only borrowing information on the

control group, we may create imbalances in risk factors between the treated and control subjects. That is to say, by incorporating historical control subjects that were not randomized alongside new study subjects, the benefit of randomization in the new study is lost and it becomes important to account for known risk factors in the statistical analysis. Thus, choosing risk factors for stratification (or covariate adjustment) is critically important in the SBH scenario.

A.3 A Simulation Study Comparing Inference Based on the PWC-PH model with Inference Based on the Weighted Cox Partial Likelihood

For the simulation studies in this appendix we used the same historical dataset that was used in the example application from Section 6 of the primary text. The number of baseline hazard components and the change points for the analysis model where taken from the best historical study model determined by DIC. The parameter values for the new study model were fixed at the posterior means reported in Table 2 of the primary text. The following characteristics were varied in the stimulation studies:

1. Number of new events $\nu = 100$, 300, and 500 with $n = 3\nu$ enrolled subjects
2. $K_s = 3$, 10, and 50 for both $s = 1$ and $s = 2$
3. $a_0 = 0.0$, 0.5, and 1.0

Enrollment was assumed to be uniform over a 2 year period. Censorship times were simulated from a scaled beta distribution with shape parameters equal to 3 and 1 with support on the interval $(0, 10)$. Administrative censoring occurred at the time when the targeted number of events had been reached. Thus, there was a substantial amount of censoring in the simulated datasets. We performed $B = 1,000$ simulation studies at each of the 27 combinations of the characteristics being varied and computed two summaries of agreement. First, we computed the R^2 value based on regressing the $\log\left[\mathrm{P}\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)\right]$ values computed using the normal approximation to the PWC-PH posterior onto the corresponding quantities computed using the weighted Cox partial likelihood. Second, we computed the percentage of study decisions that were concordant between the two approaches. Study decisions were concordant if $1\left\{\mathrm{P}\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\}$ was equal

Table A.1: Summary of inference agreement using the PWC-PH model and the weighted Cox partial likelihood

		— $K_s = 3$ —		— $K_s = 10$ —		— $K_s = 50$ —	
ν	a_0	R^2	% Concordant	R^2	% Concordant	R^2	% Concordant
100	0.0	0.98907	96.7	0.99725	98.7	0.99855	99.2
100	0.5	0.99060	96.7	0.99435	97.0	0.99474	96.9
100	1.0	0.97206	94.1	0.97506	94.2	0.97519	94.2
300	0.0	0.98663	96.8	0.99767	98.8	0.99957	99.4
300	0.5	0.99493	97.4	0.99871	98.6	0.99928	98.6
300	1.0	0.98399	97.9	0.98939	98.6	0.98924	98.5
500	0.0	0.99289	98.3	0.99932	99.5	0.99985	99.9
500	0.5	0.99214	99.1	0.99775	99.3	0.99887	99.7
500	1.0	0.98792	99.2	0.99461	98.9	0.99500	99.2

for both approaches. Analysis using the weighted Cox partial likelihood were performed using the SAS PHREG Procedure (v9.4). Appendix Table A.1 presents the results of our simulation studies. We see that there is strong agreement in the posterior probabilities used for inference despite the the fact that the assumptions used to establish the connection between the PWC-PH model and the weighted Cox partial likelihood do not hold for the simulation studies. We note a slight decrease in both metrics of agreement as a_0 increases for the case where $\nu = 100$. This is likely attributable to the fact that these analyses include more historical data than new study data and the fact that the historical data is not perfectly consistent with a proportional hazards assumption. It is clear that when the proportional hazards assumption holds definitively ($a_0 = 0$) and the number of baseline hazard components is large, there is near perfect agreement between the methods at all samples sizes considered.

Lastly, since there is strong agreement between inference based on the PWC-PH model and the weighted Cox partial likelihood for both large and small K, one can infer that inference on the hazard ratio regression parameters is robust to the model for the baseline hazard.This suggests that even a fairly coarse partition of the time axis will suffice when one in not concerned with estimating the baseline hazard (such as in the design phase of a trial).

A.4 A Simulation Study Comparing Exact Bayesian Inference Through MCMC with the Laplace Approximation

This appendix is divided into two sections. In the first section we describe our procedure for analysis of the PWC-PH model using MCMC methods. Our approach attempts to make the model fitting step as efficient as possible for scenarios where using MCMC might be preferred in design. Our MCMC implementation also serves as a tool to fit a model to historical data in order to obtain samples from the posterior that can be used to create discrete approximations of the sampling priors. In the second section we provide results from our simulation studies comparing analyses using MCMC to analyses using the proposed normal approximation to the marginal posterior of the hazard ratio parameters.

