Statistical Methods in Surrogate Marker Research for Clinical Trials

by

Yun Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2008

Doctoral Committee:

Professor Jeremy M.G. Taylor, Chair
Professor Roderick J.A. Little
Assistant Professor Michael R. Elliott
Associate Research Scientist Ananda Sen

ACKNOWLEDGEMENTS

First of all, I would like to thank Dr. Jeremy Taylor for his tremendous support, help and patience throughout my dissertation research. Jeremy has always been available to discuss my research, give me comments and meticulously edit my work. The experience I have gained as a graduate research assistant under Jeremy's supervision has been invaluable. I am very thankful for having Jeremy as my role model for both statistical methodology and applications research. I also very much appreciate his financial support during my entire PhD study. I would also like to thank Dr. Mike Elliott for his valuable input on the fourth chapter of my dissertation and his thoughtful career advice. I appreciate Michael Passarelli for his work in reconstructing the colorectal cancer data used in the fourth chapter. I am very grateful to Dr. Rod Little for initiating the ridge regression idea for the third chapter, his many insightful comments, his encouragement and his willingness to help. I want to thank Dr. Brenda Gillespie for providing us with the CIGTS data. I appreciate the great opportunity and experience working with both Drs. Ananda Sen and Mousumi Banerjee on an interesting competing risk problem. I would also like to thank many of my fellow students and faculty in the Biostatistics department for their helpful discussion and feedback, particularly, Laila Poisson, Ronglai Shen, Hyungwon Choi, Yun Li, Sinae Kim, Bhramar Mukherjee and Rebecca Andridge. I am also deeply indebted to my previous mentors, particularly Drs. Larry Kupper of UNC-Chapel Hill, Kevin Weinfurt and Kevin Anstrom of Duke. Personally, I feel enormously

thankful to my husband, Doug Schaubel, for his support, encouragement and unyielding confidence in me. Last, but definitely not the least, I would like to thank my parents, brother, Shirley, Earl and friends for their constant moral support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

.

ABSTRACT

Statistical Methods in Surrogate Marker Research for Clinical Trials

by
Yun Li

Chair: Jeremy M.G. Taylor

A surrogate marker (S) is often an intermediate physical or laboratory indicator in a disease progression process. It can be measured earlier and cost less than the true endpoint (T). A surrogate marker may be able to facilitate early prediction of the treatment (Z) effect on T and thus can be very useful in reducing the duration and cost of a clinical trial. In practice, it can either serve as a substitute for T or as an auxiliary variable. One part of my dissertation focuses on its role as an auxiliary variable. We aim to directly investigate its usage in predicting the treatment effect and identify the situations when S can be beneficial in improving the precision in both single- and multiple-trial settings when T is not completely observed. When the individual-level correlation is relatively high, there is substantial efficiency gain by using S, particularly in a multiple-trial setting. We also study the extent of efficiency gain with respect to different model assumptions that are used to describe the relationship among S, T and Z. The results motivate a generalized ridge regression method which strikes a balance between bias reduction and efficiency gain without the need to specify correct models. The other part of the dissertation directly models

the relationship of T, S and Z in a causal framework. Previous work on surrogate markers often requires one to fit models for the distribution of T given S and Z. It is well known that it usually does not have a causal interpretation because the models condition on a post randomization variable S. To solve this problem, we adapt a causal framework using the principal stratification approach introduced by Frangakis and Rubin (2002). We propose a Bayesian method to estimate the causal associations between the potential outcomes of S and T. To not only overcome some non-identifiability problems but also improve the precision of the statistical inference, we incorporate assumptions that are plausible in the surrogate context into prior distributions. The method is explored in both single trial and multiple trial settings.

CHAPTER I

Introduction

Surrogate markers (S) are often intermediate physical or laboratory indicators in a disease progression process that can be measured earlier and often easier to collect than the true endpoint (T). Examples of surrogate markers include CD4 counts and viral load for HIV infection, blood pressure and serum cholesterol level for cardiovascular disease and prostate-specific antigen for prostate cancer. A good surrogate marker has enormous potential benefits if it can reliably facilitate early prediction of the treatment effect in a clinical trial. When the true endpoints are rare, late-occurring or costly to obtain, the use of good surrogate markers can substantially reduce the trial duration and size, lower the expense and lead to earlier decision making. As more and more biomarkers are being discovered and many of which are suggested as surrogate markers, there have been continuous interest in surrogate markers in the clinical research community, particularly in pharmaceutical companies.

Generally, the research on surrogate markers has focused on their two roles. First, a surrogate marker could serve as a surrogate endpoint and be used to replace the true endpoint for early treatment efficacy evaluation. Second, it could serve as an auxiliary outcome to potentially improve the efficiency of the treatment effect estimate on T

when we observe S on more patients than T. Most previous research has been devoted to developing surrogacy validation measures to quantify how well S can replace T. In a landmark paper, Prentice (1989) proposed a formal definition for perfect surrogacy and provided validation criteria. To measure less than perfect surrogacy, the proportion of the treatment effect explained by surrogate markers was proposed by Freedman *et al* (1992) and further studied and extended by several other authors (Lin *et al*, 1997; Bycott and Taylor, 1998; Wang and Taylor, 2002). Two major drawbacks have emerged in the current literature. First, although these surrogacy measures are useful to understand to what degree the effect of Z on S accounts for the effect of Z on T, they are not practically useful for predicting the effect of treatment on T even when a substitute for T is found. Second, the quest for a substitute for T often fails. As a result, most existing reports on the use of surrogate markers have been discouraging.

In this dissertation, the objective in both Chapter II and Chapter III is to explore the role of a surrogate marker as an auxiliary variable in improving the precision of the early treatment effect prediction. In Chapter II, we aim to directly identify the situations when S can be beneficial in increasing the precision in the setting where T is not completely observed. We examine factors that impact the amount of efficiency gain, such as the trial-level correlation, individual-level correlation (defined by Buyse *et al* (2000)) and the fraction of missing T. We focus on S and T being continuous in both single trial and multiple trial settings. While the trial-level correlation is identified as the key factor that impacts the degree of efficiency gain from S in the research by Buyse *et al.* (2000) and Gail *et al* (2000), we find that the individual-level correlation plays an even more important role than the trial-level correlation in obtaining substantial efficiency gain from S with respect to the estimated treatment effect on

T when T is partially observed. In Chapter III, we examine the extent of precision improvement through the use of S with respect to different model assumptions that are used to describe the relationship among T, S and Z. When S satisfies Prentice's definition for perfect surrogacy, there is substantial gain in precision by using S to estimate the treatment effect. When S is not close to having perfect surrogacy, it can provide substantial information only under special circumstances. We propose a generalized ridge regression to avoid the need to make a correct surrogacy assumption. As simulations will show, it can strike a balance between bias and efficiency gain depending on the surrogacy nature.

Both surrogacy measures and the use of S in predicting the effect of Z on T often require one to fit models for the distribution of T given S and Z. It is well known that these measures and predicted treatment effects often do not have causal interpretations because the models used condition on a post randomization variable S (Rosenbaum, 1984). An alternative approach is to directly model the relationship between S and T in a causal framework. Frangakis and Rubin (2002) suggested a framework to study surrogacy through the association between potential outcomes of T and potential outcomes of S. In contrast to the Prentice (1989) and Freedman (1992) criteria, these association measures and quantities derived always have causal interpretations.

While Frangakis and Rubin (2002) have laid out a causal framework, no methods are currently available for estimation. In Chapter IV, we propose a Bayesian estimation method to evaluate the causal probabilities associated with the combinations of different sequence of potential outcomes for S and T for each individual when S and T are both binary. To overcome non-identifiability and increase the precision of the statistical inference, we incorporate assumptions that are plausible in the sur-

rogate context into prior distributions. We also explore the relationship among the surrogacy measures based on the traditional models and the counterfactual models. We use the causal probabilities to predict the treatment effect when T is partially observed. We then extend the method to the multiple trial setting using hierarchical modeling.

1.1 References

Buyse M., Molenberghs G., Burzykowski T., Renard D. and Geys H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. **1**, 49-67.

Bycott P.W., Taylor J.M.G. (1998). An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. *Controlled Clinical Trials*. **19**: 555-568.

Frangakis, C.E. and Rubin, D.B. (2002). Principal stratification in casual inference. *Biometrics*, **58**, 21–29.

Freedman L.S., Graubard B.I., Schatzkin A (1992). Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine*. **11**, 167-178.

Lin D.Y., Fleming T.R., DeGruttola V. (1997). Estimating the proportion of treatment effect captured by a surrogate marker. *Statistics in Medicine*. **16**, 1515-1527.

Prentice R.L. (1989). Surrogate endpoints in clinical trials, definition and operational criteria. *Statistics in Medicine*. **8**, 431-440.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *The Journal of the Royal Statistical Society, Series A*, **147**, 656–666.

Wang Y., Taylor J.M.G. (2003). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*. **58**, 803-812.

CHAPTER II

Predicting Treatment Effects Using Surrogate Markers in Clinical Trials

Summary. A surrogate marker (S) is a variable that is measured after treatment in a randomized clinical trial. It is typically easier to measure than the true endpoint and may be useful to help to shorten the length of the trial or reduce its costs. A potential use of a surrogate marker is to completely replace the true endpoint and evaluate whether the treatment is effective. A second potential use of a surrogate marker is to help to predict the treatment effect on the true endpoint (T) in situations where T is not completely observed. Thus the surrogate marker may serve as an auxiliary outcome to improve the efficiency of the treatment effect estimate. Most previous research has focused on the first role. The objective of this report is to focus on the potential use of surrogate markers as auxiliary variables and to identify situations when surrogate markers can be useful to increase efficiency when the true outcome is not completely observed. We consider the situations where both S and T are continuous variables. In a single-trial setting, the efficiency gain is small unless S and T are very highly correlated and the amount of missingness is substantial. In a multiple-trial setting, higher efficiency gain is associated with higher trial-level correlation but not individual-level correlation when only S, but not T is measured in a new trial; but, the amount of information recovery from S is negligible. However,

when T is partially observed in the new trial and the individual-level correlation is relatively high, there is substantial efficiency gain by using S and one can extract most of the information on the treatment effect. For design purposes, our results suggest that it is important to collect markers that have high adjusted individual-level correlation with T and at least a small amount of data on T. The results are illustrated using simulations and an example from a glaucoma clinical trial.

Keywords: Auxiliary variable, randomized trial, Meta analysis, Empirical Bayes.

2.1 Introduction

A surrogate marker (S) in a clinical trial is a type of biomarker intended to provide information about the true endpoint (T) and give valid inference on the effect of treatment (Z). Surrogate markers are often intermediate physical or laboratory indicators in a disease progression process, and can be measured earlier and are often easier to collect than the true endpoint. Examples of candidate surrogate markers include CD4 counts in AIDS studies, blood pressure and serum cholesterol level in cardiovascular disease, and prostate-specific antigen in prostate cancer studies. A surrogate marker could serve as a surrogate endpoint and be used to replace T. Alternatively it could serve as an auxiliary outcome to enhance the efficiency of the estimator of the treatment effect on T. When the true endpoints, (e.g. survival time), are rare, later-occurring or costly to obtain, the proper use of good surrogate markers can substantially reduce trial size and duration, hence lower the expense and lead to earlier decision making.

Most previous research on surrogate markers has focused on the potential role of S as a substitute for T. In a landmark article, Prentice (1989) proposed a formal definition for perfect surrogacy and provided validation criteria for the single trial setting. The criteria require that changes in S fully capture the effect of treatment on T. This paper inspired much research in the field, but the criteria are considered too restrictive for practical use. To relax the criteria, a surrogacy measure based on the proportion of the treatment effect explained (PTE) by S was proposed by Freedman *et al* (1992) and further studied and extended by several other authors (e.g., Lin *et al*, 1997; Bycott and Taylor, 1998; Wang and Taylor, 2002). Freedman *et al* (1992) also suggested that the PTE confidence interval's lower bound be > 0.75

for a marker to be acceptable as a surrogate marker. However, this requires the treatment effect on T to be very strong, which is rarely observed in practice (Buyse *et al*, 2000; Bycott *et al*, 1998). The PTE estimator is also highly variable and can be out of the [0,1] range (Lin *et al*, 1997; De Gruttola *et al* 1997); hence, its practical use is limited.

From a biological aspect, there are often multiple causal pathways leading to disease and complex mechanisms by which the treatment functions; hence, a biomarker may or may not mediate the effect of the treatment on T and the surrogacy measures are often not directly transferable from one study to another. Another problem is that S may not capture the harmful side effect of the treatment. These associated uncertainties in the use of S in replacing T to test a new treatment can lead to incorrect, even harmful conclusions (Fleming, Lin and Coombs, 1996; Fleming and DeMets, 1996). As a result, very few biomarkers have been accepted as valid substitutes for T and their potential use as the substitutes has been less than optimistic.

Nonetheless, the clinical research community is still extremely interested in surrogate markers. New biomarkers are being developed at a phenomenal rate, with many being suggested as possible surrogate markers for clinical trials. In this paper, we focus on the use of S as an auxiliary outcome in helping predicting the treatment effect on T. As we shall see, this role of a surrogate marker proves to be more promising. One of the most common scenarios for S to be useful as an auxiliary outcome is when one has more information on S than that on T for a study population. This occurs often in practice, since patients are usually recruited into a trial sequentially in calender time and S is observed more often and early than T, particularly on those who are enrolled early. Previous surrogacy measures are proposed based on summary statistics in order to identify a replacement for T, and they are not usually

suggested explicitly for the purpose of prediction. In the presence of individual-level data, a surrogate marker may actually be effective as an auxiliary outcome in enhancing inference, but not be identified as such using existing surrogacy measures. When S and T are strongly associated, this does not suffice for S to be a substitute for T; as Baker and Kramer (2003) state, "a correlate does not make a surrogate". However, when individual data exist, the existence of strong association between S and T can inform and increase the efficiency of treatment effect estimation, as we demonstrate.

A number of authors have explored the role of surrogate markers as auxiliary outcomes. Much of previous work has focused on situations when the true endpoint is time to failure (Pepe *et al* (1994), Robins and Rotnitzky (1992), Hsu *et al* (2006), Murray and Tsiatis (1996), Fleming *et al.* (1994) and Kosorok and Fleming (1993)). When S and T are continuous data, Venkatraman and Begg (1999) proposed fully nonparametric tests that incorporate the information from S in a single trial setting. However, the opinions in the previous research on the value of surrogate markers have been mixed.

In this paper, we aim to directly investigate the role of S in predicting the treatment effect on T when T is not completely observed. The missing mechanism for T is missing at random (Little and Rubin, 2002). We examine the factors, particularly, the correlation between S and T and the fraction of missing T, that impact the extent of increase in the precision of the treatment effect estimate resulted from utilizing S and identify the situations when S can be beneficial. The results are intended to be of practical value and directly applicable to clinical trials. We consider both single-trial and multiple-trial settings where S and T are continuous and Z is binary. In a single trial setting, the goal is to predict the treatment treatment using

S when T is partially observed. In a multiple-trial setting, we examine the situation when T is either completely missing or partially missing in a new trial when we have information on S, T and Z in the previous trials. The objective is to predict the treatment effect on T in the new trial. In Section 2.2, we examine the efficiency gain of the treatment effect estimate by using S in a single trial setting. In Section 2.3, we consider a multiple-trial setting. We first review several methods used to predict the effect of Z on T in a new trial when T is either completely missing or partially missing in the new trial. Then we evaluate the extent of information recovery from S regarding the treatment effect in the new trial and the associated factors through analytical calculations, simulations and data analysis. In Section 2.4, we present conclusions.

2.2 Single Trial Setting

Suppose that the total number of patients is $m = m_0 + m_1$, with m_0 in the placebo group ($Z = 0$) and m_1 in the treatment group ($Z = 1$). We have information on S on all patients, however, T is observed for only $r = r_0 + r_1$ patients with r_0 in the placebo group and r_1 in the treatment group. For individual j, we assume a commonly used bivariate normal distribution for S_j and T_j given Z_j as follows:

$$\begin{pmatrix} S_j \\ T_j \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_{0Sn} + \delta_{Sn} Z_j \\ \mu_{0Tn} + \delta_{Tn} Z_j \end{pmatrix}, \begin{pmatrix} \sigma_{ss} & \sigma_{st} \\ & \sigma_{tt} \end{pmatrix}\right). \tag{2.1}$$

We want to predict the treatment effect on T, δ_{Tn}, in this trial with the use of S. Let R^2_{indiv} denote the treatment adjusted individual-level correlation between S and T. Under this model assumption, $R^2_{indiv} = \sigma^2_{st}/\sigma_{ss}\sigma_{tt}$. Using a factored likelihood method (Little and Rubin, 2002), we can obtain the maximum likelihood estimates and the corresponding inverse information matrix. Under missing completely at

random assumption, the large-sample variance of the estimated treatment effect, $\hat{\delta}_{Tn}$, can be approximated by

$$(2.2) \quad var(\hat{\delta}_{Tn}) \approx \frac{\sigma_{tt}}{r_0}\left\{1 - R^2_{indiv}\frac{m_0 - r_0}{m_0}\right\} + \frac{\sigma_{tt}}{r_1}\left\{1 - R^2_{indiv}\frac{m_1 - r_1}{m_1}\right\}.$$

If we were to observe all T_1, ..., T_m, then the variance of the corresponding treatment effect estimator $\hat{\delta}^o_{Tn}$ is given by $var(\hat{\delta}^o_{Tn}) = \sigma_{tt}/m_0 + \sigma_{tt}/m_1$, while the simple estimator based on the observed data that ignores S would have variance $\sigma_{tt}/r_0 + \sigma_{tt}/r_1$. The relative efficiency (RE) of $\hat{\delta}_{Tn}$ compared with $\hat{\delta}^o_{Tn}$ equals $var(\hat{\delta}^o_{Tn})/var(\hat{\delta}_{Tn})$. When σ_{tt} and the percentage of missingness are fixed, the single most important factor in the relative efficiency is R^2_{indiv}. The higher the absolute correlation, the greater the extent of information recovery from S and the more useful S can be in predicting the treatment effect. Figure 2.1A plots the efficiency of $\hat{\delta}_{Tn}$ relative to $\hat{\delta}^o_{Tn}$ against different levels of the correlation and missing T when $m_0 = m_1$ and $r_0 = r_1$. When the correlation is high, (e.g. $R^2_{indiv} > 0.8$) and the proportion missing is less than 40%, through the utilization of S, we can obtain the estimate $\hat{\delta}_{Tn}$ with precision close to that based on completely observed T. When the correlation is greater than 0.9, coupled with even only 10% of available T, we can retrieve much of the information on the treatment effect from S. However, when S and T are not highly correlated (e.g., $R^2_{indiv} < 0.3$), the extent of the information recovery is small. As can be seen from equation (2.2), there is no gain in efficiency from using S if $R^2_{indiv} = 0$.

2.3 Multiple Trial Setting

When we can identify a group of trials which have similar treatment groups and patient populations, it is natural to use a meta-analytic approach to predict the treatment effect in a new trial. This approach could allow one to account for the heterogeneity among different trials and borrow information from previous trials to

improve the efficiency.

2.3.1 The Model

Suppose we have n randomized trials, $i = 1, ..., n$, where the nth trial is new. Let (S_{ij}, T_{ij}, Z_{ij}) represent S, T, and Z for the individual j in the trial i. We are interested in predicting the actual treatment effect on T in the new trial (δ_{Tn}) based on previous $(n-1)$ existing trials. We adopt a commonly used bivariate mixed model for the joint distribution of S_{ij}, T_{ij} and Z_{ij}:

$$\begin{aligned} S_{ij} &= \alpha_0 + \alpha_1 Z_{ij} + a_{0i} + a_{1i} Z_{ij} + \epsilon_{Sij} \\ T_{ij} &= \gamma_0 + \gamma_1 Z_{ij} + r_{0i} + r_{1i} Z_{ij} + \epsilon_{Tij} \end{aligned} \tag{2.3}$$

where

$$\begin{pmatrix} \epsilon_{Sij} \\ \epsilon_{Tij} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{ss} & \sigma_{st} \\ & \sigma_{tt} \end{pmatrix} \right), \tag{2.4}$$

and

$$\begin{pmatrix} a_{0i} \\ r_{0i} \\ a_{1i} \\ r_{1i} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, D = \begin{pmatrix} d_{ss} & d_{st} & d_{sa} & d_{sr} \\ & d_{tt} & d_{ta} & d_{tr} \\ & & d_{aa} & d_{ar} \\ & & & d_{rr} \end{pmatrix} \right). \tag{2.5}$$

With this formulation, the treatment effect in the nth trial is $\delta_{Tn} = \gamma_1 + r_{1n}$. Let $Y_i^T = (S_{ij}, T_{ij})$, $\epsilon_i^T = (\epsilon_{Sij}, \epsilon_{Tij})$, $\beta^T = (\alpha_0, \gamma_0, \alpha_1, \gamma_1)$ and $\eta_i^T = (a_{0i}, r_{0i}, a_{1i}, r_{1i})$. The model (2.3) can be written in the general mixed model notation as $Y_i = X_i\beta + U_i\eta_i + \epsilon_i$, where β denotes the fixed effects, η_i denotes the random effects, X_i and U_i are the corresponding design matrices. The vector Y_i follows a bivariate normal distribution with mean $X_i\beta$ and variance $V_i = U_i D U_i^t + \Sigma_i$.

2.3.2 Methods for Predicting the New Treatment Effect δ_{Tn}

Estimation by Buyse *et al.* (2000) (BMBRG)

Buyse, Molenberghs, Burzykowski, Renard, and Geys (BMBRG) (2000) assumed the same model and suggested a method to estimate δ_{Tn} when T is completely unobserved in the nth trial. First, they fit a bivariate mixed model to the data from trial 1 through $(n-1)$ to obtain the estimates of D and β. Second, they fit a linear regression $S_{nj} = \mu_{Sn} + \delta_{Sn} Z_{nj} + \epsilon_{Snj}$ to the surrogate marker in the nth trial. One then obtains that $\hat{\alpha}_{0n} = \hat{\mu}_{Sn} - \hat{\alpha}$ and $\hat{\alpha}_{1n} = \hat{\delta}_{Sn} - \hat{\alpha}_1$. Assuming β, D, a_{0n} and b_{0n} are known, BMBRG showed that δ_{Tn} follows a normal distribution with mean

$$E(\delta_{Tn}) = \gamma_1 + \begin{pmatrix} d_{sr} & d_{ar} \end{pmatrix} \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} a_{0n} \\ a_{1n} \end{pmatrix}, \tag{2.6}$$

and variance

$$var(\delta_{Tn}) = d_{rr} - \begin{pmatrix} d_{sr} & d_{ar} \end{pmatrix} \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sr} \\ d_{ar} \end{pmatrix}. \tag{2.7}$$

When β, D, a_{0n} and a_{1n} are unknown, their estimates can be used. However, it can lead to the underestimation of the true variance $var(\hat{\delta}_{Tn})$.

Estimation by Gail *et al.* (2000) (GPHC)

Gail, Pfeiffer, Houwelingen, and Carroll (GPHC) (2000) proposed to estimate δ_{Tn} using an estimating equation approach which does not involve modeling the joint distribution of $(S_{ij},\ T_{ij})$ at the individual level. The method also addresses the situation when T is completely missing in the new trial. Let $\mu_{Ti}^T = (\mu_{0Ti}, \mu_{1Ti})$ represent the marginal means of T in the $Z = 0$ group and the $Z = 1$ group in the ith trial, respectively and similarly for $\mu_{Si}^T = (\mu_{0Si}, \mu_{1Si})$. GPHC assume that $(\hat{\mu}_{Ti}^T, \hat{\mu}_{Si}^T)^T$ follows a multivariate normal distribution with the overall mean and covariance $\phi + \omega_i$,

where ϕ is a 4×4 matrix representing the between-trial variance and ω_i is a 4×4 matrix with two block diagonal matrices denoting the within-trial variance for each treatment group. The elements of μ_{Ti}, μ_{Si}, and ϕ are connected with the parameters in the model (2.3) in this way: $\mu_{0Ti} = \gamma_0 + r_{0i}$, $\mu_{1Ti} = \gamma_0 + r_{0i} + \gamma_1 + r_{1i}$, $\mu_{0Si} = \alpha_0 + a_{0i}$, $\mu_{1Si} = \alpha_0 + a_{0i} + \alpha_1 + a_{1i}$, $\phi_{11} = d_{tt} + d_{bb} + 2d_{tb}$, $\phi_{12} = d_{ts} + d_{ab} + d_{ta} + d_{sb}$, $\phi_{13} = d_{tt} + d_{tb}$, $\phi_{14} = d_{ts} + d_{sb}$, $\phi_{22} = d_{ss} + d_{aa} + 2d_{sa}$, $\phi_{23} = d_{st} + d_{ta}$, $\phi_{24} = d_{ss} + d_{sa}$, $\phi_{33} = d_{tt}$, $\phi_{34} = d_{st}$ and $\phi_{44} = d_{ss}$.

GPHC show that $\hat{\mu}_{Tn}$ follows a normal distribution with mean

$$E(\hat{\mu}_{Tn}) = \begin{pmatrix} \gamma_0 \\ \gamma_0 + \gamma_1 \end{pmatrix} + \begin{pmatrix} \phi_{12} & \phi_{14} \\ \phi_{23} & \phi_{34} \end{pmatrix} \begin{pmatrix} \phi_{22} + \omega_{22n} & \phi_{24} \\ \phi_{24} & \phi_{44} + \omega_{44n} \end{pmatrix}^{-1} \begin{pmatrix} a_{0n} \\ a_{0n} + a_{1n} \end{pmatrix},$$

and variance

$$var(\hat{\mu}_{Tn}) = \begin{pmatrix} \phi_{11} & \phi_{13} \\ \phi_{13} & \phi_{33} \end{pmatrix} - \begin{pmatrix} \phi_{12} & \phi_{14} \\ \phi_{23} & \phi_{34} \end{pmatrix} \begin{pmatrix} \phi_{22} + \omega_{22n} & \phi_{24} \\ \phi_{24} & \phi_{44} + \omega_{44n} \end{pmatrix}^{-1} \begin{pmatrix} \phi_{12} & \phi_{14} \\ \phi_{23} & \phi_{34} \end{pmatrix}^{T},$$

where ω_{22n} denotes the variance corresponding to $\hat{\mu}_{0Sn}$ and ω_{44n} for $\hat{\mu}_{1Sn}$.

The treatment effect on T in the new trial, δ_{Tn}, can be estimated by:

$$E(\hat{\delta}_{Tn}) = \begin{pmatrix} -1 & 1 \end{pmatrix} E(\hat{\mu}_{Tn}), \tag{2.8}$$

with variance obtained as

$$var(\hat{\delta}_{Tn}) = \begin{pmatrix} -1 & 1 \end{pmatrix} var(\hat{\mu}_{Tn}) \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \tag{2.9}$$

If we drop the terms w_{22n} and w_{44n} from the GPHC variance formula in (2.9), the expression becomes idential to the BMBRG variance derivation in (2.7). The GPHC formula takes into account the uncertainty associated with estimating a_{0n} and a_{1n} while the BMBRG formula does not. Similar to BMBRG, GPHC also assume that β and D are known.

Estimation by Henderson (1984) (HD)

While both BMBRG and GPHC methods only apply to the situation when T is completely missing in the new trial, the method introduced here can be generalized to the situations when T is either completely missing, partially missing or completely observed in the new trial. Using the generalized mixed model notation, we can obtain the estimates of β and η_n by solving the mixed model equation described by Henderson (1984) and their sum follow a normal distribution with mean

$$E(\hat{\beta} + \hat{\eta}_n) = \beta + DU_n^T V_n^{-1}(Y_n - X_n\beta). \tag{2.10}$$

and variance

$$var\left(\hat{\beta} + \hat{\eta}_n - \beta - \eta_n\right) = \left(\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right)^{-1} + D - DU_n^T V_n^{-1} U_n D + DU_n^T V_n^{-1} X_n$$
$$\left(\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right)^{-1} X_n^T V_n^{-1} U_n D - 2DU_n^T V_n^{-1} X_n \left(\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right)^{-1}.$$

The treatment effect for the nth trial has mean

$$E(\hat{\delta}_{Tn}) = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} (\beta + \eta_n) \tag{2.11}$$

and variance

$$var(\hat{\delta}_{Tn}) = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} var(\hat{\beta} + \hat{\eta}_n - \beta - \eta_n) \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix}^T. \tag{2.12}$$

Note that $\hat{\eta}_n$ is the best linear unbiased predictor (BLUP) and can be derived as an empirical Bayes estimator (Laird and Ware, 1982, Robinson, 1991). Different from GPHC and BMBRG, this variance formula accounts for the uncertainty associated with estimating β, but it treats D and Σ as known quantities. In an effort to account for all the uncertainties, GPHC proposed a bootstrap method.

Bayesian Estimation (denoted by Bayes)

An alternative method to obtain the distributions of the parameters of interest is a fully Bayesian estimation method which is applicable when T is either partially missing or completely missing. We assume noninformative priors for the fixed effects, i.e., $p(\alpha_0) \propto 1$, $p(\gamma_0) \propto 1$, $p(\alpha_1) \propto 1$, and $p(\gamma_1) \propto 1$, and vague priors for the rest of parameters, specifically, $\Sigma^{-1} \sim W(a, E)$ and $D^{-1} \sim W(c, F)$, where W refers to the Wishart distribution. We can choose $a = 3$, $c = 5$, $E = (a+1)^{-1} I_2$ and $F = (c+1)^{-1} I_4$. A data augmentation method can be used to implement the procedure by iterating the following step 1 and 2 until the parameters reach convergence:

Step 1: Impute missing T_{nj}'s from a normal distribution with mean and variance:

$$\begin{aligned} E(T_{nj}|S_{nj}, Z_{nj}) &= \gamma_0 + r_{0n} - \sigma_{st}\sigma_{ss}^{-1}(\alpha_0 + a_{0n}) \\ &+ (\gamma_1 + r_{1n} - \sigma_{st}\sigma_{ss}^{-1}(\alpha_1 + a_{1n}))Z_{nj} + \sigma_{st}\sigma_{ss}^{-1}S_{nj}, \\ Var(T_{nj}|S_{nj}, Z_{nj}) &= \sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1} \end{aligned}$$

Step 2: Treat the data after imputation as complete data, and apply Gibbs sampling to estimate the parameters of interest:

$$\begin{aligned} p(D^{-1}|\eta) &\propto W(n + c, (\sum_{i=1}^{n} \eta_i \eta_i^T + F^{-1})^{-1}) \\ p(\Sigma^{-1}|X, Y, Z, \beta, \eta) &\propto W(\sum_{i=1}^{n} m_i + a, (VS + E^{-1})^{-1}) \\ p(\eta_i|X, Y, Z, \Sigma, D) &\propto MVN(VE \times (\sum_{j=1}^{m_i} Z_{ij}^T \Sigma^{-1}(Y_{ij} - X_{ij}\beta)), VE) \\ p(\beta|X, Y, Z, \eta_i, \Sigma) &\propto MVN(VB \times (\sum_{i=1}^{n}\sum_{j=1}^{m_i} X_{ij} \Sigma^{-1}(Y_{ij} - U_{ij}\eta_i)), VB) \end{aligned}$$

where,

$$VS = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{ij} - X_{ij}\beta - U_{ij}\eta_i)(Y_{ij} - X_{ij}\beta - U_{ij}\eta_i)^T,$$
$$VB = (\sum_{i=1}^{n}\sum_{j=1}^{m_i} X_{ij}^T\Sigma^{-1}X_{ij})^{-1},$$
$$VE = (\sum_{j=1}^{m_i} U_{ij}^T\Sigma^{-1}U_{ij} + D^{-1})^{-1}.$$

From the distributions of β and η_i, we can obtain the distribution of δ_{Tn}. The Bayesian estimation method naturally takes into consideration of the uncertainty associated with estimating every parameter (Louis and Zelterman (1994)), but it is sensitive to the prior specifications. It is computationally challenging to conduct extensive simulations to evaluate the properties of this method, as such, we do not present the simulation results in this report; however, it is very feasible to analyze data using this method.