A.4.1 High Throughput Model Fitting with MCMC

In this section we discuss a reformulation of the marginal posterior for the hazard ratio parameters that is beneficial for MCMC model fitting when the design variables (treatment indicator and covariates) are binary and few in number. This will commonly be the case during design of a clinical trial. In this case, one can write the posterior distribution as a function of a relatively small set of sufficient statistics to which the simulated subject level data can be reduced prior to model fitting.

Let d index the set of D distinct values of the design variables observed in the combined historical study and new study datasets. Write $\phi_d = \exp\left(\gamma z_d^* + \boldsymbol{\beta}^T \boldsymbol{x}_d^*\right)$, where $(z_d^*, \boldsymbol{x}_d^*)$ are the particular values of the design variables identified by d. Let ν_{skd} and r_{skd} be the number of events and total time at risk in interval k for the set of new study subjects in stratum s that had design variable values identified by d. Let $\nu_{skd,0}$ and $r_{skd,0}$ be the analogous quantities for the historical study. The marginal posterior (3.3.4) reduces to

$$\pi\left(\gamma, \boldsymbol{\beta} \mid \boldsymbol{D}, \boldsymbol{D}_0, a_0\right) \quad \propto \quad \frac{\prod_{d=1}^{D} \phi_d^{\nu_d^*}}{\prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[\beta_{sk}^*\right]^{\alpha_{sk}^*}} \tag{A.4.1}$$

where

$$\begin{aligned}
\nu_d^* &= \sum_{s=1}^{S}\sum_{k=1}^{K_s}\left(\nu_{skd} + a_0\nu_{skd,0}\right), \\
\alpha_{sk}^* &= \sum_{d=1}^{D}\left(\nu_{skd} + a_0\nu_{skd,0}\right), \text{ and} \\
\beta_{sk}^* &= \sum_{d=1}^{D}\phi_d\left(r_{skd} + a_0 r_{skd,0}\right).
\end{aligned}$$

Thus, one only needs the set of sufficient statistics $\{(\nu_{skd}, r_{skd}, \nu_{skd,0}, r_{skd,0}) : \forall s, k, d\}$ and $\{(z_d^*, \boldsymbol{x}_d^*) : \forall d\}$ in order to evaluate the marginal posterior for MCMC sampling. When $D \ll n + n_0$, an approach using the sufficient statistics is far superior to the straight forward approach using subject level data. In our implementation we first reduce each simulated dataset to the sufficient statistics and the perform MCMC model fitting using (A.4.1). Our implementation supports an arbitrary number of binary covariates.

When the covariates are restricted to be binary it is straight forward to show that the full conditional distribution for each of the hazard ratio parameters is log-concave. Thus, one can use rejection sampling or adaptive rejection sampling (W. R. Gilks, 1992) to sample from the full conditionals. In our implementation we utilize a simple Newton-Raphson algorithm to locate the mode of the full conditional and then construct a three-part envelope function centered about the mode for rejection sampling. We use the optimal envelope under the assumption of normality for the full conditional. Since we only need to draw one sample at each Gibbs step and since so few samples actually get rejected, we find that it is not necessary to adapt the envelope.

A.4.2 A Comparison of MCMC Analysis Results with Results based on the Laplace Approximation

For the simulation studies in this section we used the same historical data set as was used in the example application in Section 3.5. The number of baseline hazard components and the change points for the new study model where taken from the best historical study model determined by DIC. The parameter values for the new study model were fixed at the posterior means reported in

Table A.2: Summary of inference agreement using MCMC and the Laplace approximation

		— $K = 3$ —		— $K = 10$ —		— $K = 25$ —	
ν	a_0	R^2	% Concordant	R^2	% Concordant	R^2	% Concordant
40	0.0	0.9996	99.4	0.9995	99.1	0.9995	99.2
40	0.5	0.9999	98.4	0.9999	98.6	0.9999	99.6
40	1.0	0.9999	98.5	0.9998	99.2	0.9998	99.4
80	0.0	0.9999	99.3	0.9998	99.3	0.9999	99.1
80	0.5	0.9999	99.5	0.9999	99.4	0.9999	98.9
80	1.0	0.9999	99.2	0.9999	99.4	0.9999	99.2
120	0.0	0.9999	99.5	0.9999	99.7	0.9999	99.3
120	0.5	0.9999	99.3	0.9999	99.6	0.9999	99.1
120	1.0	0.9999	99.3	0.9999	99.6	0.9999	99.4

Table 3.2. The following parameters were varied in the stimulation studies:

1. Number of new events $\nu = 40$, 80, and 120 with $n = 3\nu$ enrolled subjects
2. $K_s = 3$, 10, and 25 for both $s = 1$ and $s = 2$
3. $a_0 = 0.0$, 0.5, and 1.0

Enrollment was assumed to be uniform over a 2 year period. Administrative censoring occurred at the time when the targeted number of events had been reached. No other censoring mechanism was simulated. We performed $B = 1,000$ simulation studies at each of the 27 combinations of the parameters being varied and computed two summaries of agreement. First, we computed the R^2 value based on regressing the log $\left[\mathrm{P}\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)\right]$ values computed using the normal approximation onto the corresponding quantities computed using MCMC (based on 100,000 MCMC samples). Second, we computed the percentage of study decisions that were concordant between the two approaches. Study decisions were concordant if $1\left\{\mathrm{P}\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\}$ was equal for both approaches. Appendix Table A.2 presents the results of our simulation studies. For $a_0 = 0$, the analysis does not use any of the historical data and so the case where $\nu = 40$ and $a_0 = 0$ gives good insight into small sample performance. It appears the normal approximation is valid even when the sample size is much smaller than what will be encountered in an adequately powered clinical trial. For all cases the R^2 value was near 1.0 and the concordance between the methods was never below 98%.

APPENDIX B: CHAPTER 4 SUPPLEMENTAL MATERIALS

B.1 A Simulation Study Comparing Bayesian Inference using MCMC with the Weighted Maximum Likelihood Approximation

For the simulation studies in this appendix, we used the same historical data set that was used in the example application from Section 4.5 of the paper. Inference using MCMC methods was based on 50,000 samples with a 1,000 sample tuning phase for the slice sampler and a post-tuning burn-in phase of 100 samples. A uniform improper prior was used for the regression parameters and a Gamma $(0.001, 0.001)$ prior was used for each of the parameters in the promotion time model. Note that in the weighted maximum likelihood (ML) approximation, the prior is not incorporated. The following characteristics were varied in the simulation studies: sampling prior (default null and default alternative), a_0 (0.0 and 1.0), and n (550 and 800). All other characteristics of the simulations (e.g. enrollment distribution) matched that described in the paper. We performed $B = 1,000$ simulation studies at each of the 8 combinations of the characteristics being varied and computed two summaries of agreement. First, we computed the R^2 value based on regressing the $\mathrm{P}\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right)$ values computed using the weighted maximum likelihood approximation onto the corresponding quantities computed using MCMC. Second, we computed the percentage of study decisions that were concordant between the two approaches. Study decisions were concordant if $1\left\{\mathrm{P}\left(\gamma < 0 \middle| \boldsymbol{D}^{(b)}, \boldsymbol{D}_0, a_0\right) \geq \psi\right\}$ was equal for both approaches. Appendix Table B.1 presents the results of our simulation studies. We see that there is near exact agreement between the posterior probabilities as well as the simulated outcomes. This suggests that the weighted maximum

Table B.1: Summary of inference agreement using MCMC and the weighted maximum likelihood approximation

		– Default Alternative –		– Default Null –	
n	a_0	% Concordant	R^2	% Concordant	R^2
550	0.0	99.5	0.999894	99.9	0.999884
550	1.0	99.6	0.999721	99.9	0.999840
800	0.0	99.4	0.999929	99.9	0.999902
800	1.0	99.8	0.999830	99.9	0.999879

likelihood approximation is justified for the sample sizes considered in this paper.

APPENDIX C: CHAPTER 5 SUPPLEMENTAL MATERIALS

C.1 Integral Computation for the Restricted Maximal Borrowing Power Prior

In this appendix we develop a highly accurate approximation for the integral in (5.4.4) which is given by

$$\int \left[\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{a_0}\left[\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{\delta}\pi_0(\boldsymbol{\beta},\boldsymbol{\lambda})d\left(\boldsymbol{\beta},\boldsymbol{\lambda}\right)$$

where $\left[\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{\delta}$ is the likelihood for the new CVOT under the assumption of perfect homogeneity with the historical CVOT. For this derivation we take $\pi_0(\boldsymbol{\beta},\boldsymbol{\lambda}) = \pi_0(\boldsymbol{\beta})\times\pi_0(\boldsymbol{\lambda})$ with $\pi_0(\boldsymbol{\beta}) \propto 1$ and $\pi_0(\boldsymbol{\lambda}) \propto \prod_{s=1}^{S}\prod_{k=1}^{K_s}\lambda_{sk}^{-1}$. Using notation similar to (3.2.2), the integrand can be written as follows:

$$\left[\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{a_0+\delta}\pi_0(\boldsymbol{\beta},\boldsymbol{\lambda}) \propto \prod_{s=1}^{S}\prod_{k=1}^{K_s}\lambda_{sk}^{(a_0+\delta)\alpha_{sk,0}-1}e^{-(a_0+\delta)\beta_{sk,0}\lambda_{sk}} \times \left(\prod_{i=1}^{n_0}\phi_i^{\nu_{i0}}\right)^{(a_0+\delta)} \tag{C.1.1}$$

where $\alpha_{sk,0} = \sum_{i\in\mathcal{G}_s}\nu_{i0}^{k}$ is the total number of events for historical CVOT subjects from stratum s occurring in interval $I_{s,k}$, $\phi_i = \exp\left(\gamma z_i + \boldsymbol{\beta}^T\boldsymbol{x}_i\right)$ is the hazard ratio regression function for historical CVOT subject i, $\beta_{sk,0} = \sum_{i\in\mathcal{G}_s}\phi_i r_{i0}^{k}$, and $\boldsymbol{D}_0 = \{(y_{i0},\nu_{i0},z_i,\boldsymbol{x}_i) : i = 1,...,n\}$ is the observed data from the historical CVOT. It is straight forward to integrate out the baseline hazard in (C.1.1) and doing so yields the expression

$$\prod_{s=1}^{S}\prod_{k=1}^{K_s}\frac{\Gamma\left(\left(a_0+\delta\right)\alpha_{sk,0}\right)}{\left(a_0+\delta\right)^{(a_0+\delta)\alpha_{sk,0}}} \times f\left(\boldsymbol{\beta}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right) \tag{C.1.2}$$

where

$$f\left(\boldsymbol{\beta}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right) = \left(\frac{\prod_{i=1}^{n_0}\phi_i^{\nu_{i0}}}{\prod_{s=1}^{S}\prod_{k=1}^{K_s}\beta_{sk,0}^{\alpha_{sk,0}}}\right)^{a_0+\delta}.$$

When the number of combined events in the CVOTs are not very small, the term being exponentiated in $f\left(\boldsymbol{\beta}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right)$ is proportional to a multivariate normal density with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix of $\hat{\Sigma}$ that only depend on $\boldsymbol{D}_0$. It follows that $f\left(\boldsymbol{\beta}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right)$ is proportional to

a multivariate normal density with the same mean and covariance matrix $\frac{\hat{\Sigma}}{a_0+\delta}$. Thus,

$$p\left(\boldsymbol{\beta}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right) = \frac{f\left(\boldsymbol{\beta}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right)}{f\left(\hat{\boldsymbol{\beta}}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right)}\left|\hat{\Sigma}\right|^{-1/2}(a_0+\delta)^{p/2}(2\pi)^{-p/2}$$

is approximately a multivariate normal density. Based on this, one can easily integrate $\boldsymbol{\beta}$ out of (C.1.2) to obtain the following expression which discards any term that does not depend on a_0.

$$\int\left[\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{a_0}\left[\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\boldsymbol{D}_0)\right]^{\delta}\pi_0(\boldsymbol{\beta},\boldsymbol{\lambda})d\left(\boldsymbol{\beta},\boldsymbol{\lambda}\right)\propto\prod_{s=1}^{S}\prod_{k=1}^{K_s}\frac{\Gamma\left(\left(a_0+\delta\right)\alpha_{sk,0}\right)}{\left(a_0+\delta\right)^{(a_0+\delta)\alpha_{sk,0}}}\frac{f\left(\hat{\boldsymbol{\beta}}\middle|a_0,\boldsymbol{D},\boldsymbol{D}_0\right)}{\left(a_0+\delta\right)^{p/2}}$$

BIBLIOGRAPHY

Adcock, C. (1997), "Sample size determination: a review," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 261–283.

Baron, U., Turbachova, I., Hellwag, A., Eckhardt, F., Berlin, K., Hoffmüller, U., Gardina, P., and Olek, S. (2006), "DNA methylation analysis as a tool for cell typing," *Epigenetics*, 1, 56–61.

Berger, J. and Berliner, L. M. (1986), "Robust Bayes and empirical Bayes analysis with ε-contaminated priors," *The Annals of Statistics*, 14, 461–486.

Berkson, J. and Gage, R. P. (1952), "Survival curve for cancer patients following treatment," *Journal of the American Statistical Association*, 47, 501–515.