2.3.3 Efficiency Gain and Correlation

In this section, we study the precision of the predicted treatment effects ($\hat{\delta}_{Tn}$) and the factors that impact it, particularly, the correlation between S and T.

Correlation

In a multiple trial setting, with a bivariate mixed model assumption, the treatment adjusted individual-level or within-trial correlation between S and T is R^2_{indiv}, with the same definition as that in a single trial setting. The trial-level correlation between S and T is defined by Buyse *et al.* (2001) as

$$R^2_{trial} = \frac{\begin{pmatrix} d_{sr} & d_{ar} \end{pmatrix}\begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1}\begin{pmatrix} d_{sr} \\ d_{ar} \end{pmatrix}}{d_{rr}}.$$

The within-trial correlation R^2_{indiv} captures the relationship between S and T at the individual level. When $R^2_{indiv} = 1$, S has the perfect linear association with T. The between-trial correlation R^2_{trial} assesses how well the treatment effect on T in the new trial can be predicted by the treatment effects on S. When $R^2_{trial} = 1$, it implies that the treatment effect on T in the ith trial can be predicted without error from the treatment effect on S. While the trial-level correlation is identified as the key factor that impacts the degree of efficiency gain from S in the research by Buyse *et al.* (2000) and Gail *et al* (2000), as we shall see in the following, we identify that the individual-level correlation plays an even more important role than the trial-level correlation in obtaining substantial efficiency gain from S with respect to the estimated treatment effect on T when T is partially observed.

Prediction Precision and Correlation

Motivated by the fact that the estimate of $\beta + \eta_i$ in (2.10) can be obtained as the posterior mean of its posterior distribution when flat priors are assumed for β (Harville, 1976; Laird and Ware, 1982), in the following we calculate the posterior variance of δ_{Tn} for the purpose of examining the factors that impact the precision of the estimated treatment effect with the same prior specifications. Let r be the number of patients in the new trial on whom we have information on both S and T. Assume β, D and Σ are known quantities, we obtain the posterior variance of δ_{Tn} as (details in Appendix):

$$var(\delta_{Tn}) = \begin{pmatrix} 0 & 1 \end{pmatrix} \left(\Psi_d^{-1} + \Phi_e^{-1}\right)^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{2.13}$$

where, Ψ_d is a function only of the between-trial covariances given by $\Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$ and Φ_e is a function only of the within-trial covariances given by $\Phi_e = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$.

The elements of Ψ_d and Φ_e are listed below:

$$\begin{aligned}
\Psi_{11} &= \begin{pmatrix} d_{tt} & d_{tr} \\ d_{tr} & d_{rr} \end{pmatrix}, \quad \Psi_{12} = \begin{pmatrix} d_{st} & d_{ta} \\ d_{sr} & d_{ar} \end{pmatrix}, \quad \Psi_{21} = \begin{pmatrix} d_{st} & d_{sr} \\ d_{sr} & d_{ar} \end{pmatrix}, \\
\Psi_{22} &= \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}, \quad \phi_{11} = \frac{(\sigma_{tt}(1-R^2_{indiv}))\sum_{j=1}^{r} Z^2_{nj}}{r\sum_{j=1}^{r} Z^2_{nj} - (\sum_{j=1}^{r} Z_{nj})^2}, \\
\phi_{12} &= \frac{(\sigma_{tt}(1-R^2_{indiv}))\sum_{j=1}^{r} Z_{nj}}{r\sum_{j=1}^{r} Z^2_{nj} - (\sum_{j=1}^{r} Z_{nj})^2}, \quad \phi_{22} = \frac{r(\sigma_{tt}(1-R^2_{indiv}))}{r\sum_{j=1}^{r} Z^2_{nj} - (\sum_{j=1}^{r} Z_{nj})^2}.
\end{aligned}$$

When there is equal number of patients per treatment group in the new trial, the estimates of Φ_e simplifies to

$$\phi_{11} = \frac{2\sigma_{tt}(1-R^2_{indiv})}{r}, \quad \phi_{12} = \frac{2\sigma_{tt}(1-R^2_{indiv})}{r}, \; \phi_{22} = \frac{4\sigma_{tt}(1-R^2_{indiv})}{r}.$$

When T is completely missing in the nth trial, i.e., $r = 0$, the conditional variance simplifies to:

$$var(\delta_{Tn}) = \begin{pmatrix} 0 & 1 \end{pmatrix} \Psi_d \begin{pmatrix} 0 & 1 \end{pmatrix}^T = d_{rr}\left(1 - R^2_{trial}\right), \tag{2.14}$$

an expression equivalent to the variance formula in (2.7). From this, when T is completely missing in the new trial, the factors that determine the precision of the predictor of the treatment effect on T are R^2_{trial} and d_{rr}. When T is partially observed, the additional important factors are within-trial level including R^2_{indiv}, σ_{tt} and r. We also find that since the within-trial covariances in Φ_e are usually significantly smaller than the between-trial covariances in Ψ_d, Φ_e dominates and Ψ_d has almost negligible impact on the conditional variance of $\hat{\delta}_{Tn}$ in (2.13).

Note that the conditional posterior variance in equations (2.13) and (2.14) underestimate the prediction variance because they treat β, D, and Σ as known quantities. Morris (1983) and Ghosh and Rao (1994) showed that a better estimate of the prediction variance can be obtained by adding to the conditional posterior variance a

second term that takes into account the uncertainty about all parameters. However, our simulation studies show that the conditional variance usually accounts for the majority of the total variance, and a comparison between (2.13) and (2.14) should suffice to provide algebraic intuition about the prediction variance.

2.3.4 Simulations

The Setup

We conduct simulation studies to evaluate the bias, efficiency and coverage rates of the confidence intervals for the predicted treatment effect in a new trial using the above methods. For comparison purposes, we also estimate δ_{Tn} based on all of T before any missingness occurs in the new trial using two approaches: 1) the simple estimate without any distributional assumption, (denoted by ALLSE). That is, $\hat{\delta}_{Tn} = \sum_k T_{nk1}/m_{n1} - \sum_l T_{nl0}/m_{n0}$, where T_{nk1} represents T on patient k in the $Z = 1$ group in the nth trial and similarly for T_{nl0}, m_{n1} represents the number of patients in the $Z = 1$ group in the nth trial and similarly for m_{n0}. 2) the estimate and its variance obtained using formulae in (2.11) and (2.12) assuming the bivariate mixed model (denoted by ALLHD). When appropriate, we also estimate δ_{Tn} solely based on the observed and incomplete T in the nth trial using the simple estimation method in the same way as in 1) (denoted by INSE).

We generate 800 data sets based on the bivariate mixed model in (2.3). The parameter specifications are: $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$. To examine the impact of the trial-level correlation, we vary the correlation matrices for the random effects: $\begin{pmatrix} 1 & 0.57 & 0.37 & 0.22 \\ 0.57 & 1 & 0.24 & 0.21 \\ 0.37 & 0.24 & 1 & 0.3 \\ 0.22 & 0.21 & 0.3 & 1 \end{pmatrix}$,

$$\begin{pmatrix} 1 & 0.57 & 0.37 & 0.22 \\ 0.57 & 1 & 0.24 & 0.21 \\ 0.37 & 0.24 & 1 & 0.7 \\ 0.22 & 0.21 & 0.7 & 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 1 & 0.57 & 0.37 & 0.22 \\ 0.57 & 1 & 0.24 & 0.21 \\ 0.37 & 0.24 & 1 & 0.9 \\ 0.22 & 0.21 & 0.9 & 1 \end{pmatrix},$$ which correspond to the trial-level correlations of $R^2_{trial} = 0.1, 0.5$ and 0.8, respectively. To examine the impact of the individual-level correlation, we vary R^2_{indiv} from 0.1, 0.5, to 0.9. We vary n, m, and the percentage of missingness in the new trial (denoted by p). For each data set, we have a true treatment effect δ_{Tn} and its average across 800 data sets is $\bar{\delta}_{Tn}$. For each data set and each method used, we obtain the estimate of $\hat{\delta}_{Tn}$, its standard error and an indicator variable for whether the 95% confidence interval contains δ_{Tn} or not. We examine the method's performance by its average bias (Bias $= \bar{\hat{\delta}}_{Tn} - \bar{\delta}_{Tn}$), the average standard error (SE), the rooted mean squared error (RMSE $= \sqrt{\sum(\hat{\delta}_{Tn} - \delta_{Tn})^2/800}$), and the coverage rate (CR) over all simulated data sets. The relative efficiency (RE) of two estimators is approximated as the inverse of the ratio of the two corresponding RMSE^{2}s, because we will see all estimates are unbiased.

Information Recovery from S When T is Completely Missing in the New Trial

In Tables 2.1, 2.2, and 2.3, we present Bias, RMSE, SE and CR of the estimates of δ_{Tn} using all five methods: ALLSE, ALLHD, BMBRG, GPHC, and HD from simulations with various combinations of n, m, R^2_{indiv}, and R^2_{trial}. We vary m and R^2_{indiv} in Table 2.1, n and R^2_{indiv} in Table 2.2, R^2_{indiv} and R^2_{trial} in Table 2.3. All five methods produce unbiased estimates. When T is completely missing in the new trial, BMBRG, GPHC and HD all underestimate the variances of $\hat{\delta}_{Tn}$ which lead to lower than the 95% nominal level of coverage rates. HD appears to give slightly higher CRs than GPHC; GPHC generates slightly higher CRs than BMBRG. When the number

of trials n is relatively small, the extent of underestimation can be more severe, with less than 90% coverage rate. Comparing ALLSE with ALLHD, we have a slightly more precise estimate from ALLHD than that from ALLSE, and the efficiency gain from the bivariate normal assumption is small. There is a minor gain in the precision when the patient size per trial and the number of trials increase. Figure 2.1B shows the relatively efficiency of $\hat{\delta}_{Tn}$ when T is completely missing in the new trial compared to the estimate before any deletion in T occurs using the HD method as we vary R^2_{indiv} and R^2_{trial}. We find that while the increases in R^2_{indiv} have negligible impact on the precision, the increase in R^2_{trial} can improve the precision more than any other factor. These findings agree with the algebraic intuition from the variance formula in (2.14). Relative to the estimate based on completely observed data, the relative efficiency varies from 0.7%, 1.2% to 3.4% as we increase R^2_{trial} from 0.1, 0.5 to 0.8. As a result, when we completely rely on S and summary statistics from previous trials to predict the treatment effect on T in the new trial, the extent of information recovery is very limited and the precision of $\hat{\delta}_{Tn}$ is often insufficient to be clinically useful.

Information Recovery from S when T is Partially Observed in the New Trial

We consider the situation when 50% of T in the new trial are observed. The HD method still applies although GPHC and BMBRG are not applicable anymore. Tables 2.4, 2.5 and 2.6 list Bias, RMSE, SE and CR of $\hat{\delta}_{Tn}$ from simulations under various combinations of n, m, R^2_{indiv} and R^2_{trial}. We vary m and R^2_{indiv} in Table 2.4, n and R^2_{indiv} in Table 2.5, R^2_{indiv} and R^2_{trial} in Table 2.6. Different from before, when T is 50% observed, the underestimation of the variance of $\hat{\delta}_{Tn}$ using HD is negligible and the CR is close to or at the 95% nominal level, even when the number of trials is relatively small, such as $n = 10$. While a greater number of trials does not lead to

higher precision of the estimate, the increase in the number of patients per trial can improve the precision when R^2_{indiv} is high. Figure 2.1C presents the relative efficiency of $\hat{\delta}_{Tn}$ when T is 50% missing compared with the estimate before any deletion of T using the HD method. We find that higher R^2_{indiv} can lead to a large gain of efficiency from the use of S. When R^2_{indiv} is large (e.g., 0.7 or 0.9), most of the information on δ_{Tn} is recovered from S and the precision of the estimate is close to the estimate when T is completely observed. On the other hand, the magnitude of R^2_{trial} does not have any impact on the amount of efficiency gain from S. The observations here are in agreement with the variance formula in (2.13).

Information Recovery and Percentage of Missingness

We examine further the relationship between the extent of information recovery by incorporating S and the percentage of observed T using the HD method. Let $R^2_{trial} = 0.5$, $n = 40$, $m = 100$ and the percentage of observed T varies from 0%, 10%, 30%, 50%, 80%, to 100%. Table 2.7 presents Bias, RMSE, SE and CR of $\hat{\delta}_{Tn}$. Figure 2.1D lists the relative efficiency of $\hat{\delta}_{Tn}$ when T is partially missing compared with the estimate before any deletion of T. Naturally, the higher the proportion of available T, the smaller the RMSE, and thus the greater the precision for the treatment effect prediction. Interestingly, we find that there is a substantial efficiency gain from the information on S with even a small fraction of observed T, particularly when R^2_{indiv} is high. For example, when 30% T are observed, the lost information due to missingness is almost completely recovered from S when $R^2_{indiv} = 0.9$.

2.3.5 Data Analysis: a Glaucoma Study

The evaluation of the extent of information recovery from S in predicting the treatment effect on T in a new trial is illustrated using the Collaborative Initial

Glaucoma Treatment Study (CIGTS) (Musch *et al*, 1999). Glaucoma is a group of diseases that cause vision loss and is a leading cause of blindness. High pressure in the eyes, i.e. intraocular pressure (IOP), is a major risk factor of glaucoma. The CIGTS is a randomized multicenter clinical trial to compare the effects of two types of treatments, surgery and medicine, on reducing IOP among glaucoma patients. Patients are enrolled between 1993 and 1997. A total of 607 patients are included in the study and among them, 307 are randomly assigned into the medicine group. IOP (recorded in mmHg) has been measured at different time points following the treatment. For the purpose of this paper, we take the true endpoint to be the IOP measurement at month 96 and the surrogate marker to be the IOP measurement at the 12th month. We assume that the IOP measurements are normally distributed. To evaluate the situation of a meta-analysis where data are from different trials; we treat the different centers in the CIGTS study as independent trials testing a similar group of treatments. A preliminary analysis of these data shows that the estimate of the between-trial variances, $\hat{D}$, is non-positive definite. Mimicking the approach of Gail *et. al.* (2000), we rescale up the data size by simulating S_{ij} and T_{ij} from bivariate normal distributions for each trial and treatment group with the trial-specific and treatment-specific means and variance-covariances from the real data. The CIGTS study includes 14 centers and from which we delete five centers (i.e., 5, 7, 12, 13, 14) either because they had too few observations or because of non-positive definite covariance matrices within center. We also deleted two outliers that are greater than 35mmHg. For the $n = 9$ centers included, we increase the sample sizes to 335, 176, 385, 264, 539, 368, 286, 528, and 319. The trial-specific and treatment-specific means and correlations for S and T are listed in Table 2.8. The HD method is used to fit the rescaled data with $\hat{D}$ as being positive definite,

$\hat{R}^2_{trial} = 0.25$ and $\hat{R}^2_{indiv} = 0.15$. We randomly select Center 8 as the new trial and randomly delete some proportion of T in Center 8 to examine the extent of efficiency gain through the use of S. The missing mechanism is missing completely at random (Little and Rubin, 2002). The results are listed in Table 2.9. Without missing T, $\hat{\delta}_{Tn}$ is -2.45 with the standard error of 0.29. When T is completely missing, $\hat{\delta}_{Tn}$ is -1.58 with the standard error of 0.79. When 20% or 50% of T are missing, the precision of $\hat{\delta}_{Tn}$ using S is comparable to that based on completely observed T. Even with 80% missing, the SE is substantially smaller than that when 100% T is missing. For further illustration, when Center 9 is treated as a new trial and we obtain similar results.

2.4 Discussion

In this report, we examine the role of surrogate markers as auxiliary variables in predicting the treatment effect and identify situations when surrogate markers can be beneficial when S and T are continuous in either a single-trial or multiple-trial setting. While previous literature on the use of surrogate markers as substitutes for the true endpoints has been mostly negative and the proposed surrogate measures are often not useful in practice, we show that it is possible for surrogate markers to be useful as auxiliary variables in enhancing the inference on the true endpoint. Although high correlation between S and T does not qualify S as a good surrogate (Baker and Kramer, 2003), we show that the correlation is a critical measure in determining the extent of information recovery from S.

In a single trial, the amount of efficiency gain through S is small except in rare occasions when the correlation between S and T is extremely high. In a multiple trial setting, when T is completely unobserved, R^2_{indiv} plays little role in the amount

of information recovered from S; on the other hand, the higher the R^2_{trial}, the higher the efficiency gain from S. However, even with a high R^2_{trial}, the predicted treatment effect based on data from other trials and surrogate markers in the new trial solely can be too imprecise to be clinically useful. On the contrary, when T is partially observed in the new trial, we find that a high R^2_{indiv} is a very important determinant in increasing the precision of the predicted treatment effect from S but the impact of R^2_{trial} is negligible. With even a small fraction of T and a high R^2_{indiv}, the information on the treatment effect is mostly recovered and the prediction precision is close to that when T is completely observed. It appears that some data on T are essential to provide the basis for individual-level predictions of T from S using the distributional assumption, and hence to give a much more efficient treatment estimate. The importance of R^2_{indiv} in prediction differs from that in a single trial setting. In a single trial setting, in general the amount of information that can be recovered from S when T is partially observed is limited, and only noteworthy if the correlation between S and T is much higher than that in a multiple trial setting. When T is completely missing, we compare the BMBRG, GPHC and HD methods. Each method gave unbiased estimates; but the variances were underestimated, particularly when the number of the trials was small. Either a bootstrap or fully Bayesian methods could remedy this problem. When T is partially observed, the underestimation from the HD method becomes negligible.

The data example used for illustration purposes could be more ideal as we treat the multi-center glaucoma data as a multi-trial data and rescale up the real data. While the sharing of the clinical trial data is currently limited, our results can be generalizable. When it becomes more common for pharmaceutical companies or universities to share clinical trial data publicly, further exploration of the use of

surrogate markers would be of great interest and importance. Acknowledging the limitations associated with finite number of simulations, we do provide evidence that the use of the surrogate markers can be promising in terms of early treatment effect predictions.

In our study, we consider continuous S and T. It is likely that our findings can apply to cases when S and T are other types of data such as binary, categorical and time-to-event. Efficiency gains from S are also possibly substantial with parametric models that assume a very close and structural relationship between a surrogate marker and a true endpoint. For example, with a three-stage model specification for a time-to-event surrogate marker and a true endpoint, Cook and Lawless (2001) showed that one can achieve significant efficiency gains by using a surrogate marker.

In conclusion, surrogate markers would seem to have a more useful role as auxiliary variables than as replacements of the true endpoint. Future research should focus on the role of surrogate markers as auxiliary variables, to identify scenarios when the surrogate marker can increase the precision of the treatment effect. For design purposes, our results suggest that it is important to collect at least some data on the true endpoint and more information on the surrogate marker which has high adjusted individual-level correlation with the true endpoint. With appropriate utilization of high quality surrogate markers in estimating the treatment effect when the true endpoint is not completely observed, one can reach a desired level of precision earlier, hence shortening the study period and reducing the cost of a study.

2.5 Appendix: Conditional Posterior Variance of δ_{Tn}

The HD estimate of δ_{Tn} in (2.10) can be interpreted as a posterior mode estimate when we assume flat priors for the fixed effects and multivariate normal priors for the random effects. When β, D and Σ are known, the conditional posterior variance of δ_{Tn} can approximate the variance of $\hat{\delta}_{Tn}$ (Ghosh and Rao, 1994). Let $\alpha_0 + a_{0i} = \mu_{0Si}$, $\gamma_0 + r_{0i} = \mu_{0Ti}$, $\alpha_1 + a_{1i} = \delta_{Si}$ and $\gamma_1 + r_{1i} = \delta_{Ti}$, we can rewrite the model (2.3) as

$$\begin{aligned} S_{ij} &= \mu_{0Si} + \delta_{Si} Z_{ij} + \epsilon_{Sij} \\ T_{ij} &= \mu_{0Ti} + \delta_{Ti} Z_{ij} + \epsilon_{Tij}. \end{aligned}$$

Assume there are r observations with both S and T observed and $m_n - r$ observations with just S observed in the nth trial. The likelihood can be written as:

$$
\begin{aligned}
& L(\phi|S,T,Z) \\
= \; & \left\{\prod_{i=1}^{n-1}\left[\prod_{j=1}^{m_i} N(Y_{ij}|\beta,\eta_i,\Sigma,D,Z_{ij})\right]\right\} \\
& \left\{\prod_{j=1}^{r} N(Y_{ij}|\eta_i,\beta,Z_{ij},\Sigma,D)\prod_{j=1}^{m_n-r} N(S_{nj}|\mu_{0sn},\delta_{sn},Z_{nj},\sigma_{ss})\right\} \\
= \; & \prod_{i=1}^{n-1}\left[\prod_{j=1}^{m_i}\frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}\begin{pmatrix} S_{ij}-\mu_{0Si}-\delta_{Si}Z_{ij} \\ T_{ij}-\mu_{0Ti}-\delta_{Ti}Z_{ij}\end{pmatrix}^T \Sigma^{-1}\begin{pmatrix} S_{ij}-\mu_{0Si}-\delta_{Si}Z_{ij} \\ T_{ij}-\mu_{0Ti}-\delta_{Ti}Z_{ij}\end{pmatrix}\right\}\right] \\
& \prod_{j=1}^{r}\frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}\begin{pmatrix} S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj} \\ T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}\end{pmatrix}^T \Sigma^{-1}\begin{pmatrix} S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj} \\ T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}\end{pmatrix}\right\} \\
& \prod_{j=1}^{m_n-r}\frac{1}{\sqrt{2\pi}\sigma_{ss}^{1/2}} exp\left\{-\frac{1}{2}(S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj})^2\right\} \\
& \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}|D|^{1/2}}\exp\left\{-\frac{1}{2}\begin{pmatrix} \mu_{0Si}-\alpha_0 \\ \mu_{0Ti}-\gamma_0 \\ \delta_{Si}-\alpha_1 \\ \delta_{Ti}-\gamma_1\end{pmatrix}^T D^{-1}\begin{pmatrix} \mu_{0Si}-\alpha_0 \\ \mu_{0Ti}-\gamma_0 \\ \delta_{Si}-\alpha_1 \\ \delta_{Ti}-\gamma_1\end{pmatrix}\right\},
\end{aligned}
$$

which is equivalent to the expression of $likelihood \times prior$ when assuming flat priors for the fixed effects and multivariate normal distributions for the random effects. The conditional posterior distributions of μ_{0Tn} and δ_{Tn} given the data and all other

parameters are proportional to:

$$
\begin{aligned}
\begin{matrix}\mu_{0Tn}\\ \delta_{Tn}\end{matrix} \mid \cdot \;\propto\; & \prod_{j=1}^{r} \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2} ME^T \times \Sigma^{-1} \times ME\right\} \\
& \frac{1}{\sqrt{2\pi}|D|^{1/2}} \exp\left\{-\frac{1}{2}\begin{pmatrix} \mu_{0Sn}-\alpha_0 \\ \mu_{0Tn}-\gamma_0 \\ \delta_{Sn}-\alpha_1 \\ \delta_{Tn}-\gamma_1 \end{pmatrix}^T D^{-1} \begin{pmatrix} \mu_{0Sn}-\alpha_0 \\ \mu_{0Tn}-\gamma_0 \\ \delta_{Sn}-\alpha_1 \\ \delta_{Tn}-\gamma_1 \end{pmatrix}\right\} \\
\propto\; & \exp\left\{-\frac{1}{2}\sum_{j=1}^{r}\left[T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj}-\sigma_{st}\sigma_{ss}^{-1}(S_{nj}-\mu_{0Sn}-\alpha_n Z_{nj})\right]^2 \times q^{-1}\right\} \\
& \times \exp\left\{-\frac{1}{2} MD^T \times (\Psi_{11}-\Psi_{12}\Psi_{22}^{-1}\Psi_{21})^{-1} \times MD\right\} \\
(2.15) = \;& A \times B.
\end{aligned}
$$

where

$$
\Psi_{11} = \begin{pmatrix} d_{tt} & d_{tr} \\ d_{tr} & d_{rr} \end{pmatrix}, \Psi_{12} = \begin{pmatrix} d_{st} & d_{ta} \\ d_{sr} & d_{ar} \end{pmatrix}, \Psi_{21} = \begin{pmatrix} d_{st} & d_{sr} \\ d_{ta} & d_{ar} \end{pmatrix}, \Psi_{22} = \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix},
$$

$$
ME = \begin{pmatrix} S_{nj}-\mu_{0Sn}-\delta_{Sn}Z_{nj} \\ T_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj} \end{pmatrix}, MD = \begin{pmatrix} \mu_{0Tn}-\gamma_0 \\ \delta_{Tn}-\gamma_1 \end{pmatrix} - \Psi_{12}\Psi_{22}^{-1}\begin{pmatrix} \mu_{0Sn}-\alpha_0 \\ \delta_{Sn}-\alpha_1 \end{pmatrix}.
$$

and $q = \sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1}$.

The covariance contribution for μ_{0Tn} and δ_{Tn} from term B is $\Psi_d = \Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$.

We define $Q_{nj} = T_{nj} - \sigma_{st}\sigma_{ss}^{-1}(S_{nj} - \mu_{0Sn} - \alpha_n Z_{nj})$. From (2.15),

$$
\begin{aligned}
A &= \exp\left\{-\frac{1}{2}\sum (Q_{nj}-\mu_{0Tn}-\delta_{Tn}Z_{nj})^2 q^{-1}\right\} \\
&= exp\left\{-\frac{1}{2}\left[\frac{\sum Z_{nj}^2}{q}\delta_{Tn}^2 + \frac{r}{q}\mu_{0Tn}^2 + \frac{\sum Q_{nj}^2}{q} - 2\frac{\mu_{0Tn}\sum Q_{nj}}{q}\right.\right. \\
&\left.\left. -2\frac{\delta_{Tn}\sum Z_{nj}Q_{nj}}{q} + 2\frac{\mu_{0Tn}\delta_{Tn}\sum Z_{nj}}{q}\right]\right\}.
\end{aligned}
$$

A is proportional to a bivariate normal density. The covariance contribution from term A is defined as $\Phi_e = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$, where

$$\phi_{11} = \frac{(\sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1}) \sum_{j=1}^{r} Z_{nj}^2}{r \sum_{j=1}^{r} Z_{nj}^2 - (\sum_{j=1}^{r} Z_{nj})^2},$$
$$\phi_{12} = \frac{(\sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1}) \sum_{j=1}^{r} Z_{nj}}{r \sum_{j=1}^{r} Z_{nj}^2 - (\sum_{j=1}^{r} Z_{nj})^2},$$
$$\phi_{22} = \frac{a(\sigma_{tt} - \sigma_{st}^2\sigma_{ss}^{-1})}{r \sum_{j=1}^{r} Z_{nj}^2 - (\sum_{j=1}^{r} Z_{nj})^2}.$$

Combining the variance contributions from terms A and B, we can obtain the posterior conditional covariance for μ_{0Tn} and δ_{Tn} as: $\left(\Phi_e^{-1} + \Psi_d^{-1}\right)^{-1}$. The corresponding conditional posterior variance for $\hat{\delta}_{Tn} - \delta_{Tn}$ is $\begin{pmatrix} 0 & 1 \end{pmatrix} \left(\Phi_e^{-1} + \Psi_d^{-1}\right)^{-1} \begin{pmatrix} 0 & 1 \end{pmatrix}^T$.