Bernardo, P. and Ibrahim, J. G. (2000), "Group sequential designs for cure rate models with early stopping in favour of the null hypothesis," *Statistics in medicine*, 19, 3023–3035.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010), "The NIH roadmap epigenomics mapping consortium," *Nature Biotechnology*, 28, 1045–1048.

Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010), *Bayesian adaptive methods for clinical trials*, CRC press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent dirichlet allocation," *Journal of machine Learning research*, 3, 993–1022.

Box, G. E. (1980), "Sampling and Bayes' inference in scientific modelling and robustness," *Journal of the Royal Statistical Society. Series A (General)*, 143, 383–430.

Breslow, N. (1974), "Covariance analysis of censored survival data," *Biometrics*, 30, 89–99.

Brown, B. W., Herson, J., Atkinson, E. N., and Rozell, M. E. (1987), "Projection from previous studies: a Bayesian and frequentist compromise," *Controlled clinical trials*, 8, 29–44.

Brutti, P., De Santis, F., and Gubbiotti, S. (2008), "Robust Bayesian sample size determination in clinical trials," *Statistics in Medicine*, 27, 2290–2306.

Campbell, G. (2011), "Bayesian Statistics in Medical Devices: Innovation Sparked by the FDA." *Journal of Biopharmaceutical Statistics*, 21, 871–887.

Charlton, J., Williams, R. D., Weeks, M., Sebire, N. J., Popov, S., Vujanic, G., Mifsud, W., Alcaide-German, M., Butcher, L. M., Beck, S., et al. (2014), "Methylome analysis identifies a Wilms tumor epigenetic biomarker detectable in blood," *Genome Biology*, 15, 434.

Chen, M.-H., Harrington, D. P., and Ibrahim, J. G. (2002a), "Bayesian cure rate models for malignant melanoma: a case-study of Eastern Cooperative Oncology Group trial E1690," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51, 135–150.

Chen, M.-H., Ibrahim, J. G., Amy Xia, H., Liu, T., and Hennessey, V. (2014a), "Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program," *Statistics in medicine*, 33, 1600–1618.

Chen, M.-H., Ibrahim, J. G., Lam, P., Yu, A., and Zhang, Y. (2011), "Bayesian design of noninferiority trials for medical devices using historical data," *Biometrics*, 67, 1163–1170.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999), "A new Bayesian model for survival data with a surviving fraction," *Journal of the American Statistical Association*, 94, 909–919.

— (2002b), "Bayesian inference for multivariate survival data with a cure fraction," *Journal of Multivariate Analysis*, 80, 101–126.

Chen, M.-H., Ibrahim, J. G., Zeng, D., Hu, K., and Jia, C. (2014b), "Bayesian design of superiority clinical trials for recurrent events data with applications to bleeding and transfusion events in myelodyplastic syndrome," *Biometrics*, 70, 1003–1013.

Clarke, B. and Yuan, A. (2006), "Closed form expressions for Bayesian sample size," *The Annals of Statistics*, 34, 1293–1330.

Cox, D. R. (1975), "Partial likelihood," *Biometrika*, 62, 269–276.

De Santis, F. (2004), "Statistical evidence and sample size determination for Bayesian hypothesis testing," *Journal of Statistical Planning and Inference*, 124, 121–144.

— (2006), "Sample size determination for robust Bayesian analysis," *Journal of the American Statistical Association*, 101, 278–291.

— (2007), "Using historical data for Bayesian sample size determination," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 95–113.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, 39, 1–38.

Dolzhenko, E. and Smith, A. D. (2014), "Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments," *BMC Bioinformatics*, 15, 215.

Duan, Y. (2005), "A modified bayesian power prior approach with applications in water quality evaluation," Ph.D. thesis, Virginia Polytechnic Institute and State University.

Duan, Y., Smith, E. P., and Ye, K. (2006), "Using power priors to improve the binomial test of water quality," *Journal of agricultural, biological, and environmental statistics*, 11, 151–168.

Dudley, R. M. and Haughton, D. (2002), "Asymptotic normality with small relative errors of posterior probabilities of half-spaces," *The Annals of Statistics*, 30, 1311–1344.

Evans, M., Moshonov, H., et al. (2006), "Checking for prior-data conflict," *Bayesian Analysis*, 1, 893–914.

Feng, H., Conneely, K. N., and Wu, H. (2014), "A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data," *Nucleic Acids Research*, 42, e69.

Food and Drug Administration (2008), "Guidance for Industry: Diabetes Mellitus – Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes," [Online; last accessed 08-August-2016].

— (2010), "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials," [Online; last accessed 27-January-2016].

— (2016), "Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices," [Online; last accessed 11-July-2016].

Gagnon-Bartsch, J. A. and Speed, T. P. (2012), "Using control genes to correct for unwanted variation in microarray data," *Biostatistics*, 13, 539–552.

Geiger, M. J., Mehta, C., Turner, J. R., Arbet-Engels, C., Hantel, S., Hirshberg, . B., Koglin, J., Mendzelevski, B., Sager, P. T., Shapiro, D., et al. (2015), "Clinical development approaches and statistical methodologies to prospectively assess the cardiovascular risk of new antidiabetic therapies for type 2 diabetes," *Therapeutic Innovation & Regulatory Science*, 49, 50–64.

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and Szustakowski, J. D. (2011), "Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples," *PLoS One*, 6, e27156.

Green, J. B., Bethel, M. A., Armstrong, P. W., Buse, J. B., Engel, S. S., Garg, J., Josse, R., Kaufman, K. D., Koglin, J., Korn, S., et al. (2015), "Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes," *New England Journal of Medicine*, 373, 232–242.

Green, J. B., Bethel, M. A., Paul, S. K., Ring, A., Kaufman, K. D., Shapiro, D. R., Califf, R. M., and Holman, R. R. (2013), "Rationale, design, and organization of a randomized, controlled Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS) in patients with type 2 diabetes and established cardiovascular disease," *American Heart Journal*, 166, 983–989.

Guintivano, J., Aryee, M. J., and Kaminsky, Z. A. (2013), "A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression," *Epigenetics*, 8, 290–302.

Hebestreit, K., Dugas, M., and Klein, H.-U. (2013), "Detection of significantly differentially methylated regions in targeted bisulfite sequencing data," *Bioinformatics*, 29, 1647–1653.

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

Hobbs, B. P., Sargent, D. J., and Carlin, B. P. (2012), "Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models," *Bayesian Analysis*, 7, 639.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., et al. (2011), "Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment," *Molecular Cell*, 44, 17–28.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012), "DNA methylation arrays as surrogate measures of cell mixture distribution," *BMC Bioinformatics*, 13, 86.

Houseman, E. A., Kelsey, K. T., Wiencke, J. K., and Marsit, C. J. (2015), "Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective," *BMC Bioinformatics*, 16, 95.

Houseman, E. A., Molitor, J., and Marsit, C. J. (2014), "Reference-free cell mixture adjustments in analysis of DNA methylation data," *Bioinformatics*, 30, 1431–1439.

Ibrahim, J. G. and Chen, M.-H. (1998), "Prior distributions and Bayesian computation for proportional hazards models," *Sankhyā: The Indian Journal of Statistics, Series B*, 60, 48–64.

— (2000), "Power prior distributions for regression models," *Statistical Science*, 46–60.

Ibrahim, J. G., Chen, M.-H., and Chu, H. (2012a), "Bayesian methods in clinical trials: a Bayesian analysis of ECOG trials E1684 and E1690," *BMC Medical Research Methodology*, 12, 183.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015), "The power prior: theory and applications," *Statistics in medicine*, 34, 3724–3749.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001a), "Bayesian semiparametric models for survival data with a cure fraction," *Biometrics*, 57, 383–388.

— (2001b), *Bayesian Survival Analysis*, Springer Science & Business Media.

Ibrahim, J. G., Chen, M.-H., Xia, H. A., and Liu, T. (2012b), "Bayesian Meta-Experimental Design: Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes," *Biometrics*, 68, 578–586.

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.

Inoue, L. Y., Berry, D. A., and Parmigiani, G. (2005), "Relationship between Bayesian and frequentist sample size determination," *The American Statistician*, 59, 79–87.

Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddeloh, J. A., Wen, B., and Feinberg, A. P. (2008), "Comprehensive high-throughput arrays for relative methylation (CHARM)," *Genome Research*, 18, 780–790.

Jaffe, A. E. and Irizarry, R. A. (2014), "Accounting for cellular heterogeneity is critical in epigenome-wide association studies," *Genome Biology*, 15, R31.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186, 453–461.

Jones, P. A. and Takai, D. (2001), "The role of DNA methylation in mammalian epigenetics," *Science*, 293, 1068–1070.

Joseph, L., Wolfson, D. B., and Du Berger, R. (1995), "Sample size calculations for binomial proportions via highest posterior density intervals," *The Statistician*, 44, 143–154.

Kalbfleisch, J. D. (1978), "Non-Parametric Bayesian Analysis of Survival Time Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 40, 214–221.