R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
0.5	0.1	40	100	1.007	ALLSE	-0.002	0.111	0.109	94.9
				1.007	ALLHD	-0.002	0.109	0.107	94.1
				1.007	HD	0.005	0.958	0.886	92.8
				1.007	GPHC	0.005	0.957	0.880	92.4
				1.007	BMBRG	0.005	0.961	0.868	92.5
0.5	0.5	40	100	1.007	ALLSE	-0.004	0.109	0.109	95.8
				1.007	ALLHD	-0.004	0.108	0.108	95.3
				1.007	HD	0.005	0.957	0.886	92.6
				1.007	GPHC	0.006	0.957	0.876	92.4
				1.007	BMBRG	0.006	0.961	0.868	92.3
0.5	0.9	40	100	1.007	ALLSE	-0.006	0.109	0.109	94.8
				1.007	ALLHD	-0.005	0.107	0.107	94.8
				1.007	HD	0.006	0.957	0.886	92.4
				1.007	GPHC	0.006	0.957	0.871	92.3
				1.007	BMBRG	0.006	0.960	0.868	92.0
0.5	0.1	40	300	1.007	ALLSE	0.002	0.064	0.063	95.4
				1.007	ALLHD	0.002	0.063	0.063	95.5
				1.007	HD	0.008	0.922	0.881	93.5
				1.007	GPHC	0.008	0.922	0.872	93.5
				1.007	BMBRG	0.008	0.923	0.868	93.3
0.5	0.5	40	300	1.007	ALLSE	0.003	0.063	0.063	94.4
				1.007	ALLHD	0.003	0.062	0.063	94.8
				1.007	HD	0.008	0.922	0.881	93.5
				1.007	GPHC	0.008	0.922	0.871	93.4
				1.007	BMBRG	0.008	0.923	0.868	93.3
0.5	0.9	40	300	1.007	ALLSE	0.002	0.061	0.063	95.6
				1.007	ALLHD	0.002	0.061	0.063	95.6
				1.007	HD	0.008	0.922	0.881	93.5
				1.007	GPHC	0.008	0.922	0.869	93.3
				1.007	BMBRG	0.008	0.923	0.868	93.0

Table 2.1: Impact of R^2_{indiv} and m on $\hat{\delta}_{Tn}$ when T is Completely Missing in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
0.5	0.1	15	100	1.007	ALLSE	-0.002	0.111	0.109	94.9
				1.007	ALLHD	-0.002	0.111	0.107	93.8
				1.007	HD	0.015	1.014	0.840	88
				1.007	GPHC	0.015	1.014	0.818	87.1
				1.007	BMBRG	0.016	1.028	0.800	86.1
0.5	0.5	15	100	1.007	ALLSE	-0.004	0.109	0.109	95.8
				1.007	ALLHD	-0.004	0.109	0.107	94.9
				1.007	HD	0.016	1.013	0.841	88.1
				1.007	GPHC	0.016	1.012	0.814	86.6
				1.007	BMBRG	0.017	1.025	0.802	86
0.5	0.9	15	100	1.007	ALLSE	-0.006	0.109	0.109	94.8
				1.007	ALLHD	-0.005	0.108	0.106	93.4
				1.007	HD	0.016	1.010	0.843	88.5
				1.007	GPHC	0.016	1.010	0.810	86.3
				1.007	BMBRG	0.017	1.023	0.804	86.1
0.5	0.1	55	100	1.007	ALLSE	-0.002	0.111	0.109	94.9
				1.007	ALLHD	-0.002	0.109	0.108	93.6
				1.007	HD	-0.013	0.943	0.894	93.6
				1.007	GPHC	-0.012	0.943	0.891	93.6
				1.007	BMBRG	-0.012	0.945	0.879	93.3
0.5	0.5	55	100	1.007	ALLSE	-0.004	0.109	0.109	95.8
				1.007	ALLHD	-0.004	0.108	0.108	95.4
				1.007	HD	-0.013	0.943	0.894	93.6
				1.007	GPHC	-0.013	0.943	0.887	93.5
				1.007	BMBRG	-0.012	0.945	0.879	93.3
0.5	0.9	55	100	1.007	ALLSE	-0.006	0.109	0.109	94.8
				1.007	ALLHD	-0.005	0.107	0.107	94.5
				1.007	HD	-0.013	0.943	0.894	93.6
				1.007	GPHC	-0.013	0.944	0.882	93.3
				1.007	BMBRG	-0.012	0.946	0.880	93.3

Table 2.2: Impact of R^2_{indiv} and n on $\hat{\delta}_{Tn}$ when T is Completely Missing in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
0.1	0.1	40	100	1.002	ALLSE	-0.002	0.111	0.109	94.9
				1.002	ALLHD	-0.002	0.110	0.108	94
				1.002	HD	0.014	1.260	1.168	93
				1.002	GPHC	0.014	1.260	1.156	92.9
				1.002	BMBRG	0.014	1.263	1.152	92.6
0.1	0.5	40	100	1.002	ALLSE	-0.004	0.109	0.109	95.8
				1.002	ALLHD	-0.004	0.108	0.108	95.4
				1.002	HD	0.014	1.260	1.168	92.9
				1.002	GPHC	0.014	1.260	1.156	92.9
				1.002	BMBRG	0.015	1.263	1.152	92.9
0.1	0.9	40	100	1.002	ALLSE	-0.006	0.109	0.109	94.8
				1.002	ALLHD	-0.005	0.107	0.107	94.1
				1.002	HD	0.014	1.260	1.168	92.8
				1.002	GPHC	0.014	1.260	1.156	92.8
				1.002	BMBRG	0.015	1.263	1.152	92.6
0.8	0.1	40	100	1.010	ALLSE	-0.002	0.111	0.109	94.9
				1.010	ALLHD	-0.002	0.108	0.106	94.1
				1.010	HD	-0.001	0.585	0.538	93.1
				1.010	GPHC	-0.001	0.584	0.541	93.1
				1.010	BMBRG	-0.002	0.588	0.509	91.3
0.8	0.5	40	100	1.010	ALLSE	-0.004	0.109	0.109	95.8
				1.010	ALLHD	-0.004	0.108	0.108	95.4
				1.010	HD	-0.001	0.584	0.538	93.1
				1.010	GPHC	-0.001	0.584	0.531	92.9
				1.010	BMBRG	-0.001	0.587	0.510	91.5
0.8	0.9	40	100	1.010	ALLSE	-0.006	0.109	0.109	94.8
				1.010	ALLHD	-0.005	0.107	0.107	94.8
				1.010	HD	-0.001	0.583	0.539	92.9
				1.010	GPHC	-0.001	0.583	0.519	92.4
				1.010	BMBRG	-0.001	0.586	0.511	91.6

Table 2.3: Impact of R^2_{indiv} and R^2_{trial} on $\hat{\delta}_{Tn}$ when T is Completely Missing in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
0.5	0.1	40	100	1.007	ALLSE	0.002	0.111	0.109	94.9
				1.007	ALLHD	0.002	0.109	0.107	95.4
				1.007	INSE	0.007	0.158	0.154	94.6
				1.007	HD	0.007	0.153	0.149	94.5
0.5	0.5	40	100	1.007	ALLSE	0.004	0.109	0.109	95.8
				1.007	ALLHD	0.004	0.108	0.108	95.3
				1.007	INSE	0.008	0.157	0.154	93.4
				1.007	HD	0.008	0.144	0.141	94.1
0.5	0.9	40	100	1.007	ALLSE	0.006	0.109	0.109	94.8
				1.007	ALLHD	0.005	0.107	0.107	95.1
				1.007	INSE	0.006	0.155	0.154	94.6
				1.007	HD	0.008	0.118	0.117	94.6
0.5	0.1	40	300	1.064	ALLSE	-0.005	0.066	0.063	93.5
				1.064	ALLHD	-0.005	0.066	0.063	93.1
				1.064	INSE	-0.003	0.091	0.089	95.1
				1.064	HD	-0.003	0.090	0.088	94.4
0.5	0.5	40	300	1.064	ALLSE	-0.005	0.064	0.063	93.1
				1.064	ALLHD	-0.005	0.064	0.063	92.5
				1.064	INSE	-0.004	0.089	0.089	94.1
				1.064	HD	-0.003	0.083	0.083	95.1
0.5	0.9	40	300	1.064	ALLSE	-0.003	0.061	0.063	96.7
				1.064	ALLHD	-0.003	0.061	0.063	96.4
				1.064	INSE	-0.003	0.088	0.089	95.4
				1.064	HD	-0.002	0.066	0.069	96.4

Table 2.4: Impact of R^2_{indiv} and m on $\hat{\delta}_{Tn}$ when 50% of T is Missing in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
0.5	0.1	10	100	1.007	ALLSE	0.002	0.111	0.109	94.9
				1.007	ALLHD	0.002	0.111	0.106	94.5
				1.007	INSE	0.007	0.158	0.154	94.6
				1.007	HD	0.006	0.158	0.144	92.1
0.5	0.5	10	100	1.007	ALLSE	0.004	0.109	0.109	95.8
				1.007	ALLHD	0.003	0.109	0.106	95.1
				1.007	INSE	0.008	0.157	0.154	93.4
				1.007	HD	0.007	0.148	0.138	92.8
0.5	0.9	10	100	1.007	ALLSE	0.006	0.109	0.109	94.8
				1.007	ALLHD	0.004	0.109	0.104	93.4
				1.007	INSE	0.006	0.155	0.154	94.6
				1.007	HD	0.006	0.121	0.114	93
0.5	0.1	55	100	1.007	ALLSE	0.002	0.111	0.109	94.9
				1.007	ALLHD	0.002	0.110	0.108	95.3
				1.007	INSE	0.007	0.158	0.154	94.6
				1.007	HD	0.007	0.153	0.149	94
0.5	0.5	55	100	1.007	ALLSE	0.004	0.109	0.109	95.8
				1.007	ALLHD	0.004	0.108	0.108	95.1
				1.007	INSE	0.008	0.157	0.154	93.4
				1.007	HD	0.008	0.144	0.141	94.6
0.5	0.9	55	100	1.007	ALLSE	0.006	0.109	0.109	94.8
				1.007	ALLHD	0.005	0.107	0.107	94.9
				1.007	INSE	0.006	0.155	0.154	94.6
				1.007	HD	0.008	0.118	0.117	94.6

Table 2.5: Impact of R^2_{indiv} and n on $\hat{\delta}_{Tn}$ when 50% of T is Missing in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
0.1	0.1	40	100	1.007	ALLSE	-0.003	0.111	0.109	94.9
				1.007	ALLHD	-0.003	0.110	0.108	95.3
				1.007	INSE	0.002	0.158	0.154	94.6
				1.007	HD	0.002	0.153	0.149	94
0.1	0.5	40	100	1.007	ALLSE	-0.001	0.109	0.109	95.8
				1.007	ALLHD	-0.001	0.108	0.108	95.4
				1.007	INSE	0.002	0.157	0.154	93.4
				1.007	HD	0.003	0.144	0.141	94.4
0.1	0.9	40	100	1.007	ALLSE	0.001	0.109	0.109	94.8
				1.007	ALLHD	0.000	0.107	0.107	95.3
				1.007	INSE	0.001	0.155	0.154	94.6
				1.007	HD	0.002	0.118	0.116	94.6
0.8	0.1	40	100	1.007	ALLSE	0.005	0.111	0.109	94.9
				1.007	ALLHD	0.005	0.108	0.106	95.5
				1.007	INSE	0.010	0.158	0.154	94.6
				1.007	HD	0.010	0.150	0.145	94.1
0.8	0.5	40	100	1.007	ALLSE	0.007	0.109	0.109	95.8
				1.007	ALLHD	0.007	0.108	0.108	95.1
				1.007	INSE	0.010	0.157	0.154	93.4
				1.007	HD	0.011	0.143	0.140	94.1
0.8	0.9	40	100	1.007	ALLSE	0.008	0.109	0.109	94.8
				1.007	ALLHD	0.008	0.107	0.107	95.1
				1.007	INSE	0.008	0.155	0.154	94.6
				1.007	HD	0.010	0.118	0.117	94.8

Table 2.6: Impact of R^2_{indiv} and R^2_{trial} on $\hat{\delta}_{Tn}$ when 50% of T is Missing in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

p	R^2_{trial}	R^2_{indiv}	n	m	True	Methods	Bias	RMSE	SE	CR
100%	0.5	0.1	40	100	1.007	ALLSE	0.002	0.111	0.109	94.9
					1.007	HD	0.002	0.109	0.107	95.4
	0.5	0.5	40	100	1.007	ALLSE	0.004	0.109	0.109	95.8
					1.007	HD	0.004	0.108	0.108	95.3
	0.5	0.9	40	100	1.007	ALLSE	0.006	0.109	0.109	94.8
					1.007	HD	0.005	0.107	0.107	95.1
80%	0.5	0.1	40	100	1.007	INSE	0.005	0.123	0.122	94.5
					1.007	HD	0.005	0.121	0.119	95.4
	0.5	0.5	40	100	1.007	INSE	0.006	0.121	0.122	95
					1.007	HD	0.007	0.116	0.117	95.5
	0.5	0.9	40	100	1.007	INSE	0.005	0.121	0.122	95.4
					1.007	HD	0.007	0.109	0.110	95.4
50%	0.5	0.1	40	100	1.007	INSE	0.007	0.158	0.154	94.6
					1.007	HD	0.007	0.153	0.149	94.5
	0.5	0.5	40	100	1.007	INSE	0.008	0.157	0.154	93.4
					1.007	HD	0.008	0.144	0.141	94.1
	0.5	0.9	40	100	1.007	INSE	0.006	0.155	0.154	94.6
					1.007	HD	0.008	0.118	0.117	94.6
30%	0.5	0.1	40	100	1.007	INSE	0.013	0.202	0.199	94.5
					1.007	HD	0.012	0.190	0.187	94.9
	0.5	0.5	40	100	1.007	INSE	0.012	0.202	0.198	94.6
					1.007	HD	0.013	0.175	0.174	95.3
	0.5	0.9	40	100	1.007	INSE	0.007	0.198	0.198	94.3
					1.007	HD	0.010	0.128	0.129	95.9
10%	0.5	0.1	40	100	1.007	INSE	-0.016	0.351	0.336	91.6
					1.007	HD	-0.013	0.308	0.294	92.5
	0.5	0.5	40	100	1.007	INSE	-0.017	0.350	0.337	92.3
					1.007	HD	-0.010	0.279	0.269	92.8
	0.5	0.9	40	100	1.007	INSE	-0.013	0.347	0.339	92.3
					1.007	HD	-0.002	0.174	0.173	94.5
0%	0.5	0.1	40	100	1.007	INSE	-	-	-	-
					1.007	HD	0.014	0.961	0.886	92.8
	0.5	0.5	40	100	1.007	INSE	-	-	-	-
					1.007	HD	0.014	0.961	0.886	92.5
	0.5	0.9	40	100	1.007	INSE	-	-	-	-
					1.007	HD	0.014	0.960	0.886	92.5

Table 2.7: Impact of R^2_{indiv} and Percentage of Observed T (p) on $\hat{\delta}_{Tn}$ in a Meta-Analytic Setting. Results are from 800 simulation data. $\beta^T = (1, 2, 1, 1)$, $d_{ss} = 0.5$, $d_{tt} = 0.2$, $d_{aa} = 3.5$ and $d_{rr} = 1.6$, $\sigma_{ss} = 1$ and $\sigma_{tt} = 0.3$.

Center	Sample Size	Medicine	Surgery	**Individual-level Correlation**	
		(Means of S, T)	(Means of S, T)	Medicine	Surgery
1	670	(17.63, 16.52)	(13.76, 14.59)	0.367	0.608
2	352	(17.22, 16.42)	(14.63, 12.98)	-0.455	0.467
3	770	(19.27, 17.58)	(15.81, 16.17)	0.589	0.548
4	528	(17.17, 15.51)	(10.93, 12.88)	0.176	0.540
5	1078	(18.52, 18.67)	(14.99, 15.32)	0.435	0.407
6	736	(18.62, 18.89)	(15.13, 17.11)	-0.16	-0.0056
7	572	(18.35, 15.34)	(14.59, 14.53)	0.177	0.396
8	1056	(18.59, 16.16)	(13.60, 13.72)	0.31	0.95
9	638	(17.56, 16.82)	(14.19, 14.61)	0.042	0.756

Table 2.8: Description of Pseudodata in Glaucoma study: Treatment-Specific Means and Individual-Level Correlations for Each Center

p	Estimate	Standard Error	p-value
$center = 8$			
ALLSE	-2.45	0.29	$< .0001$
No missing[1]	-2.33	0.22	$< .0001$
100% missing[1]	-1.58	0.79	0.063
90% missing[1]	-1.50	0.47	0.0059
80% missing[1]	-2.37	0.39	$< .0001$
50% missing[1]	-2.61	0.29	$< .0001$
20% missing[1]	-2.19	0.23	$< .0001$
$center = 9$			
ALLSE	-2.21	0.30	$< .0001$
No missing[1]	-2.32	0.27	$< .0001$
100% missing[1]	-2.68	0.82	0.0053
90% missing[1]	-2.19	0.61	0.0023
80% missing[1]	-2.30	0.49	$< .0002$
50% missing[1]	-2.04	0.36	$< .0001$
20% missing[1]	-2.15	0.30	$< .0001$

Table 2.9: Estimate treatment effect on IOP at the 96th month utilizing information from early IOP measures at the $12th$ month in the glaucoma study. [1]: HD method was used.

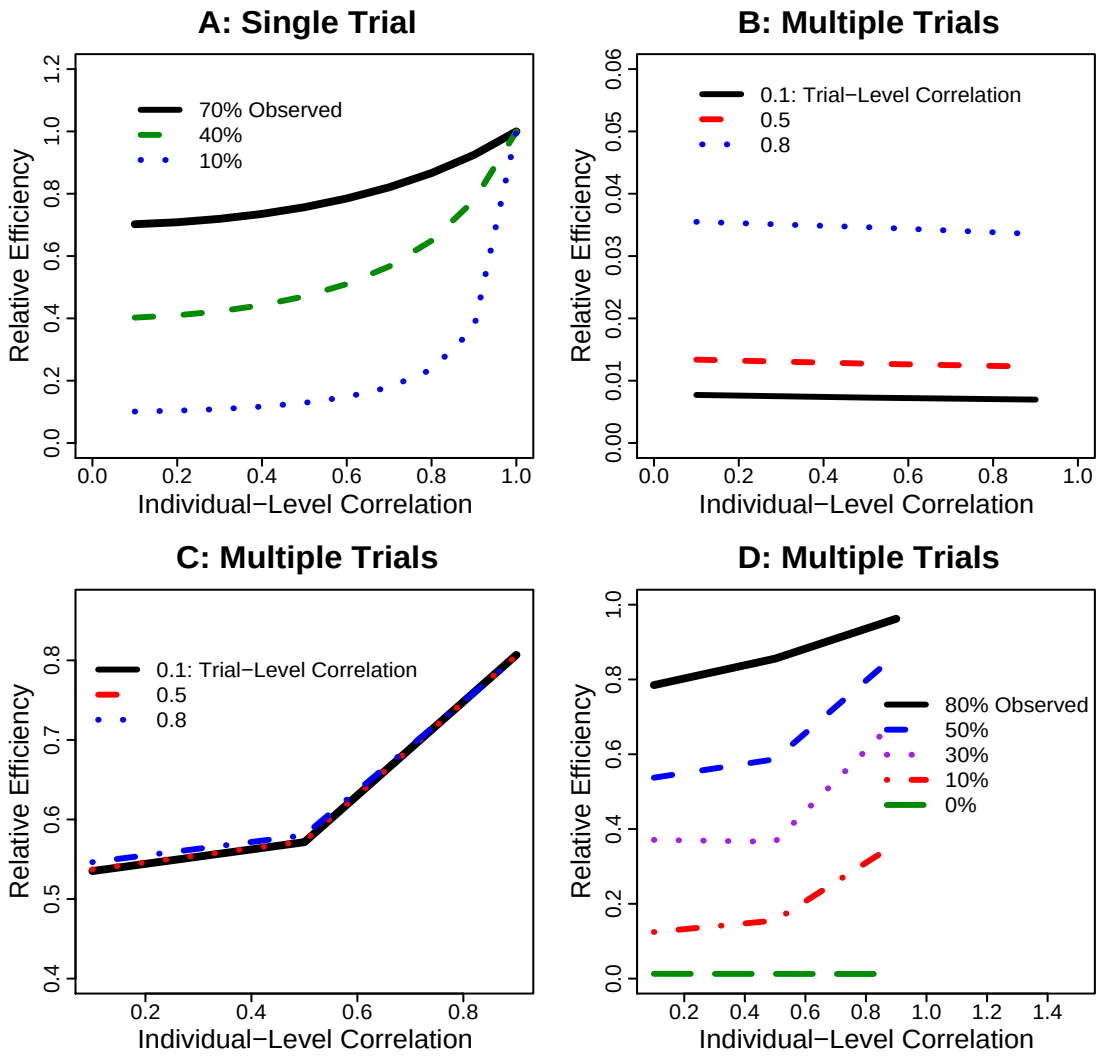

Figure 2.1: Relative efficiency of the new treatment effect estimate using S when T is not completely observed to that when T is completely observed. A: Single-Trial Setting; B: Multiple-trial setting. T is 100% missing in the new trial; C: Multiple-trial setting. T is 50% observed in the new trial; D: Multiple-trial setting. Percentage of Observed T Varies in the new trial.

2.6 References

Baker S.G., Kramer B.S. (2003). A perfect correlate does not a surrogate make. *BMC Medical Research Methodology*. **3**: 16

Begg B.B., and Leung H.Y. (1999). On the use of surrogate end points in randomized trials. *Journal Royal Statistical Society A*. **163**, 15-28.

Buyse M, Molenberghs G (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*. **54**: 1014-1029.

Buyse M., Molenberghs G., Burzykowski T., Renard D. and Geys H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. **1**, 49-67.

Bycott P.W., Taylor J.M.G. (1998). An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker. *Controlled Clinical Trials*. **19**: 555-568.

Cook R.J., Lawless J.F. (2001). Some comments on efficiency gains from auxiliary information for right-censored data. *Journal of Statistical Planning and Inference*. **96**, 191-202.

De Gruttola V, Fleming T, Lin D.Y., Coombs R. (1997). Perspective: validating surrogate markers - are we being nave? *The Journal of Infectious Diseases*. **175**, 237-246.

Fleming T.R., DeMets D.L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*. **125**, 605-613.

Freedman L.S., Graubard B.I., Schatzkin A (1992). Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine*. **11**, 167-178.

Gail M., Pfeiffer R., Houwelingen H.C.V., and Carroll R.J (2000). On Meta-analytic assessment of surrogate outcomes. *Biostatistics*. **1**, 231-246.

Ghosh M., Rao N.K. (1994). Small area estimation: an appraisal. *Statistical Science*. **9**: 55-76.

Henderson C.R. (1984). Applications of linear models in animal breeding. *University of Guelph*

Hsu, C., Taylor, J.M.G., Murray, S. and Commenges, D. (2006). Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine*. **25**: 3503-3517.

Laird N.M., Lang, N. and Stram, D. (1982). Random-effects models for longitudinal data. *Biometrics*. **38**, 963-974.

Little R.J.A and Rubin D.B. (2002). Statistical Analysis with Missing Data, 2nd Edition. Wiley: New York.

Lin D.Y., Fleming T.R., DeGruttola V. (1997). Estimating the proportion of treatment effect captured by a surrogate marker. *Statistics in Medicine*. **16**, 1515-1527.

McLean R.A. Sanders W.L. and Stroup W.W. (1991). A unified approach to mixed linear models. *The American Statistician*. **45**, 54-64.

Morris C. (1983). Parametric empirical Bayes inference: theory and application (with discussions). *Journal of American Statistical Association*. **78**, 47-65.

Murray, S and Tsiatis, A. A. (1996), Nonparametric Survival Estimation Using Prognostic Longitudinal Covariates, *Biometrics*, **52**: 137-151.

Musch D.C., Lichter P.R., Guire K.E., Standardi C.L., CIGTS Investigators (1999): The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, meth-

ods, and baseline characteristics of enrolled patients. *Ophthalmology*. **106**: 653-62.

Pepe M.S., Reilly, M and Fleming, T.R. (1994). Auxiliary outcome and the mean score method. *J. Statist. Planning Inf.* **43**: 137-160.

Prentice R.L. (1989). Surrogate endpoints in clinical trials, definition and operational criteria. *Statistics in Medicine*. **8**, 431-440.

Robins, J. M. and Rotnitzky, A. (1992), Recovery of Information and Adjustment for Dependent Censoring using Surrogate Markers, *AIDS Epidemiology: Methodological Issues, Ed. N. Jewell, K. Dietz and V. Farewell*, Boston: Birkhauser, 297-331.

Robinson G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*. **6**, 15-51.

Venkatraman E.S., Begg C.B. (1999). Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics*. **55**: 1171-1176.

Wang Y., Taylor J.M.G. (2003). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*. **58**, 803-812.

CHAPTER III

A Shrinkage Approach for Estimating a Treatment Effect Using Surrogate Marker Data in Clinical Trials

Summary. Surrogate markers (S) are often intermediate physical or laboratory indicators in a disease process in randomized clinical trials. They can be measured earlier and often cost much less than the true endpoint (T). A surrogate marker that is strongly associated with the true endpoint can provide information on the true endpoint. We examine the information recovery from S in estimating the difference between two treatment groups when S is completely observed and T is partially observed. Both S and T are continuous. When S satisfies Prentice's definition for perfect surrogacy, there is substantial gain in precision by using S to estimate the treatment effect. When S is not close to having perfect surrogacy, it can provide substantial information only under special circumstances. We propose a generalized ridge regression to avoid the need to make a correct surrogacy assumption. Simulations show that it can strike a balance between bias and efficiency gain depending on the nature of the relationship between S and T. Compared with competing methods, it has better mean squared error properties and can achieve substantial efficiency gain, particularly in small samples when S is close to being a perfect surrogate. We apply the proposed method to a glaucoma data example.

Keywords: Surrogate Marker, Auxiliary Variable, Ridge Regression, and Ran-

domized Trials.

3.1 Introduction

In clinical trials where the true endpoint (T) is rare or late-occurring, it is often very costly and takes too long to observe T for all the subjects. A surrogate marker (S) is an intermediate physical or laboratory indicator in a disease process, which can be measured earlier and easier than the true endpoint. A surrogate marker that is highly associated with the true endpoint can provide information on the effect of treatment on T when T is not completely observed. There are two potential roles of S: one as a direct substitute for T and the other as an auxiliary variable. In this article, we focus on its latter role in improving the efficiency of the estimated treatment effect on T. In practice, it often happens that S is available on all patients before T is available. Intuitively, incorporating the information from S in estimating the actual effect of Z on T should lead to narrower confidence intervals and more powerful tests. In this article, we use the terms, primary endpoint and true endpoint exchangeably.

A number of authors have explored the role of surrogate markers as auxiliary variables, mostly focusing on situations where the primary endpoint is time to failure. Many articles focus on parametric or semiparameteric modeling of the joint distribution of S and T(Pepe et al, 1992, Pepe et al., 1994, Malani, 1995, Murray and Tsiatis, 1996). When the subsample of patients for whom both S and T are observed represents the whole population, Fleming *et al* (1994) and Kosorok and Fleming (1993) proposed fully nonparametric tests that incorporate the information from S to enhance the inference. In the absence of censoring, Venkatraman and Begg (1999) proposed a fully nonparametric test to make use of the intermediate surrogate markers. The extent of efficiency gain has been the topic of the previous research,

and opinions have been mixed on the value of a surrogate marker in increasing the precision of the estimate (Murray and Tsiatis, 1996; Venkatraman and Begg, 1999; Fleming and DeMets, 1996). In many simulations, there has been little or no gain from the use of the surrogate marker; however, in rare cases when S and T have an extremely high correlation, S can provide significant additional information for the quantity of interest.

In this article, we focus on the single trial setting where T is partially observed, S and Z are measured on everyone. Both S and T are continuous and Z is binary. We assume a parametric model for the joint distribution of S and T given Z, through modeling $f(T|S,Z)$ and $f(S|Z)$ jointly, which allows more flexibility than the commonly used multivariate normal distribution for $f(S,T|Z)$. The full model for the distribution of $T|S,Z$ is $T=\beta_0+\beta_1 S+\beta_2 Z+\beta_3 SZ+\epsilon$, where ϵ is the error term. In a landmark study, Prentice (1989) proposed a formal definition for perfect surrogacy (PES) and provided validation criteria. The key criterion requires that the changes in S fully capture the effect of Z on T. In other words, $\beta_2=\beta_3=0$ assuming $\beta_1\neq 0$. When $\beta_3\neq 0$ or $\beta_2\neq 0$, S explains some, but not all, of the association between T and Z; and S is a partial surrogate. Specifically, when $\beta_2\neq 0$ and $\beta_3=0$, S is an additive partial surrogate (APAS); when $\beta_3\neq 0$, S is an interactive partial surrogate (IPAS). We can estimate the effect of Z on T under different surrogacy assumptions and different forms of the joint distribution $f(S,T|Z)$.

Our numerical studies show that different surrogacy assumptions can lead to very different degree of efficiency gain. When the assumptions are incorrect, substantial bias can occur in the estimated treatment effect. In practice, when one is unsure whether S satisfies PES, APAS or IPAS assumption, model selection methods are often used to choose the correct assumption. However, it is difficult to formally

account for the model uncertainty in the inference. The common practice of ignoring such uncertainty can lead to high type I errors (Albert et al, 2001). In addition, different sample sizes may lead to different conclusions on the nature of the surrogacy. Moreover, for any given study the power for testing the assumption is limited. In fact, even when the data do not contradict the assumption that S is a perfect surrogate, we still cannot conclude that the assumption is true.

From a biological point of view, there are often multiple pathways through which the treatment can affect T and it is seldom that a marker can capture all the effect on T. In the past two decades, the quest for perfect surrogates has been less than successful. An incorrect perfect surrogacy assumption can lead to erroneous conclusions (Fleming and DeMets, 1996). For example, the inappropriate use of arrhythmia as a surrogate for cardiac death led to the drug approval by FDA in the 1980s and caused much harm and even death among patients. However, it is very plausible for a partial surrogate to capture most but not all of the treatment effect. Examples include time to disease progression as a surrogate marker for the survival time in early stage cancer; prostate-specific antigen for prostate cancer; and the intraocular pressure for the long-term visual acuity. A model selection approach that either retains or discards a variable may be inappropriate and, as we shall see through simulations, can lead to substantial prediction error.

In Sections 3.2 and 3.3, we conduct analytic and numerical studies under these three surrogacy assumptions to explore the efficiency gain from S under the various surrogacy assumptions. In section 3.4, we propose a generalized ridge regression model to utilize the information in S. Ridge regression models have been commonly used in high-dimensional data to reduce collinearity, but their application in incorporating the auxiliary variable S in treatment prediction is novel. This shrinkage

approach can allow for the uncertainty associated with the surrogacy nature and avoid the need to make a correct surrogacy assumption. It compromises between the perfect surrogacy and partial surrogacy models and strikes a data-driven balance between bias and variance. We first introduce a hierarchical Bayes version of the generalized ridge regression model and then an empirical Bayes version as an alternative. In Section 3.5, we conduct simulation studies to examine the properties of this approach in terms of bias, efficiency gain, mean squared errors and coverage rates of confidence intervals. We compare this method with competing methods such as the model selection method and an inverse probability weighted method. In Section 3.6, we apply the proposed methods to a glaucoma data set. In Section 3.7, we summarize our findings and present a discussion.

3.2 Treatment Effect Estimation and Surrogacy Assumptions

Our goal is to estimate the treatment effect (denoted by Q) on T in the setting where T is partially available and S is completely observed. In this section, we investigate the estimates of the treatment effect using S under three different models that describe a perfect surrogate, an additive partial surrogate and an interactive partial surrogate, respectively.

Suppose that the total number of patients is $n = n_0+n_1$ with n_0 in the $Z = 0$ group and n_1 in the $Z = 1$ group. To simplify the following calculations, we assume that the clinical trial is a balanced trial; i.e., $E(Z) = 0.5$. Nonetheless, the conclusions are generalizable without this assumption. The surrogate, S, is measured on all n patients; T is available for a subset of $r = r_0+r_1$ patients, with r_0 in the $Z = 0$ group and r_1 in the $Z = 1$ group. The fraction of the subjects for whom T is not observed is p. We assume the missingness mechanism to be missing at random (MAR) (Little

and Rubin, 2002), meaning that the missingness probability may depend on observed data.

3.2.1 Interactive Partial Surrogate

We call S is an interactive partial surrogate (IPAS), when the joint distribution $f(T_i, S_i|Z_i)$ for individual i can be expressed using two models:

$$\begin{aligned} T_i &= \beta_0 + \beta_1 S_i + \beta_2 Z_i + \beta_3 S_i Z_i + \epsilon_{ti} \\ S_i &= \alpha_0 + \alpha_1 Z_i + \epsilon_{si} \end{aligned} \tag{3.1}$$

where $\epsilon_{ti} \sim N(0, \sigma_t^2)$ and $\epsilon_{si} \sim N(0, \sigma_{ss}^2)$. Under this surrogacy assumption, the marginal treatment effect has mean

$$\begin{aligned} E(\hat{Q}_{IPAS}) &= E(T|Z=1) - E(T|Z=0) = EE(T|S, Z=1) - EE(T|S, Z=0) \\ &= \beta_0 + \beta_1(\alpha_0 + \alpha_1) + \beta_2 + \beta_3(\alpha_0 + \alpha_1) - (\beta_0 + \beta_1\alpha_0) \\ &= \beta_1\alpha_1 + \beta_2 + \beta_3\alpha_0 + \beta_3\alpha_1. \end{aligned}$$

The likelihood of $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \alpha_0, \alpha_1)$ based on the observed data is given by:

$$\begin{aligned} L(\theta|S, T, Z) &= \prod_{i=1}^{r} f(T_i|S_i, Z_i, \theta) \prod_{i=1}^{n} f(S_i|Z_i, \theta) \\ &= \prod_{i=1}^{r} \frac{1}{\sqrt{2\pi}\sigma_t} \exp^{(T_i - \beta_0 - \beta_1 S_i - \beta_2 Z_i - \beta_3 S_i Z_i)^2/2\sigma_t^2} \\ &\quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_s} \exp^{-(S_i - \alpha_0 - \alpha_1 Z_i)^2/2\sigma_{ss}^2}, \end{aligned} \tag{3.2}$$

from which, we can obtain the large-sample covariance matrix of $\hat{\theta}$ by calculating the inverse of the expected information matrix $I_{IPAS}(\theta)$ (see Appendix A). Let $D_{IPAS}(Q) = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \beta_2}, \frac{\partial Q}{\partial \beta_3}, \frac{\partial Q}{\partial \alpha_0}, \frac{\partial Q}{\partial \alpha_1}) = (0, \alpha_1, 1, \alpha_0 + \alpha_1, \beta_3, \beta_1 + \beta_3)$. The asymptotic variance of $\hat{Q}_{IPAS}$ can be calculated using the delta method as

$$V(\hat{Q}_{IPAS}) = D_{IPAS}(Q)^T I_{IPAS}(\theta)^{-1} D_{IPAS}(Q).$$

Its estimate $\hat{V}(\hat{Q}_{IPAS})$ can be obtained by replacing θ with the maximum likelihood estimates $\hat{\theta}$ and replacing $I_{IPAS}(\theta)$ with the observed information matrix.