Katsis, A. and Toman, B. (1999), "Bayesian sample size calculations for binomial experiments," *Journal of Statistical Planning and Inference*, 81, 349–362.

Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M., and Blum, R. H. (2000), "High-and low-dose interferon alfa-2b in high-risk melanoma: first analysis of intergroup trial E1690/S9111/C9190," *Journal of clinical oncology*, 18, 2444–2458.

Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C., and Blum, R. H. (1996), "Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684." *Journal of clinical oncology*, 14, 7–17.

Koch, G. (2015), "Comment," *Statistics in Biopharmaceutical Research*, 7, 267–271.

Krueger, F. and Andrews, S. R. (2011), "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications," *Bioinformatics*, 27, 1571–1572.

Kulis, M. and Esteller, M. (2010), "DNA methylation and cancer," *Advances in Genetics*, 70, 27–56.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013), "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis," *Nature Biotechnology*, 31, 142–147.

Lo, Y. D., Tein, M. S., Pang, C. C., Yeung, C. K., Tong, K.-L., and Hjelm, N. M. (1998), "Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients," *The Lancet*, 351, 1329–1330.

Lun, F. M., Chiu, R. W., Sun, K., Leung, T. Y., Jiang, P., Chan, K. A., Sun, H., and Lo, Y. D. (2013), "Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA," *Clinical Chemistry*, 59, 1583–1594.

Marchenko, O., Jiang, Q., Chakravarty, A., Ke, C., Ma, H., Maca, J., Russek-Cohen, E., Sanchez-Kam, M., C. Zink, R., and Chuang-Stein, C. (2015), "Evaluation and Review of Strategies to Assess Cardiovascular Risk in Clinical Trials in Patients with Type 2 Diabetes Mellitus," *Statistics in Biopharmaceutical Research*, 7, 253–266.

M'Lan, C. E., Joseph, L., Wolfson, D. B., et al. (2008), "Bayesian sample size determination for binomial proportions," *Bayesian Analysis*, 3, 269–296.

Montaño, C. M., Irizarry, R. A., Kaufmann, W. E., Talbot, K., Gur, R. E., Feinberg, A. P., and Taub, M. A. (2013), "Measuring cell-type specific differential methylation in human brain tissue," *Genome Biology*, 14, 1–9.

Neal, R. M. (2003), "Slice sampling," *The Annals of Statistics*, 31, 705–767.

Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012), "Cure models as a useful statistical tool for analyzing survival," *Clinical Cancer Research*, 18, 3731–3736.

Pennello, G. and Thompson, L. (2007), "Experience with reviewing Bayesian medical device trials," *Journal of Biopharmaceutical Statistics*, 18, 81–115.

Pezeshk, H. (2003), "Bayesian techniques for sample size determination in clinical trials: a short review," *Statistical Methods in Medical Research*, 12, 489–504.

Pham-Gia, T. and Turkkan, N. (2003), "Determination of exact sample sizes in the Bayesian estimation of the difference of two proportions," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52, 131–150.

Rahme, E. and Joseph, L. (1998), "Exact sample size determination for binomial experiments," *Journal of Statistical Planning and Inference*, 66, 83–93.

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D., Söderhäll, C., Scheynius, A., and Kere, J. (2012), "Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility," *PLoS One*, 7, e41361.

Robertson, K. D. (2005), "DNA methylation and human disease," *Nature Reviews Genetics*, 6, 597–610.

Rondeau, V. (2010), "Statistical models for recurrent events and death: Application to cancer events," *Mathematical and Computer Modelling*, 52, 949–955.

Rubin, D. B. and Stern, H. S. (1998), "Sample size determination using posterior predictive distributions," *Sankhyā: The Indian Journal of Statistics, Series B*, 60, 161–175.

Rubin, D. B. et al. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics*, 12, 1151–1172.

Scirica, B. M., Bhatt, D. L., Braunwald, E., Steg, P. G., Davidson, J., Hirshberg, B., Ohman, P., Frederich, R., Wiviott, S. D., Hoffman, E. B., Cavender, M. A., Udell, J. A., Desai, N. R., Mosenzon, O., McGuire, D. K., Ray, K. K., Leiter, L. A., and Raz, I. (2013), "Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus," *New England Journal of Medicine*, 369, 1317–1326.

Scirica, B. M., Bhatt, D. L., Braunwald, E., Steg, P. G., Davidson, J., Hirshberg, B., Ohman, P., Price, D. L., Chen, R., Udell, J., et al. (2011), "The design and rationale of the Saxagliptin Assessment of Vascular Outcomes Recorded in patients with diabetes mellitus–Thrombolysis in Myocardial Infarction (SAVOR-TIMI) 53 Study," *American Heart Journal*, 162, 818–825.

Simon, R. (1999), "Bayesian design and analysis of active control clinical trials," *Biometrics*, 55, 484–487.

Sinha, D., Ibrahim, J. G., and Chen, M.-H. (2003), "A Bayesian justification of Cox's partial likelihood," *Biometrika*, 90, 629–641.

Spiegelhalter, D. and Freedman, L. (1986), "A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion," *Statistics in medicine*, 5, 1–13.

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

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

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014), "MOABS: model based analysis of bisulfite sequencing data," *Genome biology*, 15, 1–12.

Sun, K., Jiang, P., Chan, K. A., Wong, J., Cheng, Y. K., Liang, R. H., Chan, W.-k., Ma, E. S., Chan, S. L., Cheng, S. H., et al. (2015), "Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments," *Proceedings of the National Academy of Sciences*, 112, 503–512.

The ACCORD Study Group (2011), "Long-Term Effects of Intensive Glucose Lowering on Cardiovascular Outcomes," *New England Journal of Medicine*, 364, 818–828.

The ADVANCE Collaborative Group (2008), "Intensive Blood Glucose Control and Vascular Outcomes in Patients with Type 2 Diabetes," *New England Journal of Medicine*, 358, 2560–2572.

Tsodikov, A. (2002), "Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage," *Statistics in Medicine*, 21, 895–920.

Tsodikov, A., Ibrahim, J., and Yakovlev, A. (2003), "Estimating cure rates from survival data," *Journal of the American Statistical Association*, 98, 1063–1078.

Tsodikov, A., Loeffler, M., and Yakovlev, A. (1998), "A cure model with time-changing risk factor: an application to the analysis of secondary leukaemia. A report from the International Database on Hodgkin's Disease," *Statistics in Medicine*, 17, 27–40.

Van den Meersche, K., Soetaert, K., and Van Oevelen, D. (2009), "xsample (): an R function for sampling linear inverse problems," *Journal of Statistical Software*, 30, 1–15.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. A., Stamatoyannopoulos, J. A., Crawford, G. E., et al. (2013), "Dynamic DNA methylation across diverse human cell lines and tissues," *Genome Research*, 23, 555–567.

W. R. Gilks, P. W. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41, 337–348.

Wang, F. and Gelfand, A. E. (2002), "A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models," *Statistical Science*, 17, 193–208.

Weiss, R. (1997), "Bayesian sample size calculations for hypothesis testing," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 185–191.

White, W. B., Bakris, G. L., Bergenstal, R. M., Cannon, C. P., Cushman, W. C., Fleck, P., Heller, S., Mehta, C., Nissen, S. E., Perez, A., et al. (2011), "EXamination of cArdiovascular outcoMes with alogliptIN versus standard of carE in patients with type 2 diabetes mellitus and acute coronary syndrome (EXAMINE): a cardiovascular safety study of the dipeptidyl peptidase 4 inhibitor alogliptin in patients with type 2 diabetes with acute coronary syndrome," *American Heart Journal*, 162, 620–626.

White, W. B., Cannon, C. P., Heller, S. R., Nissen, S. E., Bergenstal, R. M., Bakris, G. L., Perez, A. T., Fleck, P. R., Mehta, C. R., Kupfer, S., et al. (2013), "Alogliptin after acute coronary syndrome in patients with type 2 diabetes," *New England Journal of Medicine*, 369, 1327–1335.

Woods, L., Rachet, B., Lambert, P., and Coleman, M. (2009), ""Cure" from breast cancer among two populations of women followed for 23 years after diagnosis," *Annals of Oncology*, 20, 1331–1336.

Yakovlev, A. Y. (1996), "Threshold models of tumor recurrence," *Mathematical and Computer Modelling*, 23, 153–164.

Yakovlev, A. Y., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefediere, A., and Tsodikov, A. (1993), "A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer," *Biometrie et analyse de donnees spatio-temporelles*, 12, 66–82.

Zaider, M., Zelefsky, M. J., Hanin, L. G., Tsodikov, A. D., Yakovlev, A. Y., and Leibel, S. A. (2001), "A survival model for fractionated radiotherapy with an application to prostate cancer," *Physics in Medicine and Biology*, 46, 2745–2758.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., et al. (2013), "Charting a dynamic DNA methylation landscape of the human genome," *Nature*, 500, 477–481.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014), "Epigenome-wide association studies without the need for cell-type composition," *Nature Methods*, 11, 309–311.