To gain intuition on the factors that impact the variance, we use a different form of this variance formula. We denote the correlation between S and T in the $Z = 0$ group as ρ_0 and that in the $Z = 1$ group as ρ_1. After some calculation, we obtain $\rho_0^2 = \frac{\beta_1^2\sigma_{ss}^2}{\sigma_t^2+\beta_1^2\sigma_{ss}^2}$ and $\rho_1^2 = \frac{(\beta_1+\beta_3)^2\sigma_{ss}^2}{\sigma_t^2+(\beta_1+\beta_3)^2\sigma_{ss}^2}$. Denote $\mathrm{V}(T|Z=0) = \sigma_{tt0}^2$ and $\mathrm{V}(T|Z=1) = \sigma_{tt1}^2$. We have $\sigma_{tt0}^2 = \sigma_t^2/(1-\rho_0^2)$ and $\sigma_{tt1}^2 = \sigma_t^2/(1-\rho_1^2)$. Under the missing completely at random assumption, the large-sample variance of $\hat{Q}_{IPAS}$ can also be approximated by

$$\frac{\sigma_{tt0}^2}{r_0}(1-\rho_0^2\frac{n_0-r_0}{n_0}) + \frac{\sigma_{tt1}^2}{r_1}(1-\rho_1^2\frac{n_1-r_1}{n_1}). \tag{3.3}$$

Suppose we observe all of T, or the data before any missingness occurs. Without any distributional assumption, the estimated treatment effect $\hat{Q}_{ALL} = \sum_{i=1}^{n_1} T_i/n_1 - \sum_{i=1}^{n_0} T_i/n_0$ with variance $V(\hat{Q}_{ALL}) = \sigma_{tt0}^2/n_0 + \sigma_{tt1}^2/n_1$. When T is partially observed and the fraction of missingness is p, the treatment effect estimated solely based on the observed T is $\hat{Q}_{CC} = \sum_{i=1}^{r_1} T_i/r_1 - \sum_{i=1}^{r_0} T_i/r_0$ and its variance $V(\hat{Q}_{cc}) = \sigma_{tt0}^2/(n_0(1-p)) + \sigma_{tt1}^2/(n_1(1-p))$. Comparing $V(\hat{Q}_{cc})$ with the variance formula in (3.3), we find that when the percentage of missingness and the variance of $T|Z$ are known, the single most important factor that impacts the precision of the estimate and the extent of the information recovery from S is the treatment-adjusted correlation between S and T.

3.2.2 Additive Partial Surrogate

When S is an additive partial surrogate (APAS), the association between S and T is the same regardless of Z, (i.e., $\beta_3 = 0$). The joint distribution $f(T_i, S_i|Z_i)$ is

expressed as:

$$\begin{aligned} T_i &= \beta_0 + \beta_1 S_i + \beta_2 Z_i + \epsilon_{ti} \\ S_i &= \alpha_0 + \alpha_1 Z_i + \epsilon_{si} \end{aligned}$$

where $\epsilon_{ti} \sim N(0, \sigma_t^2)$ and $\epsilon_{si} \sim N(0, \sigma_{ss}^2)$. This model is equivalent to the commonly used bivariate normal distribution of S_i and T_i:

$$\begin{pmatrix} T_i \\ S_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} \gamma_0 + \gamma_1 Z_i \\ \alpha_0 + \alpha_1 Z_i \end{pmatrix}, \begin{pmatrix} \sigma_{tt}^2 & \sigma_{st}^2 \\ \sigma_{st}^2 & \sigma_{ss}^2 \end{pmatrix} \right).$$

With this model assumption, $\rho_0 = \rho_1 = \rho$, $\rho^2 = \frac{\sigma_{st}^2}{\sigma_{ss}^2 \sigma_{tt}^2} = \frac{\beta_1^2 \sigma_{ss}^2}{\sigma_t^2 + \beta_1^2 \sigma_{ss}^2}$, while $\mathrm{V}(T|Z) = \sigma_{tt}^2 = \sigma_t^2/(1-\rho^2)$.

The estimated treatment effect on T has mean

$$E(\hat{Q}_{APAS}) = \beta_2 + \beta_1 \alpha_1.$$

The expected information matrix $I_{APAS}(\theta)$ under this assumption is computed as in Appendix A. We can use the delta method to compute the asymptotic variance of $\hat{Q}_{APAS}$. Denote $D_{APAS}(Q) = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \beta_2}, \frac{\partial Q}{\partial \alpha_0}, \frac{\partial Q}{\partial \alpha_1}) = (0, \alpha_1, 1, 0, \beta_1)$. The large-sample variance is given by $V(\hat{Q}_{APAS}) = D_{APAS}(Q)^T I_{APAS}(\theta)^{-1} D_{APAS}(Q)$. Its estimate $\hat{V}(\hat{Q}_{APAS})$ can be obtained by replacing θ with the maximum likelihood estimates and $I_{APAS}(\theta)$ with the observed information matrix. The large-sample variance can also approximated by

$$\frac{\sigma_{tt}^2}{r_0}(1 - \rho^2 \frac{n_0 - r_0}{n_0}) + \frac{\sigma_{tt}^2}{r_1}(1 - \rho^2 \frac{n_1 - r_1}{n_1}). \tag{3.4}$$

As under the interactive partial surrogacy assumption, when σ_{tt}^2 and the percentage of missingness are fixed, ρ^2 is the single most important factor that determines the precision of the treatment effect estimate and the extent of efficiency gain from S.

3.2.3 Perfect Surrogate

When S is a perfect surrogate (PES), S captures all of the treatment effect on T (Prentice, 1989). Under the PES assumption, the treatment effect on T disappears after one adjusts for S, i.e., $\beta_2 = \beta_3 = 0$ in model (3.1). In this case, the marginal treatment effect on T has mean

$$E(\hat{Q}_{PES}) = \beta_1\alpha_1.$$

The details of the expected information matrix, $I_{PES}(\theta)$, are described in Appendix A. Denote $D_{PES}(Q) = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \alpha_0}, \frac{\partial Q}{\partial \alpha_1}) = (0, \alpha_1, 0, \beta_1)$. The variance of $\hat{Q}_{PES}$ is $V(\hat{Q}_{PES}) = D_{PES}(Q)^T I_{PES}(\theta)^{-1} D_{PES}(Q)$. Under this assumption, the correlation between S and T is $\rho^2 = \frac{\beta_1^2\sigma_{ss}^2}{\sigma_t^2+\beta_1^2\sigma_{ss}^2}$, while $\mathrm{V}(T_i|Z_i) = \sigma_{tt}^2 = \sigma_t^2/(1-\rho^2)$. The variance can be approximated by

$$\frac{\alpha_1\sigma_{tt}^2(1-\rho^2)}{r\sigma_{ss}^2} + \frac{4\sigma_{st}^2}{n}. \tag{3.5}$$

Under the perfect surrogacy assumption, when σ_{tt}^2 and the percentage of missingness are fixed, the factors that impact the variance of the treatment effect estimate and the extent of efficiency gain from S include not only the correlation and the factors associated with the correlation, but also α_1.

3.3 Numerical Study on Information Recovery and Surrogacy Assumptions

We conduct numerical studies based on the asymptotic variances to examine the efficiency gain from S under different surrogacy assumptions. In the studies, the variances of the treatment effect on T are calculated in five scenarios:

1. $V(\hat{Q}_{ALL})$ which uses all data or the original data on T before any missingness occurs.

2. $V(\hat{Q}_{CC})$ which uses data of only complete cases, or observed T only.
3. $V(\hat{Q}_{IPAS})$ when the fitted model is the IPAS model, i.e., the saturated model.
4. $V(\hat{Q}_{APAS})$ when the fitted model assumes the APAS model, i.e., $\beta_3 = 0$.
5. $V(\hat{Q}_{PES})$ when the fitted model assumes PES, i.e., $\beta_2 = \beta_3 = 0$.

We calculate the asymptotic variances in the scenarios when the true model is either PES, APAS or IPAS model. The results in the scenario when the true model is the PES model and the fitted models assume either IPAS, APAS or PES are presented in Tables 1 and 2 and Figure 1. To illustrate the extent of efficiency gain from S, we list the relative efficiency defined by the ratios of the variance of $V(\hat{Q}_{ALL})$ to those from 2 through 5.

Based on the numerical calculations, we find that generally there is some improvement in the efficiency of the estimate of Q by incorporating S. The factors that impact the efficiency gain depend on which fitted model is used and are consistent with the variance formulae in (3.3), (3.4) or (3.5). When we assume IPAS or APAS, the higher the correlation between S and T, the higher the extent of efficiency gain from S. When we assume PES, the higher the correlation and the bigger the value of α_1, the higher the amount of information recovery from S.

The extent of efficiency gain from S also highly depends on which surrogacy assumption holds. In large samples, the amount of information recovery is almost the same whether the fitted model assumes IPAS or APAS, i.e., assuming $\beta_3 = 0$ does not result in any obvious efficiency gain over $\beta_3 \neq 0$. Under either IPAS or APAS assumption, the extent of efficiency gain is only modest except for the cases when the correlation is unusually high. However, when we assume PES in the fitted model, we can uniformly improve the efficiency gain to a much greater extent and can even result in smaller variances than $V(\hat{Q}_{ALL})$ in some cases (see Tables 3.1, 3.2

and Figure 3.1). This happens because our estimate utilizes the information that S completely captures the effect on T while the estimate obtained based on T solely does not take advantage of such an assumption. In other words, if Prentice's criteria hold, we can achieve substantial efficiency gain by using the information in S. This phenomenon was also observed by Day and Duffy (1996), Begg and Leung (2000) and Baker *et al.* (2000). When the PES assumption is incorrectly assumed, the estimates of the treatment effects can be substantially biased.

Thus, the surrogacy assumption plays a central role in the extent of efficiency gain that we could obtain from S. As mentioned in Section 3.1, a possible approach is to use a model selection method to choose which assumption to make. However, as we shall see in simulations, the method is subject to the restriction of sample size, power and difficulty in accounting for model uncertainty. In the next section, we propose a shrinkage approach that avoids the need to make a correct surrogacy assumption.

3.4 Generalized Ridge Regression

In this section, we propose a generalized ridge regression model (denoted by Ridge) to utilize S to estimate the treatment effect. The general idea is that, although S is rarely a perfect surrogate, in reality, it is common that S can capture a large portion of the treatment effect on T when S is considered as a good surrogate marker. We first consider the situation when $\beta_3 = 0$ in model 3.1. A reasonable assumption is that β_2 is close to but not necessarily be exactly equal to 0. We impose a prior distribution on β_2 such that $\beta_2 \sim N(0, \sigma^2_{b_2})$, where $\sigma^2_{b_2}$ is used to capture the uncertainty about the departure from the perfect surrogacy assumption. By assuming this prior distribution, the generalized ridge regression model induces a shrinkage effect on $\hat{\beta}_2$, which will data-adaptively shrink $\hat{\beta}_2$ towards 0 when S is close to being perfect.

Next, we introduce two versions of the generalized ridge regression method. The first is the full Bayes version, where we treat $\sigma^2_{b_2}$ as a hyper-parameter with its own prior distribution; the second is the empirical Bayes version, where $\sigma^2_{b_2}$ is estimated directly from the data.

3.4.1 Fully Bayes Estimator

When $\beta_3 = 0$, the Bayes version of the generalized ridge regression model is expressed as follows:

$$\begin{aligned} T_i &= \beta_0 + \beta_1 S_i + \beta_2 Z_i + \epsilon_{ti} \\ S_i &= \alpha_0 + \alpha_1 Z_i + \epsilon_{si} \end{aligned}$$

where

$$\begin{aligned} &\epsilon_{ti} \sim N(0, \sigma_t^2), \quad \epsilon_{si} \sim N(0, \sigma_{ss}^2), \quad \beta_0 \sim N(0, a = 100^2), \quad \beta_1 \sim N(0, a = 100^2), \\ &\beta_2 \sim N(0, \sigma_{b2}^2), \quad \sigma_t^{-2} \sim \text{Gamma}(c, d), \quad \sigma_{b2}^{-2} \sim \text{Gamma}(c, d), \\ &\alpha_0 \sim N(0, a = 100^2), \quad \alpha_1 \sim N(0, a = 100^2), \quad \sigma_{ss}^{-2} \sim \text{Gamma}(c, d). \end{aligned}$$

Note that the parametrization of Gamma(c, d) is defined such that the expected value equals cd and the variance is cd^2. We choose nearly noninformative values of $c = 0.001$ and $d = 1000$. The posterior distributions of the parameters are proportional to *Likelihood* $\times$ *Prior* given by

$$\left[\prod_{i=1}^{r} N(T_i|\beta_0, \beta_1, \beta_2, \sigma_t^2)\right] N(\beta_0|a) N(\beta_1|a) N(\beta_2|\sigma_{b2}^2) \text{Inv-Gamma}(\sigma_{b2}^2|c, d)$$
$$\text{Inv-Gamma}(\sigma_t^2|c, d) \left[\prod_{i=1}^{n} N(S_i|\alpha_0, \alpha_1, \sigma_{ss}^2)\right] N(\alpha_0|a) N(\alpha_1|a) \text{Inv-Gamma}(\sigma_{ss}^2|c, d)$$

We can use Gibbs sampling to make draws from the following conditional posterior

distributions

$$\sigma_t^{-2}|\cdot \sim \text{gamma}\left(\frac{r}{2}+c, \frac{\sum_{i=1}^{r}(t_i-\beta_0-\beta_1 s_i-\beta_2 z_i)^2+2d}{2}\right)$$

$$\sigma_{b_2}^{-2}|\cdot \sim \text{gamma}\left(0.5+c, \frac{\beta_2^2+2d}{2}\right),$$

$$\beta_0|\cdot \sim \text{normal}\left(\left(\frac{r}{\sigma_t^2}+\frac{1}{a^2}\right)^{-1}\frac{\sum_{i=1}^{r}(t_i-\beta_1 s_i-\beta_2 z_i)}{\sigma_t^2}, \left(\frac{r}{\sigma_t^2}+\frac{1}{a^2}\right)^{-1}\right),$$

$$\beta_1|\cdot \sim \text{normal}\left(\left(\frac{\sum_{i=1}^{r}s_i^2}{\sigma_t^2}+\frac{1}{a^2}\right)^{-1}\frac{\sum_{i=1}^{r}s_i(t_i-\beta_0-\beta_2 z_i)}{\sigma_t^2}, \left(\frac{\sum_{i=1}^{r}s_i^2}{\sigma_t^2}+\frac{1}{a^2}\right)^{-1}\right),$$

$$\beta_2|\cdot \sim \text{normal}\left(\left(\frac{\sum_{i=1}^{r}z_i^2}{\sigma_t^2}+\frac{1}{\sigma_{b_2}^2}\right)^{-1}\frac{\sum_{i=1}^{r}z_i(t_i-\beta_0-\beta_1 s_i)}{\sigma_t^2}, \left(\frac{\sum_{i=1}^{r}z_i^2}{\sigma_t^2}+\frac{1}{\sigma_{b_2}^2}\right)^{-1}\right),$$

$$\sigma_{ss}^{-2}|\cdot \sim \text{gamma}\left(\frac{n}{2}+c, \frac{\sum_{i=1}^{n}(s_i-\alpha_0-\alpha_1 z_i)^2+2d}{2}\right),$$

$$\alpha_0|\cdot \sim \text{normal}\left(\left(\frac{n}{\sigma_{ss}^2}+\frac{1}{a^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n}(s_i-\alpha_1 z_i)}{\sigma_{ss}^2}\right), \left(\frac{n}{\sigma_{ss}^2}+\frac{1}{a^2}\right)^{-1}\right),$$

$$\alpha_1|\cdot \sim \text{normal}\left(\left(\frac{\sum_{i=1}^{n}z_i^2}{\sigma_{ss}^2}+\frac{1}{a^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n}(s_i-\alpha_0)z_i}{\sigma_{ss}^2}\right), \left(\frac{\sum_{i=1}^{n}z_i^2}{\sigma_{ss}^2}+\frac{1}{a^2}\right)^{-1}\right),$$

where $\cdot$ represents the rest of the parameters and observed data. Based on the posterior distributions of these parameters, we can easily obtain the posterior distribution of the treatment effect estimate, $\hat{Q}_{Ridge-FB} = \hat{\beta}_1\hat{\alpha}_1 + \hat{\beta}_2$.

We can extend the ridge regression model to the situation when $\beta_3 \neq 0$. We assume the prior distribution for β_2 is $N(0, \sigma_{b_2}^2)$ and that for β_3 is $N(0, \sigma_{b_3}^2)$, with which $\sigma_{b_2}^2$ captures the uncertainty of the departure from $\beta_2 = 0$ and $\sigma_{b_3}^2$ captures the departure from $\beta_3 = 0$. The rest of the computation procedures follows very similarly to those just described.

3.4.2 Empirical Bayes Estimator

The advantage of fully Bayes estimation is that it accounts for all the uncertainty associated with estimating any parameter (Louis and Zelterman (1994)). However, it is computationally intensive, particularly when the sample size is large. In this

section, we consider an empirical Bayes version of the generalized ridge regression as an alternative which is computationally faster.

First, we consider the situation when $\beta_3 = 0$, and the model $T|S, Z$ is given by:

$$T_i = \beta_0 + \beta_1 S_i + \beta_2 Z_i + \epsilon_{ti},$$

where $\epsilon_{ti} \sim N(0, \sigma_t^2)$ and $\beta_2 \sim N(0, \sigma_{b2}^2)$. Let $\beta^T = (\beta_0, \beta_1, \beta_2)$, $X_t = (1, S, Z)$, $K = \text{diag}(0, 0, k_2)$ where $k_2 = \sigma_t^2/\sigma_{b_2}^2$. Suppose σ_{b2}^2 and σ_t^2 are known and noninformative prior distributions are assumed for β_0 and β_1. The posterior distribution of β follows a normal distribution with mean and variance expressed by:

$$\text{E}(\hat{\beta}|X_t, T) = (X_t^T X_t + K)^{-1} X_t^T T,$$
$$\text{V}(\hat{\beta}|X_t, T) = (X_t^T X_t + K)^{-1} \sigma_t^2.$$

The idea of the empirical Bayes approach is to use the data to estimate σ_{b2}^2 and σ_t^2. First, we want to find an estimate of σ_{b2}^2. Given β_2, $\hat{\beta}_2 \sim N(\beta_2, \sigma_{\beta_2}^2)$. We can obtain the joint distribution of $(\hat{\beta}_2, \beta_2)$ by multiplying the densities of $\hat{\beta}_2|\beta_2$ and β_2 together, then obtaining the marginal density of $\hat{\beta}_2$ as $N(0, \sigma_{\beta_2}^2 + \sigma_{b2}^2)$. The quantity $\sigma_{\beta_2}^2$ can be estimated from the maximum likelihood fit to $T_i = \beta_0 + \beta_1 S_i + \beta_2 Z_i + \epsilon_{ti}$. Since $E(\hat{\beta}_2) = 0$, $E((\hat{\beta}_2)^2) = \sigma_{\beta_2}^2 + \sigma_{b_2}^2$. This suggests an estimate of $\sigma_{b_2}^2$ is given by $\max(0, (\hat{\beta}_2)^2 - \hat{\sigma}_{\beta_2}^2)$; hence, $\hat{\beta}_2^2$ can be considered a conservative estimate of σ_{b2}^2 which we used in our simulations. Second, we can use the maximum likelihood fit to the model $T|S, Z$ to find an estimate of σ_t^2. We can replace the parameters in $\text{E}(\beta)$ and $\text{V}(\beta)$ with their estimates to obtain the estimated mean and variance for β.

As before, the model $S|Z$ is given by $S_i = \alpha_0 + \alpha_1 Z_i + \epsilon_{si}$, where $\epsilon_{si} \sim N(0, \sigma_{ss}^2)$. Let $\alpha^T = (\alpha_0, \alpha_1)$ and $X_s = (1, Z)$, then $\hat{\alpha}$ follows a normal distribution with its

mean and variance:

$$\mathrm{E}(\hat{\alpha}|X_s, S) = (X_s^T X_s)^{-1} X_s^T S,$$
$$\mathrm{V}(\hat{\alpha}|X_s, S) = (X_s^T X_s)^{-1} \sigma_{ss}^2.$$

Let $D_{Ridge-EB}(Q) = (\frac{\partial Q}{\partial \beta_0}, \frac{\partial Q}{\partial \beta_1}, \frac{\partial Q}{\partial \beta_2}, \frac{\partial Q}{\partial \alpha_0}, \frac{\partial Q}{\partial \alpha_1}) = (0, \alpha_1, 1, 0, \beta_1)$. We can obtain that the treatment effect $\hat{Q}_{Ridge-EB}$ follows a normal distribution with mean and variance estimated by:

$$\hat{E}(\hat{Q}_{Ridge-EB}) = \hat{\beta}_1 \hat{\alpha}_1 + \hat{\beta}_2,$$
$$\hat{\mathrm{V}}(\hat{Q}_{Ridge-EB}) = D(\hat{Q}_{Ridge-EB})^T \begin{bmatrix} \hat{\mathrm{V}}(\hat{\beta}) & \mathbf{0} \\ \mathbf{0} & \hat{\mathrm{V}}(\hat{\alpha}) \end{bmatrix} D(\hat{Q}_{Ridge-EB}),$$

where the parameter estimates are obtained from the empirical Bayes estimation.

The estimation method can be easily extended to the situation where $\beta_3 \neq 0$. We use $\hat{\beta}_2^2$ as an estimate of $\sigma_{b_2}^2$, $\hat{\beta}_3^2$ for $\sigma_{b_3}^2$ and $\hat{\sigma}_t^2$ for σ_t^2, where $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\sigma}_t^2$ can be obtained from the maximum likelihood fit to the saturated model $T_i = \beta_0 + \beta_1 S_i + \beta_2 Z_i + \beta_3 S_i Z_i + \epsilon_{ti}$. The rest of the argument follows closely to that above.

3.5 Simulation Studies

3.5.1 The Setup

We conducted extensive simulations to examine the properties of the proposed methods and compare them with those of competing methods. We generated 400 data sets with the following parameter specifications: $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$ and $\alpha_1 = 2$. We varied β_2, β_3, σ_{ss}^2 and σ_t^2. For each combination of the parameters, we examine three scenarios: 1) $\sigma_{ss}^2 = 0.5$ and $\sigma_t^2 = 1$, 2) $\sigma_{ss}^2 = 0.5$ and $\sigma_t^2 = 0.1$ and 3) $\sigma_{ss}^2 = 5$ and $\sigma_t^2 = 1$. Since the results show very similar patterns across all three scenarios, we only present the results when $\sigma_{ss}^2 = 0.5$ and $\sigma_t^2 = 1$. Each data set

contains the observations from either 60, 120 or 480 subjects per treatment group. We observe all of S, but only 20% of T. For each method and each data set, we obtain the point estimate of Q and the corresponding estimated standard error (SE), and an indicator variable for whether or not the 95% confidence interval contains the true value. We examine the method's performance by its average bias (Bias), the average standard error (SE), the empirical standard deviation (ESD), the empirical mean squared error (MSE = ESD^2 + Bias^2) and the coverage rate (CR). For the fully Bayesian method, the standard error is given by the standard deviation of the posterior distribution.

3.5.2 Methods Compared

We examine the estimators of Q and their variances in the following scenarios:

1. $\hat{Q}_{ALL}$ and $\hat{V}(\hat{Q}_{ALL})$ from the simple estimation before any deletion on T;
2. $\hat{Q}_{CC}$ and $\hat{V}(\hat{Q}_{CC})$ from the simple estimation based on complete cases;
3. $\hat{Q}_{PES}$ and $\hat{V}(\hat{Q}_{PES})$ under the perfect surrogacy assumption;
4. $\hat{Q}_{APAS}$ and $\hat{V}(\hat{Q}_{APAS})$ under the additive partial surrogacy assumption;
5. $\hat{Q}_{IPAS}$ and $\hat{V}(\hat{Q}_{IPAS})$ under the interactive partially surrogacy assumption;
6. $\hat{Q}_{Ridge-FB}$ and $\hat{V}(\hat{Q}_{Ridge-FB})$ using fully Bayes ridge regression where $\hat{V}(\hat{Q}_{Ridge-FB})$ is given by the variance of the posterior distribution;
7. $\hat{Q}_{Ridge-EB}$ and $\hat{V}(\hat{Q}_{Ridge-EB})$ using empirical Bayse ridge regression;
8. $\hat{Q}_{IPW}$ using the inverse probability weighted method which is a competing method that can also be applied to utilize the information from auxiliary variables (Horvitz and Thompson, 1952; Zhao and Lipsitz, 1992; Zhao, Lipsitz and Lew, 1996). Let Δ_i be the indicator for whether T_i is observed or not (1 being observed and 0 for not being observed). Denote $\pi_i = Pr(\Delta_i = 1)$. We obtain

the estimated π_i ($\hat{\pi}_i$) by fitting the saturated model:

$$\text{logit}(Pr(\Delta_i = 1)) = \delta_0 + \delta_1 S_i + \delta_2 Z_i + \delta_3 S_i Z_i.$$

The treatment effect can be estimated by:

$$\hat{Q}_{IPW} = \frac{\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} T_i I(Z_i = 1)}{\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} I(Z_i = 1)} - \frac{\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} T_i I(Z_i = 0)}{\sum_i^n \frac{\Delta_i}{\hat{\pi}_i} I(Z_i = 0)}.$$

9. $\hat{Q}_{MdlSel}$ and $\hat{V}(\hat{Q}_{MdlSel})$ using a commonly used two-stage model selection method. At the first stage, one would decide on which model is not contradicted by the data. A common model selection method is backward elimination. We first fit the saturated model, $T_i = \beta_0 + \beta_1 S_i + \beta_2 Z_i + \beta_3 S_i Z_i + \epsilon_{ti}$, and then delete the variables one at a time by examining the p-values associated with $\hat{\beta}_3$ and then $\hat{\beta}_2$. At the second stage, we use the selected model as the correct model to obtain the estimate of Q and its inference. For example, if the interactive partial surrogacy assumption holds for the selected model, we obtain the estimate of Q as $\hat{Q}_{IPAS}$ and its variance as $\hat{V}(\hat{Q}_{IPAS})$; if the additive partial surrogacy assumption holds, we use $\hat{Q}_{APAS}$ and its variance and so forth.

3.5.3 Simulation Results

We assumed $\beta_3 = 0$ in the first set of simulations. The results are presented in Tables 3.3 and 3.4 and Figures 3.2, 3.3 and 3.4. First, we compare the properties of Ridge-FB and Ridge-EB methods with PES, APAS and IPAS methods. When $\beta_2 = 0$, fitting an APAS or IPAS model can result in much smaller efficiency gain and larger MSE relatively to fitting the PES model. When β_2 becomes much different from 0, fitting the PES model can lead to increasingly larger bias, larger MSE and lower coverage rates compared to fitting the APAS and IPAS models. When β_2 is 0 or close to 0, both Ridge-FB and Ridge-EB can retain most of the efficiency gain

achieved by fitting the PES model without introducing appreciable bias. When β_2 is much different from 0, the Ridge methods give estimates with MSEs comparable (larger sample) or close (smaller sample) to those obtained by fitting an APAS or IPAS model without the bias resulted from fitting an incorrect PES model. Hence, the Ridge methods can strike a balance between efficiency gain and bias depending on the nature of the relationship between S and T. This is due to the variance-bias tradeoff and data-adaptive capability of the Ridge methods. Similar findings are observed across different sizes, although more pronounced in small samples than these in large samples.

Second, we compare Ridge-FB with Ridge-EB. Both Ridge-FB and Ridge-EB are data-adaptive and have comparable performance in terms of bias, MSE, and coverage rate. When the sample size is large, Ridge-FB and Ridge-EB have very similar, if not identical, performance. However, there are some subtle differences, which stand out more in small samples. Except for very large samples, Ridge-EB gives below nominal-level coverage rates. Ridge-FB gives uniformly higher and closer-to-nominal coverage rates than Ridge-EB and any of the other methods. Unlike its competitors, Ridge-FB accounts for all the uncertainty associated with estimating the variance parameters. Generally, there is higher extent of shrinkage towards 0 using Ridge-FB than using Ridge-EB. Ridge-FB is usually more sensitive to the prior assumptions, particularly in smaller samples; on the other hand, Ridge-EB is more robust and less biased when there is a large departure in the prior distribution from the true value.

Third, we compare Ridge with inverse probability weighted (IPW) estimator. The IPW method is robust, but it cannot take direct advantage of the various plausible surrogacy assumptions. Regardless of the magnitude of β_2, the amount of efficiency gain from utilizing S to estimate Q stays the same, and similarly for bias and MSE.

When β_2 is close to 0, Ridge has a clear advantage over IPW and gives considerably smaller MSE than IPW. When β_2 is much larger than 0, the efficiency gain from Ridge and IPW are more or less comparable in larger samples, although IPW gives smaller MSEs than Ridge-FB but comparable MSEs with Ridge-EB when the sample size is 60 per group. In our setting, IPW is comparable to the performance from fitting the APAS or IPAS models in terms of efficiency gain, particularly in large samples. However, IPW gives estimates with bigger MSE and less precision in small samples.

Fourth, we compare Ridge with model selection (MdlSel). Similar to Ridge, MdlSel is also a data-adaptive method; but, different from Ridge, its performance heavily relies on the available power from the data in choosing the correct model and making correct surrogacy assumptions. When the power is small, (e.g. when β_2 is somewhere in the middle between being too small and being too large, or when the sample size is small), Ridge can achieve smaller MSE and more efficiency gain than MdlSel. On the other hand, when there is more statistical power, (e.g. when the size is 120 or 480 per group and when β_2 is either ≈ 0 or very large), MdlSel and Ridge have similar performance in terms of MSE and efficiency gain. For the MdlSel method, the common practice of computing $\hat{V}(\hat{Q})$ based on the selected model fails to account for the variation in the model selection process. Hence, MdlSel generally underestimates the variance (i.e., ESD $>$ SE in all simulations), more so in smaller samples. The extent of underestimation depends on the power to detect the correct model. The underestimation of $V(\hat{Q})$ results in lower-than-nominal-level CRs, which are close to being adequate in most situations, but more variable and are the lowest in a few cases among all methods.

In the next set of simulations, we assume $\beta_3 \neq 0$. The results are presented in Tables $3.5 - 3.10$ and Figure $3.5 - 3.7$. We have similar findings as those from the

first set of simulations. When both β_2 and β_3 are relatively small or the sample size is fairly small, Ridge retains some of the efficiency gain from PES and offers smaller MSEs than either APAS or IPAS. The extent of the efficiency gain is generally smaller than when $\beta_3 = 0$, since there is a cost in efficiency associated with the increased number of estimated parameters. When β_2 or β_3 is very different from 0 and the sample size is relatively large, Ridge gives estimates with comparable MSEs and CRs as APAS or IPAS and can minimize the huge bias and very low CR from fitting PES. Even when IPAS is the correct surrogacy assumption, IPAS sometimes gives bigger MSEs than APAS in small samples, while at other times gives smaller MSEs in large samples. It is the result from the compromise between two tradeoffs: one between the number of parameters and variance, and the other between bias and variance. Relative to IPW, we find that Ridge has consistently smaller MSEs, except when both β_2 and β_3 differ significantly from 0 and the sample size is very large, where MSEs from Ridge and IPW are similar.

Next we compare Ridge with MdlSel. Generally, the performance of MdlSel appears to be even more variable than that in the previous simulations in terms of MSE, bias and CR. When there is low power from the data to detect the difference of β_2 or β_3 from 0, MdlSel gives estimates with larger MSEs than Ridge and ignoring the uncertainty in the model selection process results in consistently less-than-acceptable coverage rates.

3.6 Application to a Glaucoma Study

We apply these methods to data from the Collaborative Initial Glaucoma Treatment Study (CIGTS) (Musch *et al.*, 1999). Glaucoma is a group of diseases that cause vision loss and is a leading cause for blindness. Elevated pressure in the eyes

(i.e., intraocular pressure, IOP), is a major risk factor of glaucoma. The CIGTS is a randomized trial to compare the effects of two types of treatments, surgery ($Z = 1$) and medicine ($Z = 0$), on reducing IOP among glaucoma patients. Patients were enrolled between 1993 and 1997. The IOP level (recorded in mmHg) has been measured at different time points following randomization. For the purpose of this paper we take the true endpoint to be the IOP measurements at the 102nd month and the surrogate marker to be the IOP level at the 12th month. Due to drop out, there are many fewer patients at later periods than at earlier periods. A total of 160 patients have IOP measured at months 12 and 102, and 413 patients measured only at month 12. The missingness is not significantly associated with S or Z and seems to satisfy the missing completely at random assumption. The correlation between S and T is 0.456 (p-value $< .0001$). Summary statistics are presented in Table 3.11.

For the $S|Z$ model, based on all 413 patients, we obtain the ordinary least squared (OLS) estimates of the parameters and their 95% confidence intervals (CI): $\alpha_0 = 21.90$ $(20.75, 23.04)$ and $\alpha_1 = -3.83$ $(-4.55, -3.10)$. For the $T|S, Z$ model, we obtain the OLS estimates by fitting three regression models to the data from 160 patients. By assuming IPAS, we obtain the parameters and their 95% CIs: $\hat{\beta}_1 = 0.61$ $(0.012, 1.20)$, $\hat{\beta}_2 = 0.87$ $(-4.99, 6.74)$ and $\hat{\beta}_3 = -0.094$ $(-0.44, 0.25)$. By assuming APAS, we have: $\hat{\beta}_1 = 0.45$ $(0.29, 0.61)$, $\hat{\beta}_2 = -0.69$ $(-2.16, 0.78)$. By assuming PES, we have $\hat{\beta}_1 = 0.48$ $(0.33, 0.63)$. While the two-stage model selection method would choose the perfect surrogacy model, there is much uncertainty about whether this assumption could hold because the number of complete cases is relatively small and one single study has limited power for detecting real differences. However, the preliminary analysis does indicate that S can capture most of the treatment effect on T, implying that S is a good surrogate marker. Table 3.12 shows the estimates of

the treatment difference between two groups and their 95% CIs. The Ridge method assume $\beta_3 = 0$. Fitting either the IPAS or APAS model results in CIs with widths slightly narrower than that from the CC method, showing that there is a very limited efficiency gain by utilizing S. Fitting the PES model leads to substantial efficiency gain; however, the estimate is quite different from other estimates, perhaps due to the fact that a potentially incorrect PES assumption could lead to substantial bias. Results from fitting Ridge-FB and Ridge-EB are comparable, resulting in estimates with more precision and possibly more bias than those by fitting IPAS or APAS, but less than that by fitting PES. The results reflect the data-adaptive and bias-variance tradeoffs feature of the Ridge methods.

3.7 Discussion

In this article, we propose a shrinkage approach to utilize the information from S to estimate the treatment effect on T. Without the need to make correct surrogacy assumptions, ridge regression can directly take advantage of the relationship between S and T, increase the information recovery from S and, hence, estimate precision. When S captures most of the treatment effect, the generalized ridge regression method can retain most of the considerable efficiency gain achieved under the perfect surrogacy assumption. When S only captures modest amount of the treatment effect, using S can lead to efficiency comparable to that under partial surrogacy assumptions, while limiting the bias by making an incorrect perfect surrogacy assumption. The proposed ridge-based methods is a robust estimation approach, have the bias-variance tradeoff and data-adaptive property and can strike a balance between bias and efficiency gain depending on the evidence from the data regarding the validity of a surrogacy assumption. We propose both full Bayes and empirical Bayes versions of

the generalized ridge regression. When the sample size is small to modest, the Bayes version gives smaller MSEs and better coverage rates than the EB version. On the other hand, EB is much faster than FB particularly when the sample size is very large. It is also more robust and less biased when there is a very large departure from the perfect surrogate assumption.

From a statistical point of view, we have touched upon two important areas of research: model selection and missing data. First, let us consider model selection methods in our setting. A common statistical practice is to select a parsimonious model for $T|S, Z$ and then use the selected model to predict the effect on T from S. In situations when the power to detect the correct assumption is relatively small, the uncertainty of model selection procedure is very large and ignoring such uncertainty can result in very low coverage and large bias. The ridge regression methods outperforms the model selection methods in terms of MSE, bias and CRs in such situations. In practice, when a good marker S is observed on more subjects than T, we would suggest to use S to enhance the inference of the treatment effect. We could use a model selection method as a screening tool. When there is less certainty about which surrogacy assumption holds, we would recommend to use a generalized ridge regression method. A feature of the ridge regression method is that it does not drop any variable and conducts model selection. As a result, it cannot achieve full efficiency when the true parameter β_2 is actually equal to 0, as our simulations have shown. This does not pose a problem in the surrogate marker setting as previous empirical studies have shown that it is more likely for a good surrogate marker to capture most of the treatment effect but unlikely for S to be a perfect surrogate (Fleming and DeMets, 1996). For such situations, a ridge regression method could have a very good performance and have a clear advantage over model selection meth-

ods in terms of MSE, efficiency gain and bias. In some practical settings when it is a reasonable option to drop a variable and choose a more parsimonious model, different prior distributions (such as a mixture prior with point mass at $\beta_2 = 0$) can be used instead of the normal prior. Or a hybrid method that can perform both model selection and shrinkage such as the least absolute shrinkage and selection operator method (LASSO) (Tibshirani, 1996) can be considered.

Second, let us consider our methods in the missing-data research context, since utilizing S in predicting treatment effect is essentially a missing data problem. We compared the generalized ridge regression with the inverse probability weighted method. Although the inverse probability weighted method is a robust method, it requires us to model the probability of missingness and does not have the data-adaptive feature of the generalized ridge regression and take direct advantage of the nature of the relationship between S and T. Hence, when S is close to being perfect, our method can give smaller MSEs and achieve much more substantial efficiency gain than IPW. On the other hand, a more data-adaptive solution such as multiple imputation method that uses the $T|S, Z$ model to impute missing T would likely achieve similar efficiency gain and robustness if we use the generalized ridge regression to model $T|S, Z$. A comparison with the improved IPW methods (Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999) is also worthy of investigation.

Many extensions to the generalized ridge regression method can be made in the surrogate marker context. When multiple biomarkers are considered, there could be even a stronger motivation for the use of a ridge regression method, since a greater percentage of the treatment effect may be captured by the biomarkers. Besides the data-adaptive and robust features of ridge regression, we can also take advantage of its ability to reduce the collinearity problem in the multiple biomarkers setting.

The idea can also be extended to the cases when S and T are different data types; such as time-to-event, which can be more challenging, particularly when we need to consider censoring for both S and T. In summary, ridge regression is a rich research area worthy of further study and implementation.

										Relative Efficiency		
Q	p	β_0	β_1	α_0	α_1	σ_t^2	σ_{ss}^2	ρ^2	$V(Q_{All})$	CC	IPAS	APAS	PES
2	0.7	0.5	1	0.005	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	0.5	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	5	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	10	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	20	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	50	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
0.05	0.7	0.5	1	1	0.05	1	0.5	0.333	0.006	0.3	0.391	0.391	2.975
0.5	0.7	0.5	1	1	0.5	1	0.5	0.333	0.006	0.3	0.391	0.391	1.723
1	0.7	0.5	1	1	1	1	0.5	0.333	0.006	0.3	0.391	0.391	0.931
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
5	0.7	0.5	1	1	5	1	0.5	0.333	0.006	0.3	0.391	0.391	0.418
10	0.7	0.5	1	1	10	1	0.5	0.333	0.006	0.3	0.391	0.391	0.398
20	0.7	0.5	1	1	20	1	0.5	0.333	0.006	0.3	0.391	0.391	0.393
2	0.7	0.05	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	1	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	10	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	20	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	40	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
0.1	0.7	0.5	0.05	1	2	1	0.5	0.001	0.004	0.3	0.300	0.300	0.450
1	0.7	0.5	0.5	1	2	1	0.5	0.111	0.005	0.3	0.325	0.325	0.479
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
4	0.7	0.5	2	1	2	1	0.5	0.667	0.012	0.3	0.563	0.563	0.711
10	0.7	0.5	5	1	2	1	0.5	0.926	0.054	0.3	0.853	0.853	0.917
20	0.7	0.5	10	1	2	1	0.5	0.980	0.204	0.3	0.956	0.956	0.977
40	0.7	0.5	20	1	2	1	0.5	0.995	0.804	0.3	0.989	0.989	0.994
200	0.7	0.5	100	1	2	1	0.5	1.000	20.004	0.3	1.000	1.000	1.000
2	0.7	0.5	1	1	2	0.01	0.5	0.980	0.002	0.3	0.956	0.956	0.977
2	0.7	0.5	1	1	2	0.1	0.5	0.833	0.002	0.3	0.720	0.720	0.831
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	1	2	5	0.5	0.091	0.022	0.3	0.320	0.320	0.474
2	0.7	0.5	1	1	2	10	0.5	0.048	0.042	0.3	0.310	0.310	0.462
2	0.7	0.5	1	1	2	20	0.5	0.024	0.082	0.3	0.305	0.305	0.456
2	0.7	0.5	1	1	2	40	0.5	0.012	0.162	0.3	0.303	0.303	0.453

Table 3.1: Asymptotic Variance Calculations. Relative Efficiency (RE) Compared with that Obtained from Original Data (ALL). True Model: Perfect Surrogacy ($n = 1000$)

										Relative Efficiency		
Q	p	β_0	β_1	α_0	α_1	σ_t^2	σ_{ss}^2	ρ^2	$V(Q_{All})$	CC	IPAS	APAS	PES
2	0.7	0.5	1	1	2	1	0.01	0.010	0.004	0.3	0.302	0.302	0.305
2	0.7	0.5	1	1	2	1	0.05	0.048	0.004	0.3	0.310	0.310	0.326
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.7	0.5	1	1	2	1	1	0.500	0.008	0.3	0.462	0.462	0.750
2	0.7	0.5	1	1	2	1	5	0.833	0.024	0.3	0.720	0.720	1.080
2	0.7	0.5	1	1	2	1	10	0.909	0.044	0.3	0.825	0.825	1.068
2	0.7	0.5	1	1	2	1	20	0.952	0.084	0.3	0.900	0.900	1.042
2	0.7	0.5	1	1	2	1	40	0.976	0.164	0.3	0.946	0.946	1.023
2	0.7	0.5	1	1	2	1	100	0.990	0.404	0.3	0.977	0.977	1.010
2	0.1	0.5	1	1	2	1	0.5	0.333	0.006	0.9	0.931	0.931	1.209
2	0.2	0.5	1	1	2	1	0.5	0.333	0.006	0.8	0.857	0.857	1.125
2	0.3	0.5	1	1	2	1	0.5	0.333	0.006	0.7	0.778	0.778	1.033
2	0.4	0.5	1	1	2	1	0.5	0.333	0.006	0.6	0.692	0.692	0.931
2	0.5	0.5	1	1	2	1	0.5	0.333	0.006	0.5	0.600	0.600	0.818
2	0.6	0.5	1	1	2	1	0.5	0.333	0.006	0.4	0.500	0.500	0.692
2	0.7	0.5	1	1	2	1	0.5	0.333	0.006	0.3	0.391	0.391	0.551
2	0.8	0.5	1	1	2	1	0.5	0.333	0.006	0.2	0.273	0.273	0.391
2	0.9	0.5	1	1	2	1	0.5	0.333	0.006	0.1	0.143	0.143	0.209
2	0.95	0.5	1	1	2	1	0.5	0.333	0.006	0.05	0.073	0.073	0.108
2	0.1	0.5	1	1	2	1	20	0.952	0.084	0.9	0.995	0.995	1.047
2	0.2	0.5	1	1	2	1	20	0.952	0.084	0.8	0.988	0.988	1.047
2	0.3	0.5	1	1	2	1	20	0.952	0.084	0.7	0.980	0.980	1.046
2	0.4	0.5	1	1	2	1	20	0.952	0.084	0.6	0.969	0.969	1.046
2	0.5	0.5	1	1	2	1	20	0.952	0.084	0.5	0.955	0.955	1.045
2	0.6	0.5	1	1	2	1	20	0.952	0.084	0.4	0.933	0.933	1.044
2	0.7	0.5	1	1	2	1	20	0.952	0.084	0.3	0.900	0.900	1.042
2	0.8	0.5	1	1	2	1	20	0.952	0.084	0.2	0.840	0.840	1.038
2	0.9	0.5	1	1	2	1	20	0.952	0.084	0.1	0.700	0.700	1.026
2	0.95	0.5	1	1	2	1	20	0.952	0.084	0.05	0.525	0.525	1.002

Table 3.2: Asymptotic Variance Calculations. Relative Efficiency (RE) Compared with that Obtained from Original Data (ALL). True Model: Perfect Surrogacy ($n = 1000$)

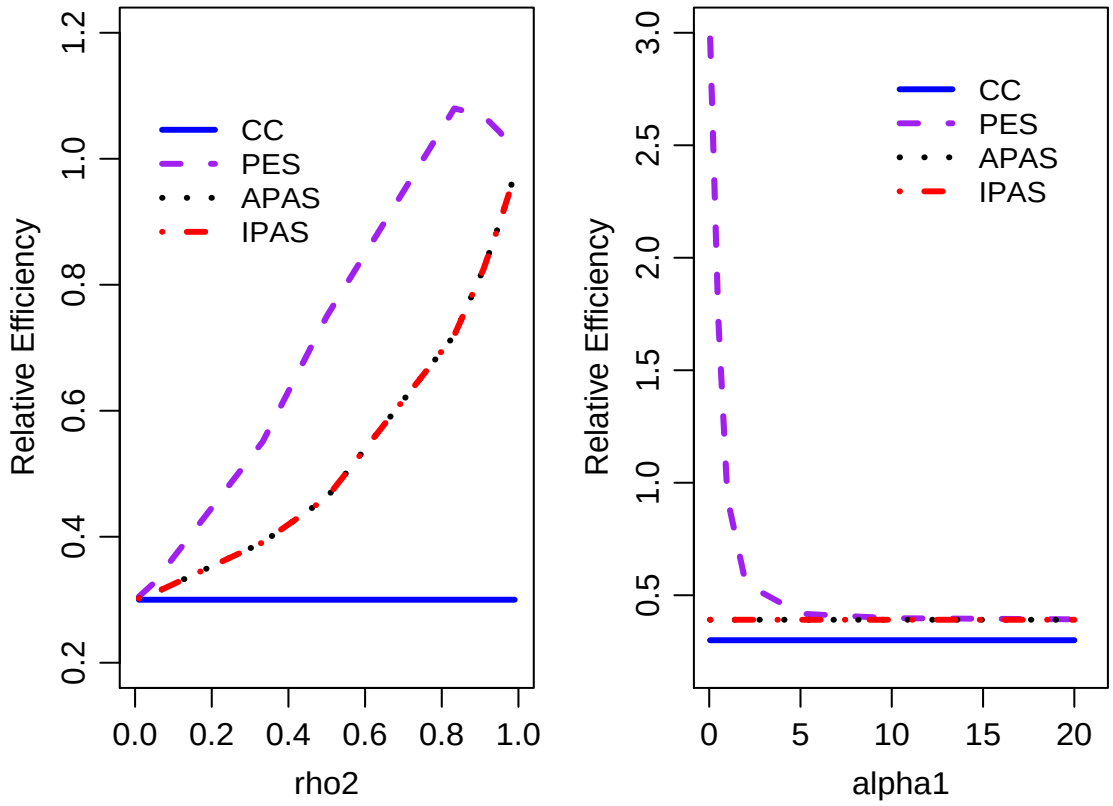

Figure 3.1: Asymptotic Relative Efficiency (RE) Compared with that Obtained from Original Data (ALL). Left: $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma_t^2 = 1$, $p = 0.7$, and ρ^2 varies. Right: $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\sigma_t^2 = 1$, $p = 0.7$, $\sigma_{ss}^2 = 0.5$, $\rho^2 = 0.333$ and α_1 varies. ($n = 1000$)

n_1, n_2	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
$60, 60$	0	2	Bias	5	36	29	42	50	37	40	43	40
			MSE	51	259	133	207	216	156	172	224	167
			ESD	226	508	364	453	462	393	413	471	406
			SE	222	511	376	450	466	429	403	-	381
			CR	94.3	92.5	95.8	93.0	94.0	96.3	93.8	-	95.3
	0.2	2.2	Bias	5	36	-34	42	50	4	20	43	-6
			MSE	51	259	135	207	216	157	173	224	170
			ESD	226	508	366	453	462	396	416	471	412
			SE	222	511	376	450	466	433	405	-	385
			CR	94.3	92.5	95.5	93.0	94.0	96.3	93.3	-	94.3
	0.6	2.6	Bias	5	36	-159	42	50	-54	-15	43	-75
			MSE	51	259	163	207	216	169	183	224	194
			ESD	226	508	371	453	462	401	427	471	434
			SE	222	511	391	450	466	444	412	-	395
			CR	94.3	92.5	92.3	93.0	94.0	95.5	92.8	-	91.3
	2	4	Bias	5	36	-598	42	50	-115	-37	43	-40
			MSE	51	259	527	207	216	236	217	224	275
			ESD	226	508	412	453	462	408	464	471	523
			SE	222	511	464	450	466	493	440	-	436
			CR	94.3	92.5	73.3	93.0	94.0	93.5	91.8	-	89.0
$120, 120$	0	2	Bias	6	16	24	14	14	19	18	14	21
			MSE	26	131	73	107	109	83	90	111	85
			ESD	161	361	269	327	330	287	299	334	291
			SE	158	358	257	310	314	290	278	-	261
			CR	94.3	94.3	92.3	92.3	91.8	93.5	90.8	-	90.3
	0.2	2.2	Bias	6	16	-41	14	14	-12	-3	14	-23
			MSE	26	131	75	107	109	84	88	111	87
			ESD	161	361	272	327	330	290	302	334	294
			SE	158	358	258	310	314	292	279	-	264
			CR	94.3	94.3	92.3	92.3	91.8	93.5	90.5	-	89.3
	0.6	2.6	Bias	6	16	-170	14	14	-65	-35	14	-82
			MSE	26	131	106	107	109	97	101	111	115
			ESD	161	361	277	327	330	305	315	334	329
			SE	158	358	268	310	314	302	286	-	274
			CR	94.3	94.3	89.0	92.3	91.8	94.5	91.3	-	89.0
	2	4	Bias	6	16	-621	14	14	-62	-31	14	-3
			MSE	26	131	481	107	109	124	112	111	119
			ESD	161	361	309	327	330	347	334	334	345
			SE	158	358	317	310	314	329	307	-	307
			CR	94.3	94.3	45.8	92.3	91.8	91.8	91.5	-	90.5

Table 3.3: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2 = 0.333$ and $p = 0.8$.

n_1, n_2	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
480, 480	0	2	Bias	-5	10	-3	-2	-2	-2	-2	-2	-5
			MSE	6	33	15	24	24	18	19	24	17
			ESD	80	181	120	155	155	134	139	155	131
			SE	79	177	127	153	153	142	136	-	130
			CR	95	93.8	95.8	94.5	94.5	96	95.0	-	94.5
	0.2	2.2	Bias	-5	10	-69	-2	-2	-29	-22	-2	-43
			MSE	6	33	19	24	24	20	21	24	23
			ESD	80	181	121	155	155	138	145	155	145
			SE	79	177	128	153	153	145	139	-	132
			CR	95	93.8	93.0	94.5	94.5	95.0	93.8	-	91.5
	0.6	2.6	Bias	-5	10	-200	-2	-2	-49	-31	-2	-41
			MSE	6	33	55	24	24	27	26	24	35
			ESD	80	181	122	155	155	158	158	155	183
			SE	79	177	132	153	153	155	148	-	145
			CR	95	93.8	67.5	94.5	94.5	92.8	92.3	-	83.5
	2	4	Bias	-5	10	-661	-2	-2	-13	-13	-2	-2
			MSE	6	33	455	24	24	24	24	24	24
			ESD	80	181	134	155	155	155	156	155	155
			SE	79	177	156	153	153	154	152	-	152
			CR	95	93.8	0.8	94.5	95	94.8	95	-	94.5

Table 3.4: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2 = 0.333$ and $p = 0.8$.

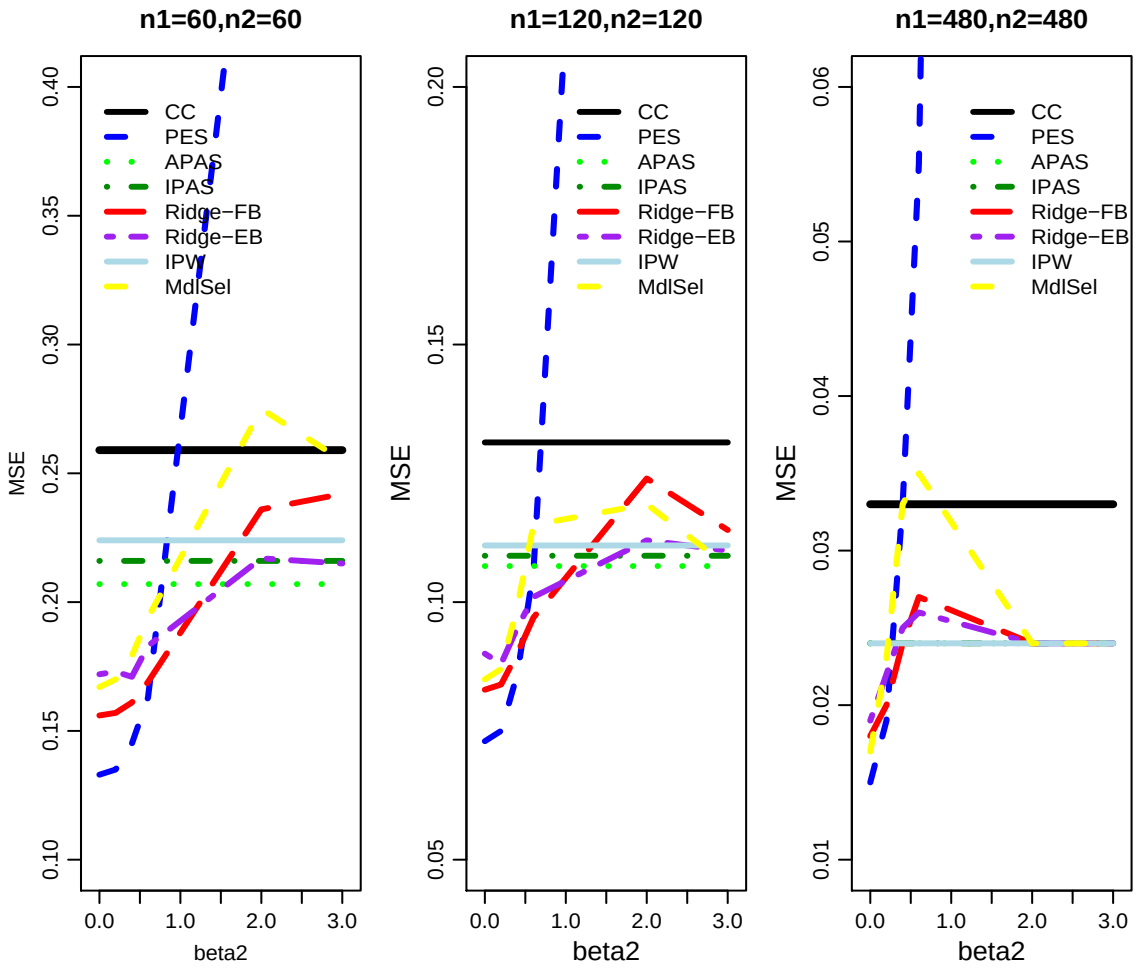

Figure 3.2: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2 = 0.333$ and $p = 0.8$.

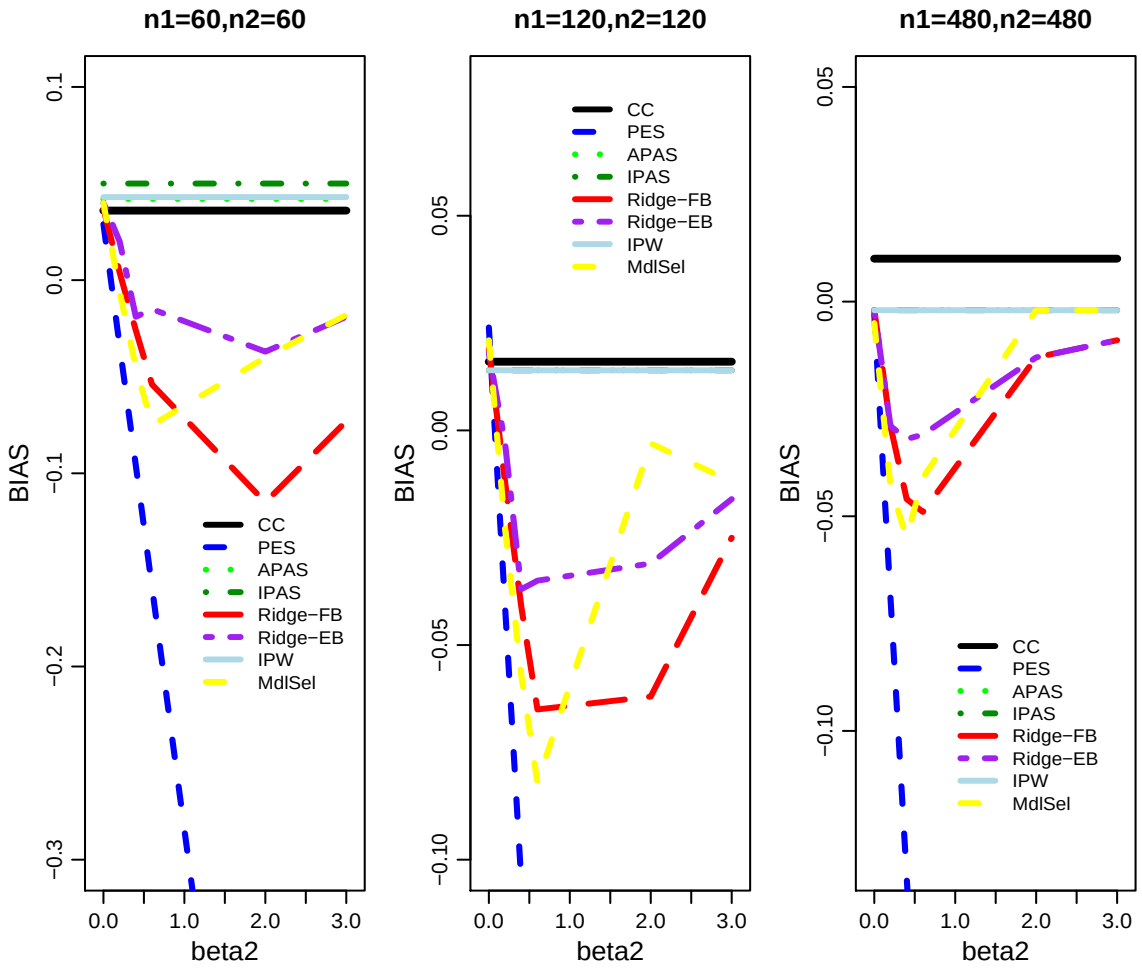

Figure 3.3: Bias by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2 = 0.333$ and $p = 0.8$.

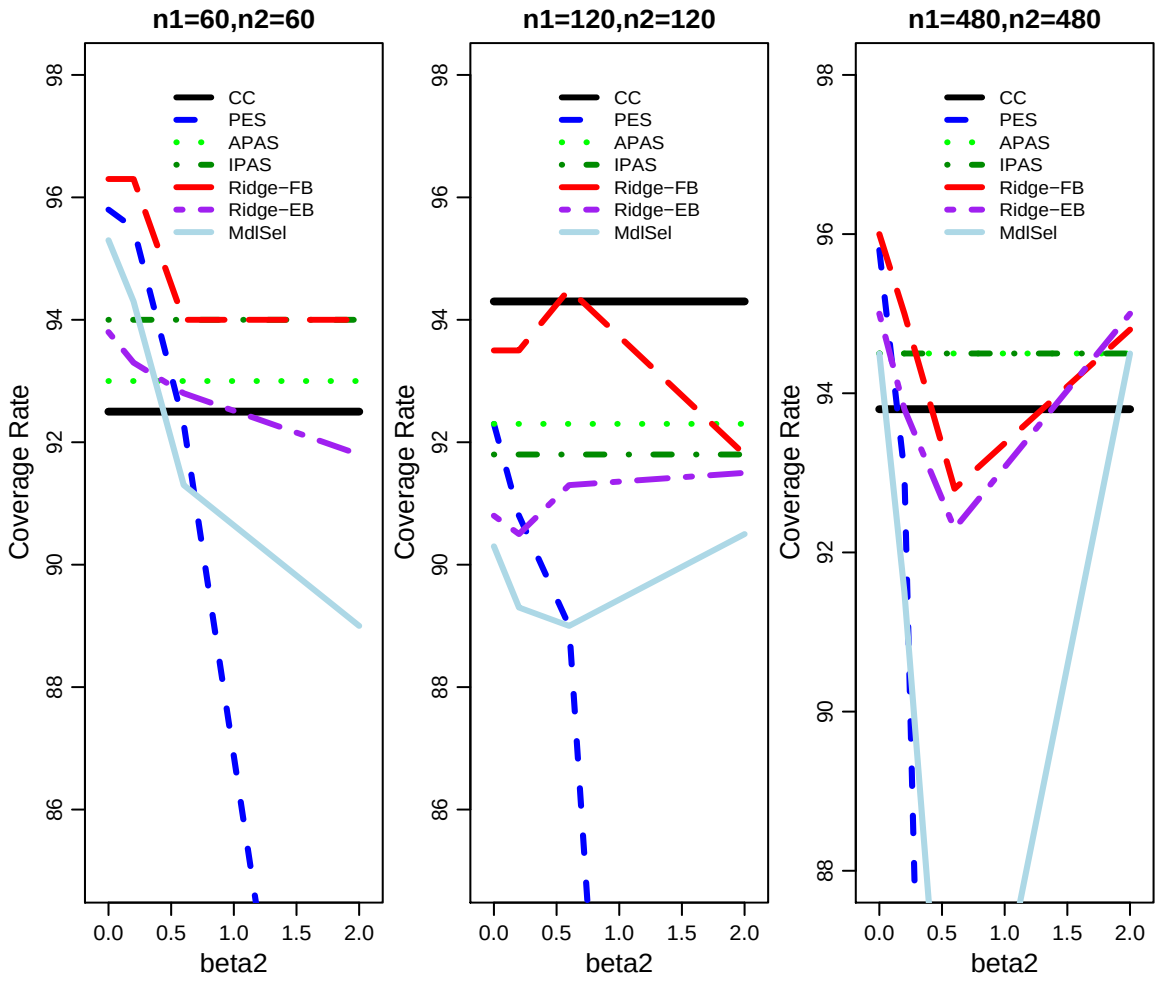

Figure 3.4: Coverage Rate by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2 = 0.333$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
60, 60	0.2	0	2.6	Bias	5	33	-97	40	49	2	23	42	-48
				MSE	54	276	149	211	220	172	190	230	179
				ESD	233	525	374	458	466	415	435	478	420
				SE	230	530	390	455	470	470	436	-	395
				CR	94.5	92.8	93.5	92.5	94.0	95.3	94.0	-	91.3
	0.2	0.2	2.8	Bias	5	33	-159	40	49	-19	12	42	-71
				MSE	54	276	167	211	220	177	193	230	199
				ESD	233	525	377	458	466	420	439	478	441
				SE	230	530	397	455	470	475	438	-	401
				CR	94.5	92.8	92.3	92.5	94.0	95.0	93.5	-	90.8
	0.2	0.6	3.2	Bias	5	33	-285	40	49	-54	-5	42	-114
				MSE	54	276	230	211	220	190	200	230	248
				ESD	233	525	386	458	466	432	448	478	484
				SE	230	530	414	455	470	486	442	-	413
				CR	94.5	92.8	87.3	92.5	94.0	95.3	92.3	-	86.3
	0.2	2	4.6	Bias	5	33	-723	40	49	-86	-19	42	-17
				MSE	54	276	714	211	220	226	216	230	274
				ESD	233	525	436	458	466	468	465	478	523
				SE	230	530	498	455	470	519	453	-	446
				CR	94.5	92.8	66.8	92.5	94.0	94.5	92.3	-	88.8
120, 120	0.2	0	2.6	Bias	6	15	-105	12	14	-12	-0.2	14	-53
				MSE	28	140	89	110	111	95	103	114	108
				ESD	167	373	280	332	333	308	321	338	324
				SE	164	372	268	313	317	319	299	-	274
				CR	95.0	94.5	90.3	92.0	91.8	95.0	91.8	-	88.5
	0.2	0.2	2.8	Bias	6	15	-170	12	14	-31	-11	14	-72
				MSE	28	140	109	110	111	99	105	114	118
				ESD	167	373	283	332	333	313	324	338	336
				SE	164	372	273	313	317	323	300	-	279
				CR	95.0	94.5	88.8	92.0	91.8	95.5	91.8	-	88.5
	0.2	0.6	3.2	Bias	6	15	-299	12	14	-57	-27	14	-91
				MSE	28	140	174	110	111	109	108	114	144
				ESD	167	373	291	332	333	325	328	338	368
				SE	164	372	284	313	317	331	302	-	291
				CR	95.0	94.5	79.0	92.0	91.8	95.3	92.5	-	85
	0.2	2	4.6	Bias	6	15	-750	12	14	-54	-32	14	6
				MSE	28	140	669	110	111	121	114	114	120
				ESD	167	373	327	332	333	344	336	338	346
				SE	164	372	341	313	317	348	311	-	312
				CR	95.0	94.5	36.3	92	91.8	95.3	91.3	-	90.5

Table 3.5: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.333$, ${\rho_1}^2 = 0.419$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
480, 480	0.2	0	2.6	Bias	-4	11	-134	-2	-2	-23	-17	-1	-48
				MSE	7	36	33	25	25	24	24	25	34
				ESD	83	188	124	157	157	152	154	157	177
				SE	82	183	132	154	153	160	148	-	141
				CR	94.3	93.5	83.0	94.5	94.3	96.3	93.5	-	83.3
	0.2	0.2	2.8	Bias	-4	11	-200	-2	-2	-32	-24	-1	-38
				MSE	7	36	55	25	25	25	25	25	36
				ESD	83	188	124	157	157	156	156	157	185
				SE	82	183	134	154	154	163	149	-	147
				CR	94.3	93.5	68.0	94.5	94.3	95.0	93.5	-	83.5
	0.2	0.6	3.2	Bias	-4	11	-331	-2	-2	-35	-27	-1	-5
				MSE	7	36	126	25	25	27	26	25	26
				ESD	83	188	127	157	157	161	159	157	163
				SE	82	183	140	154	154	167	151	-	153
				CR	94.3	93.5	31.0	94.5	94.3	96.3	94	-	92.3
	0.2	2	4.6	Bias	-4	11	-792	-2	-2	-14	-13	-1	-2
				MSE	7	36	648	25	25	25	25	25	25
				ESD	83	188	141	157	157	158	158	157	157
				SE	82	183	168	154	154	169	154	-	153
				CR	94.3	93.5	0	94.5	94.3	96.3	94.5	-	94.5

Table 3.6: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.333$, ${\rho_1}^2 = 0.419$ and $p = 0.8$.

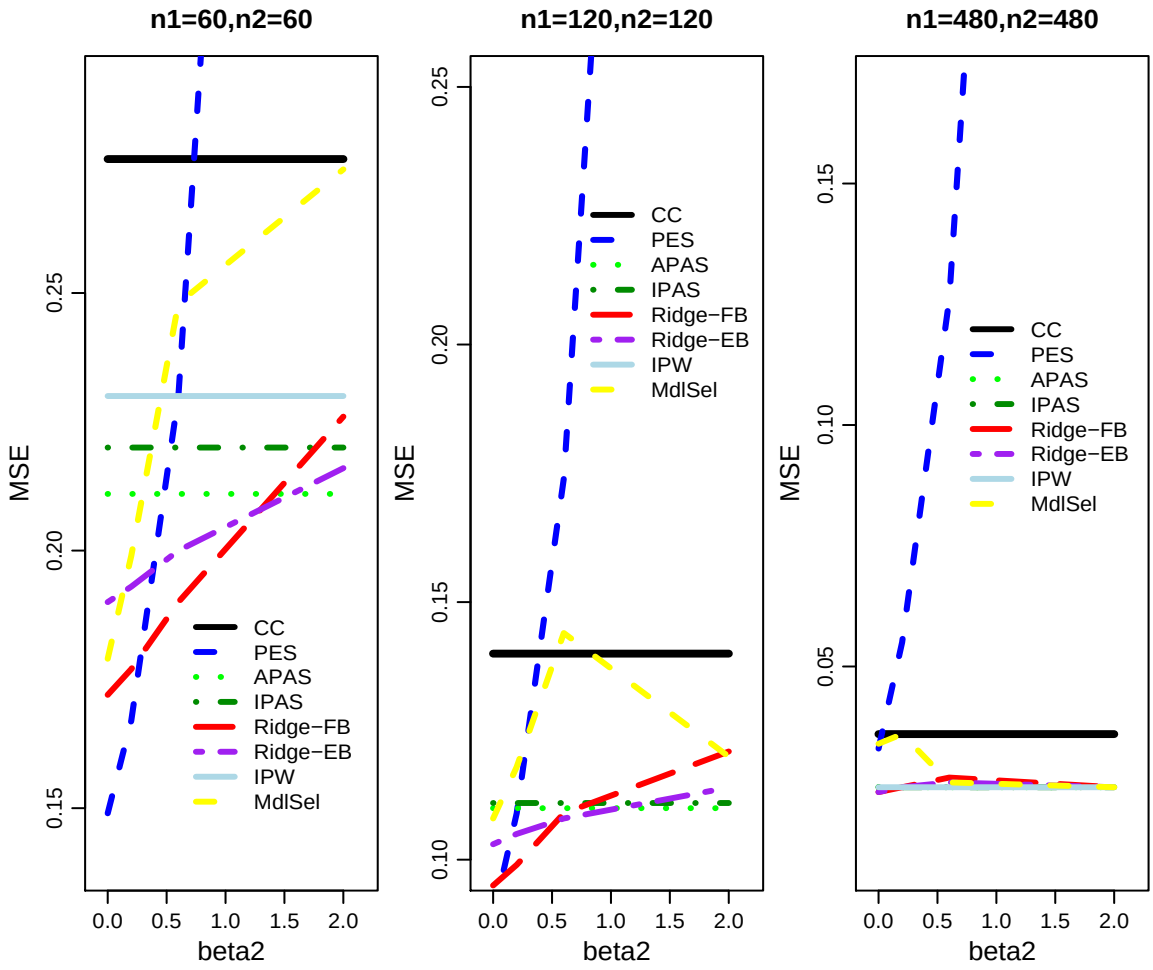

Figure 3.5: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0.2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2_0 = 0.333$, $\rho^2_0 = 0.419$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
60, 60	0.6	0	3.8	Bias	3	27	-348	35	48	-39	1	40	-94
				MSE	63	321	291	227	229	203	210	244	277
				ESD	251	566	412	475	476	449	458	493	518
				SE	249	574	441	471	480	503	453	-	437
				CR	94.3	93.8	86.3	94.0	93.5	96.0	93.3	-	85.3
	0.6	0.2	4	Bias	3	27	-411	35	48	-50	-6	40	-92
				MSE	63	321	344	227	229	208	213	244	295
				ESD	251	566	418	475	476	453	461	493	535
				SE	249	574	452	471	480	508	455	-	442
				CR	94.3	93.8	83.3	94.0	93.8	95.8	92.5	-	86.8
	0.6	0.6	4.4	Bias	3	27	-536	35	48	-64	-16	40	-60
				MSE	63	321	474	227	229	220	219	244	295
				ESD	251	566	432	475	476	465	468	493	540
				SE	249	574	475	471	480	516	458	-	452
				CR	94.3	93.8	77.3	94.0	93.8	94.8	91.8	-	87.5
	0.6	2	5.8	Bias	3	27	-975	35	48	-67	-20	40	19
				MSE	63	321	1197	227	229	241	229	244	262
				ESD	251	566	497	475	476	486	478	493	512
				SE	249	574	578	471	480	535	466	-	465
				CR	94.3	93.8	58.8	94.0	93.8	94.8	92.0	-	91.0
120, 120	0.6	0	3.8	Bias	6	13	-363	10	15	-31	-11	14	-53
				MSE	33	162	229	120	117	114	115	121	148
				ESD	181	403	312	347	342	337	338	348	381
				SE	177	403	303	325	324	341	312	-	310
				CR	94.5	95.5	74.0	93.0	91.8	95.0	91.8	-	86.8
	0.6	0.2	4	Bias	6	13	-427	10	15	-36	-17	14	-37
				MSE	33	162	283	120	117	117	116	121	145
				ESD	181	403	316	347	342	340	340	348	379
				SE	177	403	310	325	324	343	312	-	314
				CR	94.5	95.5	68.3	93.0	91.8	95.3	91.8	-	87.3
	0.6	0.6	4.4	Bias	6	13	-556	10	15	-43	-26	14	-11
				MSE	33	162	416	120	117	120	118	121	137
				ESD	181	403	327	347	342	344	342	348	371
				SE	177	403	326	325	324	347	314	-	318
				CR	94.5	95.5	56.3	93.0	91.8	95.0	91.0	-	89.3
	0.6	2	5.8	Bias	6	13	-1008	10	15	-47	-30	14	13
				MSE	33	162	1152	120	117	126	120	121	119
				ESD	181	403	370	347	342	352	345	348	345
				SE	177	403	395	325	324	355	320	-	322
				CR	94.5	95.5	25.5	93.0	91.8	95.5	92.0	-	91.5

Table 3.7: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma_{ss}^2 = 0.5$, $\sigma_t^2 = 1$, ${\rho_0}^2 = 0.333$, ${\rho_0}^2 = 0.561$ and $p = 0.8$.

										Ridge	Ridge		Mdl-
n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	FB	EB	IPW	Sel
480, 480	0.6	0	3.8	Bias	-3	14	-396	-1	-0.9	-12	-11	-0.5	-1
				MSE	8	42	175	27	26	25	27	26	26
				ESD	90	205	136	164	162	159	160	162	162
				SE	89	198	149	160	158	166	154	-	158
				CR	94.3	93.5	21.0	94.5	94.8	96.0	93.3	-	94.8
	0.6	0.2	4	Bias	-3	14	-462	-1	-0.9	-20	-16	-0.5	-1
				MSE	8	42	232	27	26	26	26	26	26
				ESD	90	205	137	164	162	159	160	162	162
				SE	89	198	153	160	158	167	154	-	158
				CR	94.3	93.5	12.8	94.5	94.8	95.5	93.3	-	94.8
	0.6	0.6	4.4	Bias	-3	14	-594	-1	-0.9	-30	-21	-0.5	-1
				MSE	8	42	372	27	26	27	27	26	26
				ESD	90	205	141	164	161	162	162	162	162
				SE	89	198	161	160	158	170	155	-	158
				CR	94.3	93.5	3.5	94.5	94.8	95.8	93.8	-	94.8
	0.6	2	5.8	Bias	-3	14	-1054	-1	-0.9	-14	-12	-0.5	-1
				MSE	8	42	1137	27	26	27	27	26	26
				ESD	90	205	160	164	161	163	163	162	162
				SE	89	198	195	160	158	174	158	-	158
				CR	94.3	93.5	0	94.5	94.8	96.3	94.3	-	94.8

Table 3.8: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.333$, ${\rho_1}^2 = 0.561$ and $p = 0.8$.

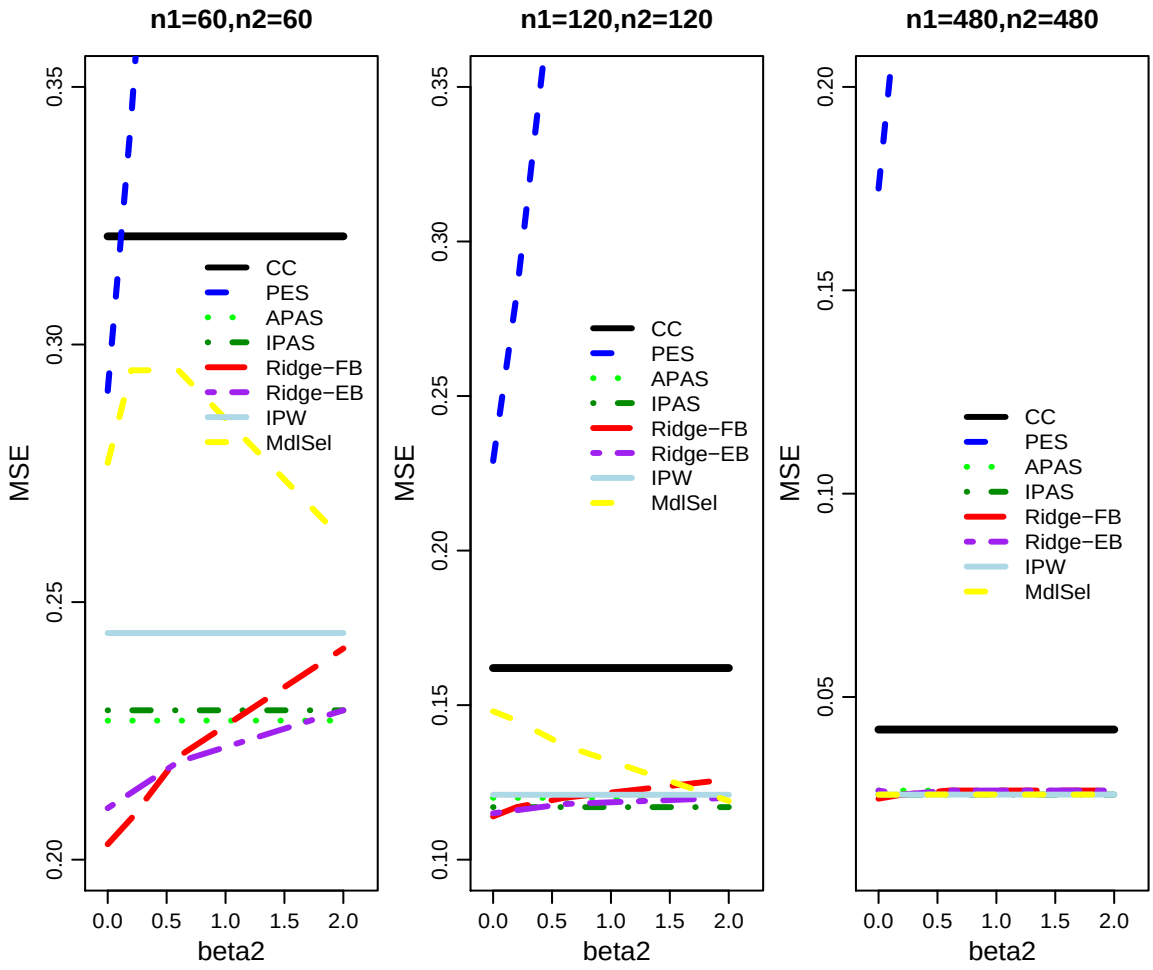

Figure 3.6: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 0.6$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2_0 = 0.333$, $\rho^2_0 = 0.561$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
$60, 60$	2	0	8	Bias	-0.3	7	-1227	21	44	4	15	33	32
				MSE	115	587	1954	369	283	264	269	327	300
				ESD	339	766	670	607	530	514	518	571	547
				SE	340	776	732	580	534	569	516	-	532
				CR	95.0	94.0	59.5	94.5	94.8	96.8	93.0	-	93.5
	2	0.2	8.2	Bias	-0.3	7	-1290	21	44	-4	10	33	35
				MSE	115	587	2126	369	283	264	270	327	299
				ESD	339	766	680	607	530	514	520	571	546
				SE	340	776	749	580	534	570	517	-	532
				CR	95.0	94.0	58.8	94.5	94.8	96.8	93.3	-	93.5
	2	0.6	8.6	Bias	-0.3	7	-1415	21	44	-19	0.5	33	38
				MSE	115	587	2495	369	283	265	274	327	294
				ESD	339	766	702	607	530	515	523	571	541
				SE	340	776	783	532	534	573	519	-	531
				CR	95.0	94.0	53.8	94.5	94.8	96.8	92.8	-	93.8
	2	2	10	Bias	-0.3	7	-1854	21	44	-55	-11	33	39
				MSE	115	587	4054	369	283	287	283	327	294
				ESD	339	766	786	607	530	533	532	571	541
				SE	340	776	909	580	534	590	526	-	531
				CR	95.0	94.0	43.0	94.5	94.8	96.3	93.3	-	93.8
$120, 120$	2	0	8	Bias	7	5	-1265	0.2	16	-8	8	15	16
				MSE	61	299	1840	197	147	141	144	158	148
				ESD	246	547	490	443	384	376	379	398	384
				SE	241	547	504	405	364	378	355	-	364
				CR	95.3	95.8	26.5	92.8	93.5	95.0	93.3	-	93.5
	2	0.2	8.2	Bias	7	5	-1329	0.2	16	-1	3	15	16
				MSE	61	299	2013	197	147	141	144	158	148
				ESD	246	547	496	443	384	376	380	398	384
				SE	241	547	515	405	364	379	355	-	364
				CR	95.3	95.8	24.0	92.8	93.5	94.8	93.5	-	93.5
	2	0.6	8.6	Bias	7	5	-1458	0.2	16	-19	-7	15	16
				MSE	61	299	2387	197	147	143	145	158	148
				ESD	246	547	510	443	384	377	381	398	384
				SE	241	547	538	405	364	381	356	-	364
				CR	95.3	95.8	21.8	92.8	93.5	94.5	93.3	-	93.5
	2	2	10	Bias	7	5	-1910	0.2	16	-52	-21	15	16
				MSE	61	299	3964	197	147	156	149	158	148
				ESD	246	547	563	443	384	392	385	398	384
				SE	241	547	624	405	364	397	362	-	364
				CR	95.3	95.8	13.5	92.8	93.5	96.0	92.5	-	93.5

Table 3.9: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.333$, ${\rho_1}^2 = 0.818$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
480, 480	2	0	8	Bias	-0.3	24	-1314	0.4	2	-2	-1	3	2
				MSE	15	79	1771	43	34	33	33	34	34
				ESD	124	280	213	208	184	181	182	185	184
				SE	120	269	250	199	178	184	174	-	178
				CR	94.3	92.8	0	94.5	93.8	95.0	93.0	-	93.8
	2	0.2	8.2	Bias	-0.3	24	1379	0.4	2	-13	-7	3	2
				MSE	15	79	1950	43	34	33	33	34	34
				ESD	124	280	216	208	184	181	183	185	184
				SE	120	269	255	199	178	185	175	-	178
				CR	94.3	92.8	0	94.5	93.8	94.8	92.8	-	93.8
	2	0.6	8.6	Bias	-0.3	24	-1511	0.4	2	-28	-15	3	2
				MSE	15	79	2333	43	34	35	34	34	34
				ESD	124	280	223	208	184	184	184	185	184
				SE	120	269	267	199	178	188	176	-	178
				CR	94.3	92.8	0	94.5	93.8	94.5	93.0	-	93.8
	2	2	10	Bias	-0.3	24	1972	0.4	0.2	-12	-8	3	2
				MSE	15	79	3951	43	34	35	34	34	34
				ESD	124	280	249	208	184	186	185	185	184
				SE	120	269	309	199	178	193	178	-	178
				CR	94.3	92.8	0	94.5	93.8	94.8	93.3	-	93.8

Table 3.10: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = 1$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2_0 = 0.333$, $\rho^2_1 = 0.818$ and $p = 0.8$.

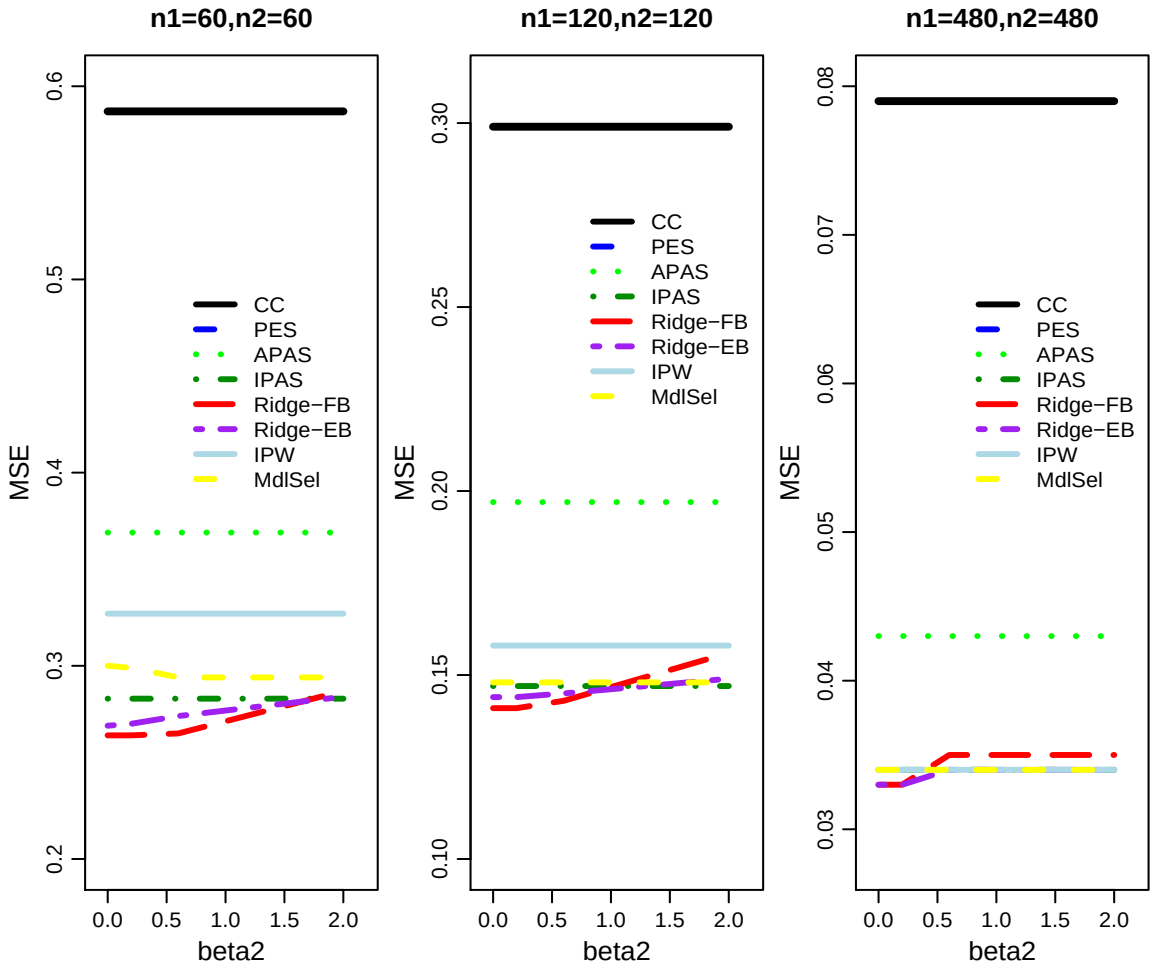

Figure 3.7: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_3 = 2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, $\rho^2_0 = 0.333$, $\rho^2_1 = 0.818$ and $p = 0.8$.

	Medicine	Surgery
IOP Observed at 12th and 102nd Month		
Number of Patients	86	74
IOP at 12th Month: Mean (SE)	17.9(3.29)	14.1(4.96)
IOP at 102nd Month: Mean (SE)	17.5(4.67)	15.1(4.61)
IOP Missing at 102nd Month		
Number of Patients	206	207
IOP at 12th Month: Mean (SE)	18.2(3.80)	14.3(5.19)

Table 3.11: Summary Statistics from CIGTS data. IOP at the 102th month as True Endpoint and IOP at the 12th month as Surrogate

Estimation Method	Estimate	95% CI	CI Width
CC	-2.391	$(-3.844, -0.937)$	2.907
IPAS	-2.419	$(-3.792, -1.046)$	2.746
APAS	-2.400	$(-3.765, -1.034)$	2.731
PES	-1.833	$(-2.490, -1.176)$	1.315
MdlSel	-1.833	$(-2.490, -1.176)$	1.315
Ridge-EB	-2.094	$(-3.138, -1.049)$	2.089
Ridge-FB	-2.019	$(-3.033, -1.006)$	2.027

Table 3.12: Quantity of Interest: Difference in the IOP Reduction at the 102nd Month between Surgery Treatment and Medicine Treatment. Estimates from Seven Methods are Presented here. IOP at the 102nd month as True Endpoint and IOP at the 12th month as Surrogate

3.8 Appendix

3.8.1 Asymptotic Variance

When S is an interactive partial surrogate, the elements of the expected information matrix are the second derivatives of the likelihood function in (3.2) and given

by:

$$
\begin{aligned}
-E\frac{\partial^2 logL}{\partial\beta_0^2} &= E\frac{r}{\sigma_t^2} = \frac{n(1-p)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_1^2} &= E\frac{\sum_{i=1}^{r} S_i^2}{\sigma_t^2} = \frac{n(1-p)(\sigma_s^2+\alpha_0^2+\alpha_0\alpha_1+0.5\alpha_1^2)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_2^2} &= E\frac{\sum_{i=1}^{r} Z_i^2}{\sigma_t^2} = \frac{0.5n(1-p)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_3^2} &= E\frac{\sum_{i=1}^{r} S_i^2 Z_i^2}{\sigma_t^2} = \frac{n(1-p)(0.5\sigma_s^2+0.5\alpha_0^2+\alpha_0\alpha_1+0.5\alpha_1^2)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} &= E\frac{\sum_{i=1}^{r} S_i}{\sigma_t^2} = \frac{n(1-p)(\alpha_0+0.5\alpha_1)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_2} &= E\frac{\sum_{i=1}^{r} Z_i}{\sigma_t^2} = \frac{0.5n(1-p)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_3} &= E\frac{\sum_{i=1}^{r} S_i Z_i}{\sigma_t^2} = \frac{0.5n(1-p)(\alpha_0+\alpha_1)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_2} &= E\frac{\sum_{i=1}^{r} S_i Z_i}{\sigma_t^2} = \frac{0.5n(1-p)(\alpha_0+\alpha_1)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_3} &= E\frac{\sum_{i=1}^{r} S_i^2 Z_i}{\sigma_t^2} = \frac{0.5n(1-p)[(\alpha_0+\alpha_1)^2+\sigma_s^2]}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\beta_2\partial\beta_3} &= E\frac{\sum_{i=1}^{r} S_i Z_i^2}{\sigma_t^2} = \frac{0.5n(1-p)(\alpha_0+\alpha_1)}{\sigma_t^2} \\
-E\frac{\partial^2 logL}{\partial\alpha_0^2} &= E\frac{n}{\sigma_s^2} = \frac{n}{\sigma_s^2} \\
-E\frac{\partial^2 logL}{\partial\alpha_1^2} &= E\frac{\sum_{i=1}^{n} Z_i^2}{\sigma_s^2} = \frac{0.5n}{\sigma_s^2} \\
-E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} &= E\frac{\sum_{i=1}^{n} Z_i}{\sigma_s^2} = \frac{0.5n}{\sigma_s^2}
\end{aligned}
$$

The expected information matrix $I_{IPAS}(\theta)$ is equal to

$$
\begin{pmatrix}
-E\frac{\partial^2 logL}{\partial\beta_0^2} & -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} & -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_2} & -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_3} & 0 & 0 \\
-E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} & -E\frac{\partial^2 logL}{\partial\beta_1^2} & -E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_2} & -E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_3} & 0 & 0 \\
-E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_2} & -E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_2} & -E\frac{\partial^2 logL}{\partial\beta_2^2} & -E\frac{\partial^2 logL}{\partial\beta_2\partial\beta_3} & 0 & 0 \\
-E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_3} & -E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_3} & -E\frac{\partial^2 logL}{\partial\beta_2\partial\beta_3} & -E\frac{\partial^2 logL}{\partial\beta_3^2} & 0 & 0 \\
0 & 0 & 0 & 0 & -E\frac{\partial^2 logL}{\partial\alpha_0^2} & -E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} \\
0 & 0 & 0 & 0 & -E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} & -E\frac{\partial^2 logL}{\partial\alpha_1^2}
\end{pmatrix}
$$

When S is an additive partial surrogate, i.e., $\beta_3 = 0$, the expected information matrix $I_{APAS}(\theta)$ is given by:

$$\begin{pmatrix} -E\frac{\partial^2 logL}{\partial\beta_0^2} & -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} & -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_2} & 0 & 0 \\ -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} & -E\frac{\partial^2 logL}{\partial\beta_1^2} & -E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_2} & 0 & 0 \\ -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_2} & -E\frac{\partial^2 logL}{\partial\beta_1\partial\beta_2} & -E\frac{\partial^2 logL}{\partial\beta_2^2} & 0 & 0 \\ 0 & 0 & 0 & -E\frac{\partial^2 logL}{\partial\alpha_0^2} & -E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} \\ 0 & 0 & 0 & -E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} & -E\frac{\partial^2 logL}{\partial\alpha_1^2} \end{pmatrix}$$

When S is a perfect surrogate, i.e., $\beta_2 = \beta_3 = 0$, the information matrix $I_{PES}(\theta)$ is expressed as:

$$\begin{pmatrix} -E\frac{\partial^2 logL}{\partial\beta_0^2} & -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} & 0 & 0 \\ -E\frac{\partial^2 logL}{\partial\beta_0\partial\beta_1} & -E\frac{\partial^2 logL}{\partial\beta_1^2} & 0 & 0 \\ 0 & 0 & -E\frac{\partial^2 logL}{\partial\alpha_0^2} & -E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} \\ 0 & 0 & -E\frac{\partial^2 logL}{\partial\alpha_0\partial\alpha_1} & -E\frac{\partial^2 logL}{\partial\alpha_1^2} \end{pmatrix}$$

3.8.2 Additional Simulation Results

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
60, 60	0.2	0	-3.4	Bias	6	48	-95	41	51	3	25	56	-46
				MSE	92	490	183	238	252	204	222	300	216
				ESD	303	699	417	487	500	452	471	545	462
				SE	306	693	422	497	512	507	478	-	431
				CR	94.0	94.5	94.8	96.3	96.5	97.3	97.0	-	93.8
	0.2	0.2	-3.2	Bias	6	48	-158	41	51	-18	14	56	-70
				MSE	92	490	197	238	252	207	224	300	233
				ESD	303	699	414	487	500	455	473	545	478
				SE	306	693	420	497	512	508	478	-	433
				CR	94.0	94.5	94.3	96.3	96.5	96.8	96.8	-	93.5
	0.2	0.6	-2.8	Bias	6	48	-283	41	51	-53	-4	56	-112
				MSE	92	490	249	238	252	217	229	300	273
				ESD	303	699	411	487	500	463	478	545	510
				SE	306	693	419	497	512	514	479	-	439
				CR	94.0	94.5	91.3	96.3	96.5	96.5	96.0	-	92.3
	0.2	2	-1.4	Bias	6	48	-722	41	51	-84	-17	56	-15
				MSE	92	490	697	238	252	254	243	300	301
				ESD	303	699	419	487	500	497	492	545	549
				SE	306	693	452	497	512	545	488	-	480
				CR	94.0	94.5	61.8	96.3	96.5	96.5	94.5	-	90.3
120, 120	0.2	0	-3.4	Bias	-4	-8	-115	3	5	-23	-10	4	-63
				MSE	48	262	105	137	138	119	128	146	130
				ESD	219	512	304	370	371	344	358	382	355
				SE	216	482	291	342	346	345	328	-	301
				CR	95.5	93.3	90.0	91.8	92.0	94.3	91.5	-	88.3
	0.2	0.2	-3.2	Bias	-4	-8	-179	3	5	-41	-21	4	-81
				MSE	48	262	124	137	138	124	131	146	145
				ESD	219	512	303	370	371	349	361	382	372
				SE	216	482	290	342	346	347	328	-	304
				CR	95.5	93.3	87.0	91.8	92.0	91.8	91.0	-	85.8
	0.2	0.6	-2.8	Bias	-4	-8	-308	3	5	-67	-36	4	-101
				MSE	48	262	186	137	138	135	135	146	177
				ESD	219	512	302	370	371	362	365	382	409
				SE	216	482	288	342	342	352	329	-	312
				CR	95.5	93.3	81.5	91.8	92.0	91.8	89.8	-	85.0
	0.2	2	-1.4	Bias	-4	-8	-759	3	5	-64	-42	4	-3
				MSE	48	262	674	137	138	148	141	146	147
				ESD	219	512	313	370	371	379	373	382	384
				SE	216	482	308	342	346	369	336	-	341
				CR	95.5	93.3	30.0	91.8	92.0	92.0	90.5	-	91.0

Table 3.13: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = -2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0.618$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
480, 480	0.2	0	-3.4	Bias	-14	-35	-143	-11	-11	-33	-27	-13	-58
				MSE	14	67	43	32	32	31	31	32	40
				ESD	118	256	150	178	178	172	174	180	192
				SE	108	242	144	170	170	174	163	-	156
				CR	92.0	92.5	82.0	94.8	94.5	94.5	93.3	-	86.5
	0.2	0.2	-3.2	Bias	-14	-35	-209	-11	-11	-42	-33	-13	-48
				MSE	14	66	67	32	32	33	32	32	43
				ESD	118	256	148	178	178	176	177	180	201
				SE	108	242	143	170	170	176	164	-	161
				CR	92.0	92.5	69.3	94.8	94.0	93.8	93.0	-	87.0
	0.2	0.6	-2.8	Bias	-14	-35	-341	-11	-11	-45	-37	-13	-15
				MSE	14	67	138	32	32	34	34	32	34
				ESD	118	256	147	178	178	180	179	180	183
				SE	108	242	143	170	170	180	166	-	169
				CR	92.0	92.5	34.5	94.8	94.0	94.3	93.8	-	93.0
	0.2	2	-1.4	Bias	-14	-35	-802	-11	-11	-24	-23	-13	-11
				MSE	14	67	664	32	32	33	32	32	32
				ESD	118	256	147	178	178	179	179	180	178
				SE	108	242	153	170	170	182	169	-	170
				CR	92.0	92.5	0	94.8	94.0	95.3	94.0	-	94.8

Table 3.14: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = -2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0.618$ and $p = 0.8$.

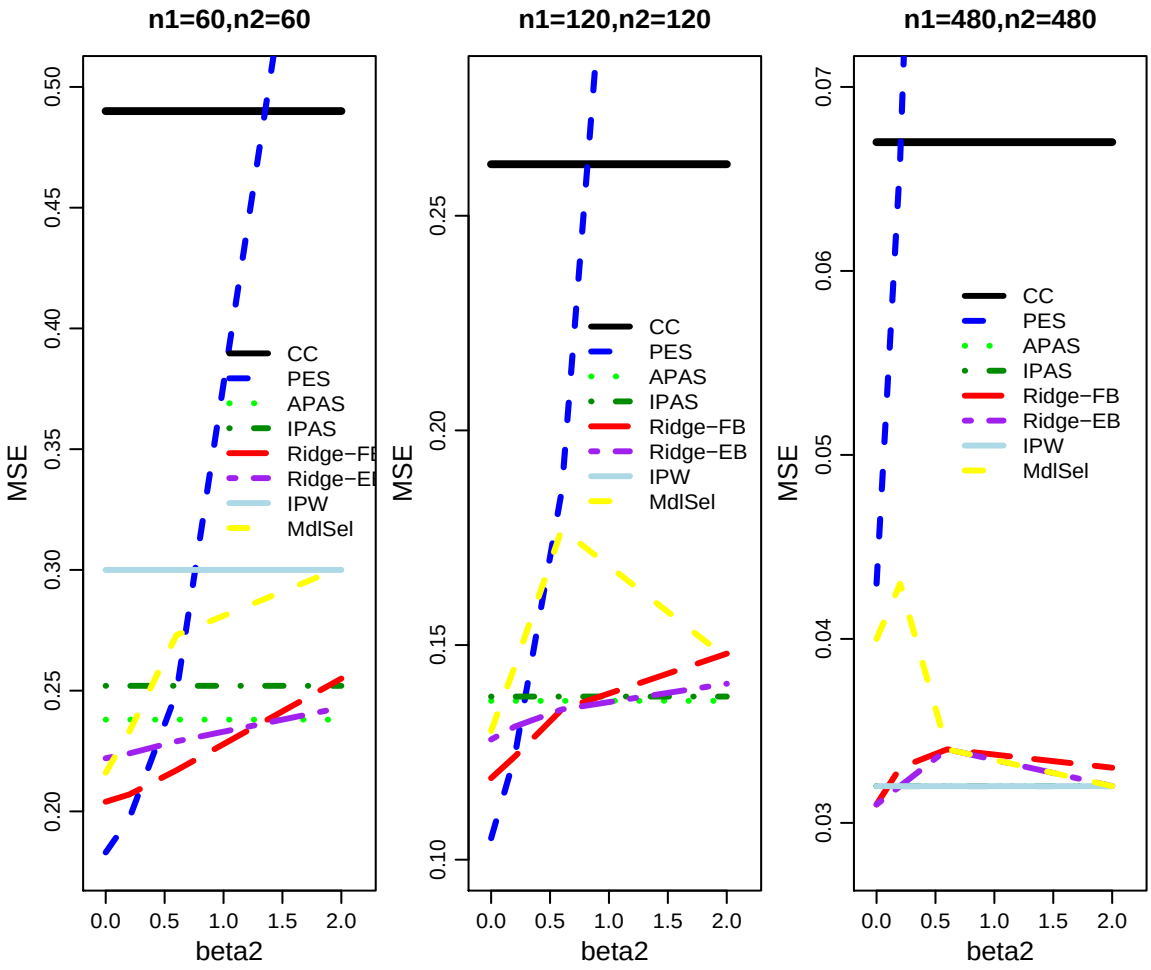

Figure 3.8: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = -2$, $\beta_3 = 0.2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0.618$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
60, 60	0.6	0	-2.2	Bias	5	42	-346	37	50	-39	3	54	-92
				MSE	81	439	285	235	242	210	221	286	280
				ESD	284	662	407	483	489	457	470	532	521
				SE	288	652	415	490	501	512	469	-	441
				CR	94.3	94.5	88.3	96.5	96.0	96.8	95.8	-	90.0
	0.6	0.2	-2	Bias	5	42	-409	37	50	-50	-4	54	-91
				MSE	81	439	333	235	242	215	223	286	290
				ESD	284	662	407	483	489	461	473	532	531
				SE	288	652	418	491	501	516	470	-	446
				CR	94.3	94.5	84.0	96.5	96.0	96.3	95.5	-	89.0
	0.6	0.6	-1.6	Bias	5	42	-534	37	50	-63	-14	54	-59
				MSE	81	439	453	235	242	227	228	286	290
				ESD	284	662	409	483	489	472	478	532	535
				SE	288	652	427	491	501	523	472	-	459
				CR	94.3	94.5	73.5	96.5	96.0	95.8	94.5	-	89.0
	0.6	2	-0.2	Bias	5	42	-973	37	50	-66	-18	54	21
				MSE	81	439	1139	235	242	247	237	286	273
				ESD	284	662	439	483	489	493	486	532	522
				SE	288	652	491	491	501	542	479	-	483
				CR	94.3	94.5	46.8	96.5	96.0	96.5	94.8	-	93.5
120, 120	0.6	0	-2.2	Bias	-3	-11	-372	0.05	5	-41	-20	5	-63
				MSE	43	238	230	137	133	132	132	140	172
				ESD	208	488	302	371	365	361	363	375	410
				SE	203	454	284	339	338	350	323	-	317
				CR	95.8	93.0	74.5	91.0	91.3	92.0	89.5	-	84.0
	0.6	0.2	-2	Bias	-3	-11	-437	0.05	5	-46	-26	5	-47
				MSE	43	238	283	137	133	135	133	140	169
				ESD	208	488	303	371	365	364	364	375	408
				SE	203	454	286	339	339	352	323	-	323
				CR	95.8	93.0	67.0	91.0	91.3	91.8	89.0	-	83.8
	0.6	0.6	-1.6	Bias	-3	-11	-566	0.05	5	-53	-35	5	-21
				MSE	43	23	414	137	133	138	134	140	154
				ESD	208	488	306	371	365	368	365	375	392
				SE	203	454	292	339	338	355	324	-	330
				CR	95.8	93.0	50.3	91.0	91.3	91.8	88.8	-	87.5
	0.6	2	-0.2	Bias	-3	-11	-1017	0.05	5	-57	-40	5	3
				MSE	43	23	1142	137	133	141	137	140	136
				ESD	208	488	328	371	365	371	367	375	369
				SE	203	454	334	339	338	362	332	-	336
				CR	95.8	93.0	13.5	91.0	91.3	92.5	89.8	-	90.0

Table 3.15: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = -2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0.495$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
480, 480	0.6	0	-2.2	Bias	-13	-32	-406	-11	-11	-22	-21	-12	-11
				MSE	12	60	185	31	30	30	30	31	31
				ESD	111	242	144	177	174	171	171	175	175
				SE	102	228	141	168	166	173	161	-	166
				CR	92.3	92.8	18.3	94.0	94.3	95	93.8	-	94.0
	0.6	0.2	-2	Bias	-13	-32	-471	-11	-11	-30	-26	-12	-11
				MSE	12	60	243	31	30	30	30	31	31
				ESD	111	242	144	177	174	171	172	175	175
				SE	102	228	142	168	166	174	161	-	166
				CR	92.3	92.8	8.5	94.0	94.3	94.5	93.5	-	94.0
	0.6	0.6	-1.6	Bias	-13	-32	-603	-11	-11	-40	-32	-12	-11
				MSE	12	60	384	31	30	32	31	31	31
				ESD	111	242	144	177	174	174	174	175	175
				SE	102	228	145	168	166	176	163	-	166
				CR	92.3	92.8	1.8	94.0	94.3	94.3	93.5	-	94.0
	0.6	2	-0.2	Bias	-13	-32	-1064	-11	-11	-28	-22	-12	-11
				MSE	12	60	1155	31	30	31	31	31	31
				ESD	111	242	150	177	174	174	175	175	175
				SE	102	228	165	168	166	181	165	-	166
				CR	92.3	92.8	0	94.0	94.3	95.0	93.8	-	94.0

Table 3.16: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = -2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0.495$ and $p = 0.8$.

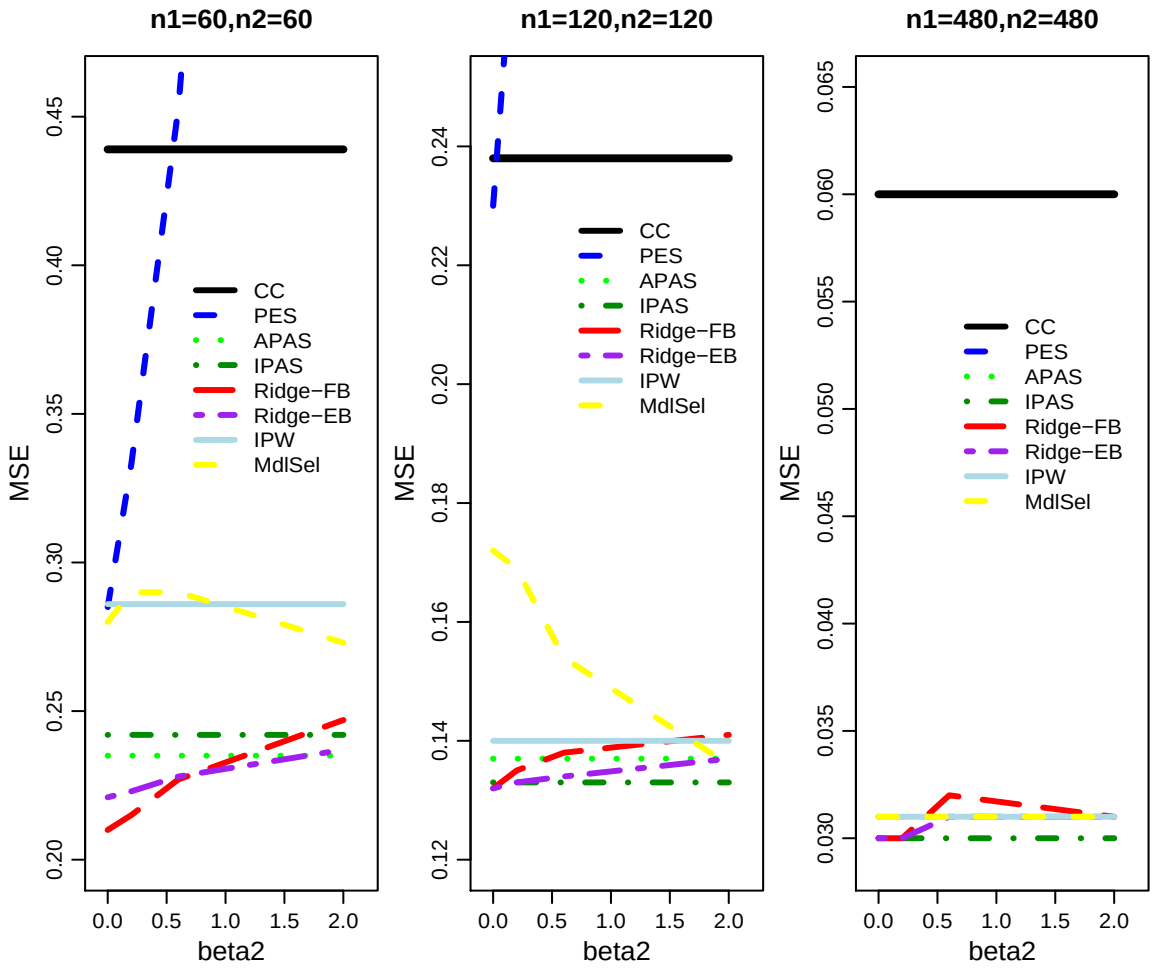

Figure 3.9: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = -2$, $\beta_3 = 0.6$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0.495$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
$60, 60$	2	0	2	Bias	1	22	-1225	22	46	5	17	47	34
				MSE	64	370	1808	307	227	209	211	268	249
				ESD	253	608	554	553	474	457	459	515	498
				SE	259	583	533	533	485	516	460	-	481
				CR	94.3	92.3	41.8	93.3	94.5	96.5	94.5		93.0
	2	0.2	2.2	Bias	1	22	-1288	22	46	-3	11	47	36
				MSE	64	370	1975	307	227	208	212	268	248
				ESD	253	608	562	553	474	456	460	515	496
				SE	259	583	601	533	485	517	461	-	481
				CR	94.3	92.3	39.8	93.3	94.5	96.3	94.5	-	93.3
	2	0.6	2.6	Bias	1	22	-1413	22	46	-18	2	47	40
				MSE	64	370	2334	307	227	208	214	268	242
				ESD	253	608	580	553	474	456	463	515	491
				SE	259	583	632	533	485	519	463	-	481
				CR	94.3	92.3	35.0	93.3	94.5	96.5	94.5	-	93.8
	2	2	4	Bias	1	22	-1852	22	46	-54	10	47	41
				MSE	64	370	3856	307	227	225	222	268	242
				ESD	253	608	653	553	474	471	471	515	491
				SE	259	583	751	533	485	535	469	-	481
				CR	94.3	92.3	27.5	93.3	94.5	97.0	93.3	-	93.8
$120, 120$	2	0	2	Bias	-2	-19	-1274	-9	6	-2	-1	6	6
				MSE	36	211	1788	179	129	123	126	134	129
				ESD	190	459	405	423	359	351	355	365	359
				SE	183	407	401	370	326	341	315		326
				CR	93.3	92.5	13.0	91.8	88.3	92.3	89.8	-	88.8
	2	0.2	2.2	Bias	-2	-19	-1339	-9	6	-11	-6	6	6
				MSE	36	211	1961	179	129	123	127	134	129
				ESD	190	459	410	423	359	350	356	365	359
				SE	183	407	411	370	326	341	315	-	326
				CR	93.3	92.5	11.5	91.8	88.8	92.3	89.8	-	88.8
	2	0.6	2.6	Bias	-2	-19	-1468	-9	6	-29	-17	6	6
				MSE	36	211	2332	179	129	124	127	134	129
				ESD	190	459	422	423	359	351	357	365	359
				SE	183	407	432	370	326	343	316	-	326
				CR	93.3	92.5	9.3	91.8	88.8	92.8	89.8	-	88.8
	2	2	4	Bias	-2	-19	-1919	-9	6	-62	-31	6	6
				MSE	36	211	3901	179	129	136	130	134	129
				ESD	190	459	467	423	359	364	360	365	359
				SE	183	407	512	370	326	357	321	-	326
				CR	93.3	92.5	4.5	91.8	88.8	92.5	88.8	-	88.8

Table 3.17: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = -2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma_{ss}^2 = 0.5$, $\sigma_t^2 = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0$ and $p = 0.8$.

n_1, n_2	β_3	β_2	Q		ALL	CC	PES	APAS	IPAS	Ridge FB	Ridge EB	IPW	Mdl-Sel
480, 480	2	0	2	Bias	-10	-22	-1323	-9	-7	-12	-11	-9	-7
				MSE	10	48	1782	39	28	27	27	28	28
				ESD	98	219	176	196	168	163	165	168	168
				SE	91	204	198	183	160	166	155	-	160
				CR	93.5	93.5	0	92.3	93.3	94.5	93.8	-	93.3
	2	0.2	2.2	Bias	-10	-22	-1389	-9	-7	-23	-17	-9	-7
				MSE	10	48	1961	39	28	27	28	28	28
				ESD	98	219	179	196	168	163	165	168	168
				SE	91	204	203	183	160	167	155	-	160
				CR	93.5	93.5	0	92.3	93.3	94.8	93.5	-	93.3
	2	0.6	2.6	Bias	-10	-22	-1521	-9	-7	-42	-24	-9	-7
				MSE	10	48	2346	39	28	29	28	28	28
				ESD	98	219	184	196	168	165	167	168	168
				SE	91	204	214	183	160	170	157	-	160
				CR	93.5	93.5	0	92.3	93.3	94.3	93.8	-	93.3
	2	2	4	Bias	-10	-22	-1982	-9	-7	-22	-18	-9	-7
				MSE	10	48	3969	39	28	29	29	28	28
				ESD	98	219	205	196	168	169	168	168	168
				SE	91	204	253	199	160	176	159	-	160
				CR	93.5	93.5	0	92.3	93.3	96.0	93.5	-	93.3

Table 3.18: Results from 400 simulated data. Bias, MSE, ESD and SE are Reported by Multiplication of 1000. $\beta_0 = 0.5$, $\beta_1 = -2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0$ and $p = 0.8$.

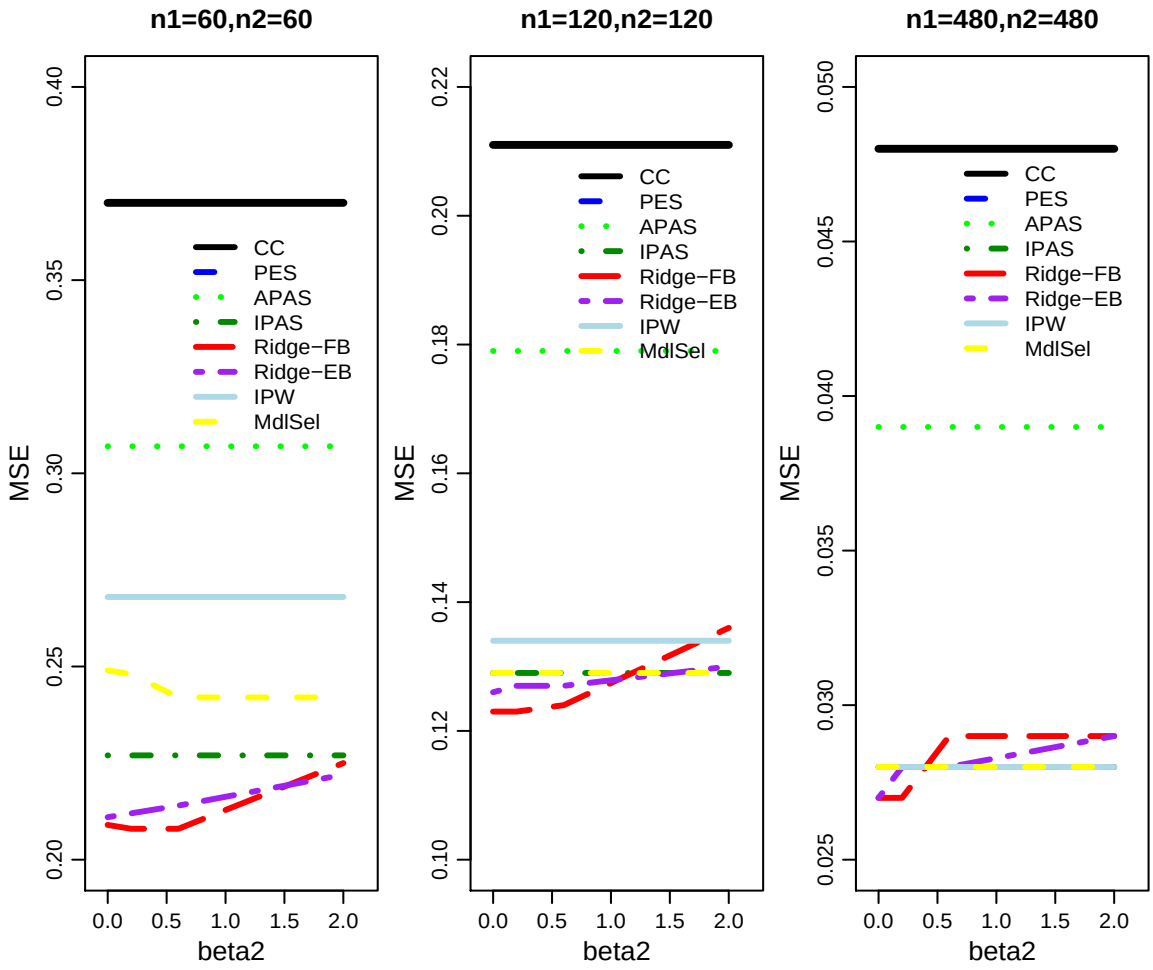

Figure 3.10: MSE by Sample Size and β_2 from 400 simulated data sets. $\beta_0 = 0.5$, $\beta_1 = -2$, $\beta_3 = 2$, $\alpha_0 = 1$, $\alpha_1 = 2$, $\sigma^2_{ss} = 0.5$, $\sigma^2_t = 1$, ${\rho_0}^2 = 0.667$, ${\rho_1}^2 = 0$ and $p = 0.8$.

3.9 References

Albert P.S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the Case-only Design for Identifying Gene-Environment Interactions. *American Journal of Epidemiology*. **154**, 687-693.

Baker S.G., Izmirlian G., and Kipnis V. (2005). Resolving paradoxes involving surrogate endpoints. *Journal of Royal Statistical Society A*. **168**, 753-762.

Begg C.B., and Leung D.H. (2000). On the use of surrogate end points in randomized trials. *Journal Of The Royal Statistical Society Series A*. **163**, 15-28.

Cook R.J., and Lawless J.F. (2001). Some comments on efficiency gains from auxiliary information for right-censored data. *Journal of Statistical Planning and Inference*. **96**, 191-202.

Day, N.E. and Duffy, S.W. (1996). Trial design based on surrogate end points : application to comparison of different breast screening frequencies. *Journal of the Royal Statistical Society. Series A. Statistics in society*. **1996**, 49-60.

Finkelstein, D.M., and Schoenfeld, D.A., (1994). Analyzing survival in the presence of an auxiliary variable. *Statistics in Medicine*. **13**, 1747-1754.

Flandre, P. (1996). On the use of auxiliary data to estimate the sruvival function and its variance: an application to acquired immunodeficiency syndrome. *Journal of Clinical Epidemiology*. **49**, 899-905.

Fleming T.R., DeMets D.L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*. **125**, 605-613.

Fleming, T.R., Prentice, R.L., Pepe, M.S. and Glidden, D., 1994. Surrogate and auxiliary endpoints in clinical trials with potential applications in cancer and AIDS

research. *Statistics in Medicine*. **13**, 955-968.

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*. **47**, 663-685.

Kosorok, M.R. and Fleming, T.R. (1993). Using surrogate failure time data to increase cost effectiveness in clinical trials. *Biometrika*. **80**. 823-833.

Lagakos, S.M. (1977). Using auxiliary variables for improved estimates of survival time. *Biometrics*. **33**, 399-404.

Little R.J.A and Rubin D.B. (2002). Statistical Analysis with Missing Data, 2nd Edition. Wiley: New York.

Louis, T.A., and Zelterman, D. (1994). Bayesian approaches to research synthesis. In H. Cooper, and L. V. Hedges (Eds.), The handbook of research synthesis. New York: Russell Sage Foundation.

Malani, H.M. (1995). A modification of the re-distribution to the right algorithm using disease markers. *Biometrika*. **82**, 515-526.

Murray, S. and Tsiatis, A.A. (1996). Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*. **52**, 137-151.

Musch D.C., Lichter P.R., Guire K.E., Standardi C.L., CIGTS Investigators (1999): The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology*. **106**: 653-62.

Pepe, M.S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*. **79**, 355-365.

Pepe, M.S., Reilly, M. and Fleming, T.R. (1994). Auxiliary outcome data and the

mean score method. *J. Statist. Planning Inf.* **42**, 137-160.

Prentice R.L. (1989). Surrogate endpoints in clinical trials, definition and operational criteria. *Statistics in Medicine*. **8**, 431-440.

Robins, J.M. Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*. **89**, 846–866.

Scharfstein D.O., Rotnitzky A, and Robins J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association*. **94**, 1096-1120.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*. **58**, 267-288.

Venkatraman E.S., and Begg C.B. (1999). Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics*. **55**, 1171-1176.

Zhao L.P., and Lipsitz S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*. *11*. 769-82.

Zhao L.P., Lipsitz S, and Lew D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics*. **52**, 1165-82.

CHAPTER IV

Assessing Surrogacy in Clinical Trials Using Counterfactual Models

Summary. A surrogate marker (S) is a variable that can be measured earlier and often easier than the true endpoint (T) in a clinical trial. It can be very useful if it can reliably facilitate early prediction of the effect of the treatment (Z) on T. Most previous research has been devoted to developing surrogacy measures to quantify how well S can replace T or examining the use of S in predicting the treatment effect. However, the research often requires one to fit models for the distribution of T given S and Z. It is well known that such models do not have causal interpretations because the models condition on a post-randomization variable S. In this paper, we directly model the relationship among T, S and Z in a causal inference framework, specifically using a potential outcomes framework introduced by Frangakis and Rubin (2002) for surrogate markers. We propose a Bayesian estimation method to evaluate the causal probabilities associated with the cross-classification of the potential outcomes of S and T when S and T are both binary. We use a log-linear model to model the odds ratios of the potential outcomes. The quantities derived from this approach always have causal interpretations. This causal model is not identifiable from data without additional assumptions. To reduce the non-identifiability problem and increase the precision for statistical inferences, we incorporate assumptions that are plausible in

the surrogate context by using prior distributions. We also explore the relationship among the surrogacy measures based on traditional models and this counterfactual model. We use the causal probabilities to predict the treatment effect when T is partially observed. Then we extend the method to the multiple trial setting using hierarchical modeling. The methods are applied to data from a glaucoma treatment study and a colorectal cancer study.

Keywords: Bayesian Estimation, Counterfactual Model, Randomized Trial, Surrogate Marker.

4.1 Introduction

Surrogate markers (S) in a randomized clinical trial are intermediate physical or laboratory indicators of a disease progression process that can be measured earlier and are often easier to collect than the true endpoint (T). A good surrogate marker will have a strong association with T. When T is rare, late-occurring or costly to obtain, we can use an effective surrogate marker to reliably extract information on the effect of the treatment (Z) on T before T is completely observed. Thus a surrogate marker can have enormous potential benefits in reducing trial duration and size, lowering the trial expense and leading to earlier decision making. Examples of potential surrogate markers include CD4 counts and viral load for HIV infection, blood pressure and serum cholesterol level for cardiovascular disease and prostate-specific antigen for prostate cancer. As more biomarkers are discovered and recommended as surrogate markers, there is continuing strong interest in surrogate markers in the clinical research community.

In order to fully realize the potential use of surrogate markers in predicting treatment effects, biological research has been conducted to understand the mechanism through which the treatment functions. Statistical methods have been proposed to quantify the value of a particular surrogate marker and complement the biological research in the hope of understanding the relationship among T, S and Z. Prentice (1989) proposed a formal definition of perfect surrogacy which requires that S fully captures the effect of the treatment on T. To measure less than perfect surrogacy, the proportion of the treatment effect explained by S was proposed by Freedman *et al* (1992) and further extended by Wang and Taylor (2002). However, these measures and the prediction of the treatment effect on T using S often require one to

utilize models for the distribution of T given S and Z. They often do not have causal interpretations because the models used condition on the post randomization variable S (Rosenbaum, 1984). Other surrogacy measures include the trial-level and individual-level correlations between S and T in a multiple-trial setting (Buyse *et al*, 2000) and measures based on entropy (Alonso *et al.*, 2003).

An alternative approach is to directly model the relationship among T, S and Z in a causal framework. The general idea in causal modeling hypothesizes the setting wherein each individual has two potential outcomes, corresponding to the two possible treatment regimes (e.g., $Z = 1$ for treatment and $Z = 0$ for placebo). The causal treatment effect would be the comparison between these two potential outcomes for the same set of individuals. In reality, we only observe one of the outcomes since either the treatment or placebo (not both) is assigned to a patient. The framework has been used to model noncompliance (Imbens and Rubin, 1997; Frangakis *et al*, 2002; Hirano *et al*, 2000; Balke and Pearl, 1997). In this article, we use the terms, counterfactual model, causal model and potential outcomes model exchangeably.

We adopt a causal framework to study surrogacy through a principal stratification approach introduced by Frangakis and Rubin (2002). The idea is to examine the distribution of the potential outcomes of T with respect to Z within each principal stratum, which is defined by each pair of possible realizations of the potential outcomes of S. Since the principal strata cannot be changed by treatment, they can be adjusted for as a pre-randomization variable. This approach has been investigated by Taylor *et al* (2005) to study the causal interpretation of the surrogacy measure developed by Freedman *et al* (1992). When both S and T are binary, the potential outcomes for S and T are denoted by $(S(Z) = 0, 1)$ and $(T(Z) = 0, 1)$ with respect

to Z. We can study the causal association between S and T through the causal probabilities associated with the combinations of different sequence of potential outcomes of S and T for each individual. We can evaluate the surrogate marker based on the degree to which the causal effect of Z on S is reflected by the causal effect of Z on T. In contrast to Prentice (1989) and Freedman *et al* (1992) criteria, these association measures and the quantities derived always have causal interpretations. Since we only observe one set of the potential outcomes, these probabilities are not fully identifiable from the data.

While Frangakis and Rubin (2002) laid out a causal framework for studying surrogate markers, there has been little work on estimation methods for this causal model. An exception to this is the paper of Gilbert and Hudgens (2007), where their context of a HIV vaccine trial allowed them to make strong assumptions. In this paper, we propose a Bayesian estimation method to evaluate the causal probabilities. In practice, data are collected in a scientific context and with a considerable amount of a priori knowledge. We incorporate our prior knowledge by imposing appropriate prior distributions and placing some reasonable constraints on the model parameters. We hope to reduce the non-identifiability problem and increase the precision for the statistical inference of interest by introducing prior beliefs. For example, we could explore the assumptions such as that S and T are closely associated and the pair $(S(0), S(1))$ more likely than not agrees with the pair $(T(0), T(1))$. Our contribution to the counterfactual literature is that we focus on directly modeling the association of the pairs of potential outcomes of S and T under an underlying ordering constraint between $(S(0), S(1))$ and $(T(0), T(1))$.

In Section 4.2, we introduce a real data example from the Collaborative Initial Glaucoma Treatment Study (CIGTS) to which we will apply our methods. In Section

4.3, we first explore the assumptions that are necessary to help identify the parameters and then introduce a Bayesian estimation method for the single-trial setting. In Section 4.4, we apply the proposed method to the glaucoma data and examine the sensitivity of the priors. In Section 4.5, we evaluate our estimation algorithm through simulations. In Section 4.6, we explore the connections among the surrogacy measures based on conventional models and the counterfactual model. In Section 4.7, we use the causal probabilities to predict the treatment effect when T is partially observed. In Section 4.8, we extend the method to a multiple-trial setting where we assume the relationship between the counterfactual S and T is the same across trials conditioning on the trial-specific principal strata. Finally, we summarize our findings and provide discussion.

4.2 Glaucoma Treatment Study

We begin by considering a single randomized clinical trial setting. We will apply the proposed method to data from the Collaborative Initial Glaucoma Treatment Study (CIGTS) (Musch *et al.*, 1999). Glaucoma is a group of diseases that cause vision loss and is a leading cause for blindness. Elevated pressure in the eyes (i.e., intraocular pressure, IOP), is a major risk factor of glaucoma. The CIGTS is a randomized trial to compare the effects of two types of treatments, surgery ($Z = 1$) and medicine ($Z = 0$), on reducing IOP among glaucoma patients. Patients were enrolled between 1993 and 1997. IOP (recorded in mmHg) has been measured at different time points following randomization. For the purpose of this paper we take the true endpoint to be the IOP measurements at the 96th month and the surrogate marker to be the IOP at the 12th month. Both S and T are defined as 1 if IOP is less than 18mmHg and 0 otherwise. Due to drop out, there are fewer patients at the

96th month than at the 12th month. A total of 228 patients have IOP measured at both month 12 and 96, and 345 patients measured only at month 12. Table 4.1 lists the number of patients and the outcomes.

4.3 Methods

4.3.1 Potential Outcomes Model and Quantities of Interest

In the counterfactual framework, for each subject i, we have two potential outcomes for each of S_i and T_i, with one potential outcome observed and the other unobserved, denoted by $S_i(Z)$ and $T_i(Z)$ with respect to the treatment option Z. The possible realizations of the potential outcomes $(S_i(0), S_i(1))$ are $(0,0)$, $(0,1)$, $(1,1)$ and $(1,0)$ and similarly for $(T_i(0), T_i(1))$. The counterfactual probabilities that are associated with the combinations of different sequences of potential outcomes for S_i and T_i are listed in Table 4.2. Note that the probabilities sum to 1 as the 16 cells are a partition of the population. Each of these probabilities is of interest because collectively they completely describe the causal relationship among T, S and Z. Frangakis and Rubin (2002) also proposed the concepts of *associative* and *dissociative* effects to evaluate the strength of the connection between the causal treatment effect on T and the causal treatment effect on S. If the causal treatment effect on T_i is reflected on the changes in S_i, the effect is *associative*. Conversely, if the causal treatment effect on outcome T_i is not in the same direction as the effect on S_i, it is *dissociative*. To evaluate the degree of surrogacy, Taylor *et al* (2005) defined associative proportion (AP) as the ratio of the associative effect relative to the total causal treatment effect.

4.3.2 Assumptions

Since one of the potential outcomes is unobserved, the counterfactual model defined above is overparameterized. To assist in the identifiability of the model, we make the following assumptions.

Ignorability of Treatment Assignment (Rubin, 1978)

This assumption requires that the patients are comparable in both treatment groups. Due to randomization, the potential outcomes of S_i and T_i are independent of the actual treatment assignment Z_i. Thus, the principal stratum each patient belongs to is not impacted by the treatment assignment, i.e., $(S_i(0), S_i(1)) \perp Z_i$.

Stable Unit Treatment Value Assumption (Rubin, 1980)

This assumption implies that the potential outcomes of each person are independent of other individuals' treatment assignments, i.e., if $i \neq i'$, $S_i, ..., T_i \perp Z_{i'}$.

Monotonicity Assumption

Angrist, Imbens and Rubin (1996) borrowed the monotonicity assumption used in the instrumental variable approach and applied it to causal inference problems. In the surrogacy setting, under this assumption, a patient who received $Z = 1$ does not become worse off than that patient if he or she received $Z = 0$. Assume $S = 1$ and $T = 1$ represent better outcomes than $S = 0$ and $T = 0$, respectively. The monotonicity assumption requires that $S_i(1) \geq S_i(0)$ and $T_i(1) \geq T_i(0)$ for all i; hence, we cannot observe $(S_i(0) = 1, S_i(1) = 0)$ and $(T_i(0) = 1, T_i(1) = 0)$ and the number of free parameters is reduced from 15 to 8 as shown in Table 4.3. Since our data can support six parameters, as the probabilities $(p(T = t, S = s|Z))$ within each treatment group add up to 1, only some of the parameters in Table 4.3 or certain parameter combinations are identifiable.

Under the monotonicity assumption, the associative effect is p_{22} and the dissociative effect is $p_{12} + p_{32}$. The overall causal treatment effect is $p_{12} + p_{22} + p_{32}$. The associative proportion is $p_{22}/(p_{12} + p_{22} + p_{32})$. In a randomized trial, the overall causal treatment effect is directly estimable from the data but the associative effect p_{22} is not.

Ordering Restriction

In clinical trials, we collect information on surrogate markers based on prior scientific knowledge. In many cases that the surrogate marker is closely related to the true endpoint, possibly because the marker is in the causal pathway leading to disease. Hence, we assume the potential outcomes $(S(0), S(1))$ are more likely to agree with the potential outcomes $(T(0), T(1))$ than not. We assume there is an ordering in the sequence of the values of the potential outcomes: $(0, 0)$, $(0, 1)$, and $(1, 1)$ which we describe as "non-responsive", "responsive", and "always responsive", respectively. The assumed close relationship between S and T implies that when S is non-responsive (responsive), T is also more likely to be non-responsive (responsive). Similarly it is unlikely that a person will be non-responsive for S and always responsive for T. One way of reflecting this positive association between the ordered potential outcome pairs is to place restriction on the probabilities in Table 4.3 by imposing appropriate prior distributions. We expect this to not only overcome some of the identifiability problems but also improve the precision of statistical inferences.

4.3.3 Observed Data, Complete Data and Likelihood

We first consider the situation where we have the information on Z, S and T on every patient. The observed data include one of the two potential outcomes of S and T corresponding to the treatment one patient received. Suppose we have

$i = 1, ..., r$ patients. Let r_0 denote the number of patients in the $Z = 0$ group and r_1 the number of patients in the $Z = 1$ group. Let r_{zst} denote the number of patients for the combination of Z, S and T. The underlying probabilities associated with the cross-tabulations of Z, S and T can be expressed in terms of the counterfactual probabilities as in Table 4.4.

The observed-data likelihood function is given by

$$\begin{aligned} L_{obs} &= (p_{11} + p_{12} + p_{21} + p_{22})^{r_{000}}(p_{13} + p_{23})^{r_{001}}(p_{31} + p_{32})^{r_{010}}p_{33}^{r_{011}} \\ &\quad p_{11}^{r_{100}}(p_{12} + p_{13})^{r_{101}}(p_{21} + p_{31})^{r_{110}}(p_{22} + p_{23} + p_{32} + p_{33})^{r_{111}}. \end{aligned}$$

The complete data consists of all potential outcomes for both S and T. Let n_{jk} denote the cell count corresponding to the counterfactual probability in the cell (j, k) for the jth row and the kth column of Table 4.3 for all patients and n_{jk}^z for the treatment group z. The complete data likelihood is

$$\begin{aligned} L_{com} &= p_{11}^{n_{11}^0+n_{11}^1}p_{12}^{n_{12}^0+n_{12}^1}p_{13}^{n_{13}^0+n_{13}^1}p_{21}^{n_{21}^0+n_{21}^1}p_{22}^{n_{22}^0+n_{22}^1}p_{23}^{n_{23}^0+n_{23}^1}p_{31}^{n_{31}^0+n_{31}^1}p_{32}^{n_{32}^0+n_{32}^1}p_{33}^{n_{33}^0+n_{33}^1} \\ &= p_{11}^{n_{11}}p_{12}^{n_{12}}p_{13}^{n_{13}}p_{21}^{n_{21}}p_{22}^{n_{22}}p_{23}^{n_{23}}p_{31}^{n_{31}}p_{32}^{n_{32}}p_{33}^{n_{33}}. \end{aligned}$$

There is a one-to-one or many-to-one correspondence between n_{jk}'s and r_{zst}'s. For example, $n_{33}^0 = r_{011}$, $n_{11}^1 = r_{100}$, and $n_{13}^0 + n_{23}^1 = r_{001}$.

Since we only observe one of the potential outcomes, the counterfactual model contains more parameters than the number of independent observations. We adopt a Bayesian approach to incorporate sensible priors to reflect our prior beliefs. We treat the unobserved potential outcomes as missing data and estimate them via imputation.

4.3.4 The Model

Let $S^* = 1, 2, 3$ denote the ordered categories of $(S(0), S(1))$: $(0, 0)$, $(0, 1)$ and $(1, 1)$ and $T^* = 1, 2, 3$ denote the categories of $(T(0), T(1))$. We consider a log-linear

model to model the cell counts corresponding to the causal probabilities in Table 4.3. Let $E(n^z_{jk}) = \mu_{jk}$, $j = 1, 2, 3$ and $k = 1, 2, 3$. The model is specified as

$$\log\mu_{jk} = \lambda + \lambda_{jS} + \lambda_{kT} + \lambda_{jk}, \tag{4.1}$$

where, λ_{jS} and λ_{kT} denote the row and column variables, respectively. For identifiability of the log-linear model, we require constraints ($\lambda_{2S} = \lambda_{2T} = \lambda_{j2} = \lambda_{2k} = 0$) which give nice and simple expressions for the log odds ratios (OR) in the four 2×2 subtables in the four corners of Table 4.3:

$$\begin{aligned}
\log(OR_1) &= \log\left(\frac{\mu_{11} \times \mu_{22}}{\mu_{12} \times \mu_{21}}\right) = \lambda_{11}, \\
\log(OR_2) &= \log\left(\frac{\mu_{12} \times \mu_{23}}{\mu_{13} \times \mu_{22}}\right) = -\lambda_{13}, \\
\log(OR_3) &= \log\left(\frac{\mu_{21} \times \mu_{32}}{\mu_{22} \times \mu_{31}}\right) = -\lambda_{31}, \\
\log(OR_4) &= \log\left(\frac{\mu_{22} \times \mu_{33}}{\mu_{23} \times \mu_{32}}\right) = \lambda_{33}.
\end{aligned}$$

A positive association between S^* and T^* implies that λ_{11} and λ_{33} are positive and λ_{13} and λ_{31} are negative. This parametrization allows us to model the association between the potential outcomes of S and T directly. There is a close relationship between the multinomial and Poisson loglinear models (Birch, 1963; Lindley, 1964; Forster, 1996). Conditional on the total counts, we can express the causal probabilities using the model parameters in (4.1) as:

$$p_{jk} = \frac{\exp(\lambda_{jS} + \lambda_{kT} + \lambda_{jk})}{\sum_j \sum_k \exp(\lambda_{jS} + \lambda_{kT} + \lambda_{jk})},$$

where, $j = 1, 2, 3$ and $k = 1, 2, 3$.

4.3.5 Prior Specifications

Since λ_{1S}, λ_{3S}, λ_{1T}, λ_{3T} and λ are identifiable quantities which estimate the relative row effects and column effects, we choose "vague" priors for these variables.

Specifically, we let

$$p(\exp(\lambda_{jS})) = \text{gamma}(a, b), \quad j = 1, 3; \quad p(\exp(\lambda_{kT})) = \text{gamma}(a, b), \quad k = 1, 3;$$
$$p(\exp(\lambda)) = \text{gamma}(a, b)$$

where the parametrization of gamma(a, b) is defined such that the expected value is ab and the variance is ab^2. We chose nearly noninformative values of $a = 0.01$ and $b = 100$.

For λ_{11}, λ_{13}, λ_{31} and λ_{33}, there is much less information provided from the data, so for these we use mildly informative priors. To encourage but not force the ordering restriction, we use the priors to suggest positive associations between the potential outcomes of S and T:

$$p(\lambda_{11}) = \text{normal}(u, v^2), \;\; p(\lambda_{13}) = \text{normal}(-u, v^2),$$
$$p(\lambda_{31}) = \text{normal}(-u, v^2), \;\; p(\lambda_{33}) = \text{normal}(u, v^2),$$

where the parameterizations of the normal(u, v^2) give the mean of u and the variance of v^2. An example of the prior parameters can be: $u = 0.7$ and $v^2 = 1.96$. This induces the relationship that $E(OR_1) > 1$, $E(OR_2) > 1$, $E(OR_3) > 1$ and $E(OR_4) > 1$ to encourage the positive association between potential outcomes S_k^* and T_k^*. With the prior specification of normal$(0.7, 1.96)$, the median is 2.0 and the 95% probability interval is $(0.2, 20)$ for all four odds ratios, which provides a reasonably wide range. We also consider gamma prior distributions on these parameters for conjugacy reasons but they gave posterior distributions with undesirable properties.

4.3.6 Estimation Procedure

We use a data augmentation method (Little and Rubin, 2002) to estimate the parameters. This method regards the missing data as parameters. To simplify

the notation, we use $r_{obs} = \{r_{000}, r_{001}, r_{010}, r_{011}, r_{100}, r_{101}, r_{110}, r_{111}\}$ for the observed data and $\theta = (\lambda, \lambda_{jS}, \lambda_{kT}, \lambda_{jk})$ for all parameters. The complete data cell counts are denoted by $n_{com} = \{n^z_{11}, n^z_{12}, n^z_{13}, n^z_{21}, n^z_{22}, n^z_{23}, n^z_{31}, n^z_{32}, n^z_{33}\}$. To implement this procedure, we iterate the following I-step and P-step until convergence:

I-step: This step consists of distributing the observed counts into the cells of the counterfactual model in Table 4.3. Given θ^{l-1} and r_{obs}, we impute n^{0l}_{11}, n^{0l}_{12}, n^{1l}_{21}, n^{0l}_{22}, n^{0l}_{12}, n^{0l}_{13}, n^{1l}_{21}, n^{1l}_{22}, n^{1l}_{23}, n^{1l}_{32}, n^{1l}_{33} and n^{0l}_{31} where, n^{0l}_{11}, is the draw of the count that contributes to n_{11} from r_{000} from the lth iteration, n^{1l}_{12} is the draw of the count that contributes to n_{12} from r_{101} from the lth iteration, and so on. Let $\omega^{l-1}_1 = p^{l-1}_{11} + p^{l-1}_{12} + p^{l-1}_{21} + p^{l-1}_{22}$ and $\omega^{l-1}_2 = p^{l-1}_{22} + p^{l-1}_{23} + p^{l-1}_{32} + p^{l-1}_{33}$.

1. $(n^{0l}_{11}, n^{0l}_{12}, n^{0l}_{21}, n^{0l}_{22}) \sim \text{Multi}\left(r_{000}, \frac{p^{l-1}_{11}}{\omega^{l-1}_1}, \frac{p^{l-1}_{12}}{\omega^{l-1}_1}, \frac{p^{l-1}_{21}}{\omega^{l-1}_1}, \frac{p^{l-1}_{22}}{\omega^{l-1}_1}\right)$
2. $n^{1l}_{12} \sim \text{Bin}\left(r_{101}, \frac{p^{l-1}_{12}}{p^{l-1}_{12} + p^{l-1}_{13}}\right)$
3. $n^{0l}_{13} \sim \text{Bin}\left(r_{001}, \frac{p^{l-1}_{13}}{p^{k-1}_{13} + p^{l-1}_{23}}\right)$
4. $n^{1l}_{21} \sim \text{Bin}\left(r_{110}, \frac{p^{l-1}_{21}}{p^{k-1}_{21} + p^{l-1}_{31}}\right)$
5. $(n^{1l}_{22}, n^{1l}_{23}, n^{1l}_{32}, n^{1l}_{33}) \sim \text{Multi}\left(r_{111}, \frac{p^{l-1}_{22}}{\omega^{l-1}_2}, \frac{p^{l-1}_{23}}{\omega^{l-1}_2}, \frac{p^{l-1}_{32}}{\omega^{l-1}_2}, \frac{p^{l-1}_{33}}{\omega^{l-1}_2}\right)$
6. $n^{0l}_{31} \sim \text{Bin}\left(r_{010}, \frac{p^{l-1}_{31}}{p^{l-1}_{31} + p^{l-1}_{32}}\right)$
7. $n^l_{11} = n^{0l}_{11} + r_{100}; \quad n^l_{12} = n^{0l}_{12} + n^{1l}_{12}; \quad n^l_{13} = n^{0l}_{13} + r_{101} - n^{1l}_{12};$
8. $n^l_{21} = n^{0l}_{21} + n^{1l}_{21}; \quad n^l_{22} = n^{0l}_{22} + n^{1l}_{22}; \quad n^l_{23} = r_{001} - n^{0l}_{13} + n^{1l}_{23};$
9. $n^l_{31} = n^{0l}_{31} + r_{110} - n^{1l}_{21}; \quad n^l_{32} = r_{010} - n^{0l}_{31} + n^{1l}_{32}; \quad n^l_{33} = r_{011} + n^{1l}_{33}$

P-step: generate θ^l from the posterior distribution derived based on complete data, $p(\theta^l | n^l_{com})$, where n^l_{com} include the counts of the complete data obtained in the

I-step from the lth iteration.

$$p(\exp(\lambda^l)|\cdot) \sim \text{gamma}\left(n^l_{1+} + n^l_{2+} + n^l_{3+} + a, \frac{1}{\sum_{j=1}^{3}\sum_{k=1}^{3}(2 \times V + \frac{1}{b}}\right),$$

$$p(\exp(\lambda^l_{1S})|\cdot) \sim \text{gamma}\left(n^l_{1+} + a, \frac{1}{2 \times V1S + \frac{1}{b}}\right),$$

$$p(\exp(\lambda^l_{1T})|\cdot) \sim \text{gamma}\left(n^l_{+1} + a, \frac{1}{2 \times V1T + \frac{1}{b}}\right),$$

$$p(\exp(\lambda^l_{3S})|\cdot) \sim \text{gamma}\left(n^l_{3+} + a, \frac{1}{2 \times V3S + \frac{1}{b}}\right),$$

$$p(\exp(\lambda^l_{3T})|\cdot) \sim \text{gamma}\left(n^l_{+3} + a, \frac{1}{2 \times V3T + \frac{1}{b}}\right),$$

$$p(\lambda^l_{11}|\cdot) \propto 2\exp(-\exp(\lambda^l + \lambda^l_{1S} + \lambda^l_{1T} + \lambda^l_{11}))\exp(\lambda^l_{11})^{n_{11}}\exp(-(\lambda^l_{11} - u)^2/(2v^2)),$$

$$p(\lambda^l_{13}|\cdot) \propto 2\exp(-\exp(\lambda^l + \lambda^l_{1S} + \lambda^l_{3T} + \lambda^l_{13}))\exp(\lambda^l_{13})^{n_{13}}\exp(-(\lambda^l_{13} + u)^2/(2v^2)),$$

$$p(\lambda^l_{31}|\cdot) \propto 2\exp(-\exp(\lambda^l + \lambda^l_{3S} + \lambda^l_{1T} + \lambda^l_{31}))\exp(\lambda^l_{31})^{n_{31}}\exp(-(\lambda^l_{31} + u)^2/(2v^2)),$$

$$p(\lambda^l_{33}|\cdot) \propto 2\exp(-\exp(\lambda^l + \lambda^l_{3S} + \lambda^l_{3T} + \lambda^l_{33}))\exp(\lambda^l_{33})^{n_{33}}\exp(-(\lambda^l_{33} - u)^2/(2v^2)),$$

$$p^l_{jk} = \frac{\exp(\lambda^l_{jS} + \lambda^l_{kT} + \lambda^l_{jk})}{\sum_j\sum_k \exp(\lambda^l_{jS} + \lambda^l_{kT} + \lambda^l_{jk})}, j, k = 1, 2, 3,$$

where,

$$V = \exp(\lambda^l_{jS} + \lambda^l_{kT} + \lambda^l_{jk})),$$

$$V1S = \exp(\lambda^l + \lambda^l_{1T} + \lambda^l_{11}) + 2\exp(\lambda^l) + 2\exp(\lambda^l + \lambda^l_{3T} + \lambda^l_{13}),$$

$$V1T = \exp(\lambda^l + \lambda^l_{1S} + \lambda^l_{11}) + 2\exp(\lambda^l) + 2\exp(\lambda^l + \lambda^l_{3S} + \lambda^l_{31}),$$

$$V3S = \exp(\lambda^l + \lambda^l_{1T} + \lambda^l_{31}) + 2\exp(\lambda^l) + 2\exp(\lambda^l + \lambda^l_{3T} + \lambda^l_{33}),$$

$$V3T = \exp(\lambda^l + \lambda^l_{1S} + \lambda^l_{13}) + 2\exp(\lambda^l) + 2\exp(\lambda^l + \lambda^l_{3S} + \lambda^l_{33}),$$

$\cdot$ represents all the rest of the parameters, n^l_{j+} denotes $\sum_{k=1}^{3} n^l_{jk}$, n^l_{+k} denotes $\sum_{j=1}^{3} n^l_{jk}$ and so on. For $\exp(\lambda)$, $\exp(\lambda_{1S})$, $\exp(\lambda_{3S})$, $\exp(\lambda_{1T})$ and $\exp(\lambda_{3T})$, the conditional draws can be made directly from gamma distributions using the Gibbs sampler. For λ_{11}, λ_{13}, λ_{31} and λ_{33}, we could not draw directly from appropriate conditional distributions, instead we use the Metropolis-Hastings algorithm (Gelman et

al, 2004). The proposal distribution in the Metropolis-Hastings is normal, with variance adjusted to give an acceptance rate of approximate 25%. We ran Markov Chain Monte Carlo (MCMC) for $400,000$ iterations and discarded the first $200,000$ iterations for burn-in. We obtained 2000 draws by saving every $100th$ sample after the burn-in period. Convergence was assessed graphically. The sensitivity towards the initial values was evaluated by comparing parameter estimates from five chains. For the quantities of interest, we obtained the Gelman-Rubin Statistic, ($\hat{R}$), which is the squared root of the ratio of the adjusted sum of the between- and within-chain variances over the within-chain variances (Gelman *et al.*, 2004). At convergence, $\hat{R} = 1$. Generally, $\hat{R} = 1.2$ is considered sufficient for convergence. In the application to the CIGTS data, for all the counterfactual probabilities and ORs, min $\hat{R} = 0.99985$ and max $\hat{R} = 1.00067$. The algorithm for estimating the parameters is also validated on simulated data.

4.4 Application to Glaucoma Data

4.4.1 The Results

We apply the method to estimate the counterfactual probabilities using the data from 228 patients in the CIGTS with whom S, T and Z are completely observed. In Table 4.5, we report the medians and their 95% credible intervals (CI) from the posterior distributions of the counterfactual probabilities. We choose $a = 0.01$, $b = 100$, $u = 0.7$, $v^2 = 1.4^2$ for the prior distributions to induce weak ordering. The estimated causal treatment effect for $p_{12} + p_{22} + p_{32}$ is 0.11 with its 95% CI $(0.013, 0.23)$. The observed treatment effect is 0.13 with its 95% confidence interval being $(0.014, 0.25)$, which is slightly bigger than that estimated using the counterfactual model. The difference is resulted from the assumptions we made in the counterfactual model. The associative effect p_{22} is estimated as 0.027 with its 95%CI $(0.0028, 0.092)$ and

the dissociative effect $p_{12} + p_{32}$ as $0.079(0.0093, 0.16)$. The associative proportion $p_{22}/(p_{12} + p_{22} + p_{32})$ is estimated as $0.27(0.078, 0.54)$ which indicates that the causal effect on the true endpoint is only partially reflected on the causal effect on the surrogate marker. Its credible interval is wide implying the associative proportion estimate is quite variable. The ratio $p_{22}/(p_{12} + p_{22} + p_{32} + p_{21} + p_{23})$ is estimated as $0.08(0.012, 0.22)$.

4.4.2 Sensitivity of Priors

We consider the impact of the priors on the posterior distributions. If there is no data available to provide information on the parameters of interest, the posterior density will exactly match the prior density. If there is a lot of information provided by the data, the posterior may differ from the prior. We evaluate identifiability by plotting the prior and posterior distributions against each other (Garrent and Zeger, 2000). Figure 4.1 shows the prior and posterior distributions for selected quantities of interest when $u = 0.7$ and $v = 1.4$. The average overlaps between the prior and posterior distributions for p_{11}, p_{33} and the causal effect are small, indicating that these parameters are likely well identified. On the other hand, we find more overlap between the prior and posteriors for p_{12}, p_{21}, and p_{32} which indicate these parameters are less identifiable. Figure 4.2 shows the prior and posterior distributions of four odds ratios. There is a substantial overlap between the prior and posterior distribution for OR_1, OR_2 and OR_3, indicating that much of the information for these ORs is provided from the prior assumptions. There is less overlap between the prior and posterior for OR_4, indicating that some information is available from the data.

To further assess the extent of the impact of the priors on the posterior distributions, we vary the variances of the prior distributions for λ_{11}, λ_{13}, λ_{31} and λ_{33}

and fix the means. Then, we vary the means of these prior distributions but fix the variances. The results are listed in Table 4.6. The second column lists the posterior medians and standard deviations when $u = 0.7$ and $v^2 = 1.96$. Columns 3, 4, 5, and 6 list the percentage of the changes in the posterior medians and standard deviations relative to the second column. When we change the prior mean, we observe bigger changes in the posterior medians than the posterior variances of the counterfactual probabilities. Relative to the medians when $u = 0.7$, with $u = 0$ or $u = 1.4$, the extent of the changes in the medians is less than 10%. When we change the prior variance, we observe more changes in the posterior variances than the posterior means for the counterfactual probabilities. Compared with the variances when $v^2 = 1.96$, with $v^2 = 1$ or $v^2 = 4$, the changes in the posterior standard deviations are generally less than 10% with only one exception. However, the odds ratios are much more sensitive to the prior assumptions, especially for OR_1, OR_2 and OR_3 (not listed).

4.5 Simulation Study

We conduct a small simulation study to examine the properties of the estimation method. We simulate 200 data sets under the parameter specification: $\lambda_{1S} = 0.15$, $\lambda_{1T} = -0.3$, $\lambda_{3S} = 0.3$, $\lambda_{3T} = -0.7$, $\lambda_{11} = 0.5$, $\lambda_{13} = -0.8$, $\lambda_{31} = -0.5$, $\lambda_{33} = 0.8$ and $\lambda = 3.5$. We analyze the simulated data assuming the correct model structure with the prior distributions for λ_{11} and λ_{33} being normal$(0.7, 1.4^2)$ and for λ_{13} and λ_{31} being normal$(-0.7, 1.4^2)$. We use a nearly noninformative prior gamma$(0.01, 100)$ for $\exp(\lambda_{1S})$, $\exp(\lambda_{1T})$, $\exp(\lambda_{3S})$, $\exp(\lambda_{3T})$ and $\exp(\lambda)$. The simulation results are listed in Table 4.7. $\text{SD}(\overline{Est})$ is the standard deviation of the posterior means from 200 data sets. $\overline{SD}$ is the mean of the posterior standard deviations from 200 data sets. The estimated parameter values are very close to the true values. Since we used

informative priors for λ_{11}, λ_{13}, λ_{31} and λ_{33}, in which the true values are included in the 95% credible intervals of the priors, we observe over-coverage and SD($\overline{Est}$) less than $\overline{SD}$ for the less identifiable quantities.

4.6 Relationship between the Counterfactual Model and Conventional Models

In this section, we examine the surrogacy measures in both the counterfactual model setup and the conventional model setup. In a more conventional model setup, we use logistic regression to model the joint distribution of S, T, and Z:

$$\begin{aligned} \text{logit}[P(T_i = 1|S_i, Z_i)] &= \beta_0 + \beta_1 S_i + \beta_2 Z_i + \beta_3 Z_i S_i, \\ \text{logit}[P(S_i = 1|Z_i)] &= \alpha_0 + \alpha_1 Z_i, \\ P(Z_i = 1) &= 0.5. \end{aligned} \tag{4.2}$$

The parameters in these models can be expressed as the functions of the counterfactual probabilities. For example, some of the parameters in model (4.2) are given by:

$$\begin{aligned} \exp(\beta_1) &= \frac{(p_{11}+p_{12}+p_{21}+p_{22})p_{33}}{(p_{13}+p_{23})(p_{31}+p_{32})}, \\ \exp(\beta_2) &= \frac{(p_{11}+p_{12}+p_{21}+p_{22})(p_{12}+p_{13})}{(p_{13}+p_{23})p_{11}}, \\ \exp(\beta_3) &= \frac{(p_{13}+p_{23})(p_{31}+p_{32})p_{11}(p_{22}+p_{23}+p_{32}+p_{33})}{(p_{11}+p_{12}+p_{21}+p_{22})p_{33}(p_{12}+p_{13})(p_{21}+p_{31})}. \end{aligned}$$

4.6.1 Perfect Surrogacy and Principal Surrogacy

Prentice (1989) proposed a formal definition for perfect surrogacy and provided validation criteria. The most essential criterion requires that changes in S fully capture the effect of Z on T, i.e., $\beta_1 \neq 0$, $\beta_2 = 0$ and $\beta_3 = 0$. Perfect surrogacy is also called statistical surrogacy. Frangakis and Rubin (2002) suggested a definition

for principal surrogacy which requires that causal treatment effect on T may only exist when the causal treatment effect on S exist; i.e., $p_{12} = p_{32} = 0$. When S and T are binary, we argue that two more restrictions, $p_{21} = p_{23} = 0$, ensure that the causal treatment effect on T is completely captured by the causal treatment effect on S, and that the causal effect on S perfectly reflects the causal effect on T. In other words, the causal effect on T is equal to that on S. Under this condition, β_2 and β_3 can be simplified to:

$$\begin{aligned} \exp(\beta_2) &= \frac{p_{11} + p_{22}}{p_{11}}, \\ \exp(\beta_3) &= \frac{p_{11}p_{33} + p_{22}p_{11}}{p_{11}p_{33} + p_{22}p_{33}}. \end{aligned}$$

For S to be meaningful as a surrogate in the counterfactual framework, we require $p_{22} > 0$, which leads to $\beta_2 > 0$. Therefore, when the causal effect on T equals the causal effect on S, S does not satisfy the criteria for perfect surrogacy.

4.6.2 Surrogacy Measures

In this section, we explore the connections among a few commonly used surrogacy measures in the counterfactual framework. When a surrogate marker does not satisfy perfect surrogacy, Freedman *et al.* (1992) proposed a measure based on the proportion of the treatment effect explained. One of the drawbacks of this measure is that it assumes there is no interaction between S and T; i.e., $\beta_3 = 0$. A measure which is free of this assumption was proposed by Wang and Taylor (2002) as $F_{WT} = \delta\gamma_a/\tau$

where,

$$\begin{aligned}
\delta &= P(S=1|Z=0) - P(S=1|Z=1) = p_{21} + p_{22} + p_{23}, \\
\tau &= P(T=1|Z=0) - P(T=1|Z=1) = p_{12} + p_{22} + p_{32}, \\
\gamma_a &= P(T=1|Z=0, S=1) - P(T=1|Z=0, S=0) \\
&= \frac{p_{33}}{p_{31} + p_{32} + p_{33}} - \frac{p_{13} + p_{23}}{p_{11} + p_{12} + p_{21} + p_{22} + p_{13} + p_{23}}.
\end{aligned}$$

Odds ratios (OR) are also often used to describe the association between S and T. The odds ratio in the $Z = 0$ group is denoted as OR_{g0} and that in the $Z = 1$ group denoted as OR_{g1}. $OR_{g0} = ((p_{11} + p_{12} + p_{21} + p_{22}) \times p_{33})/((p_{13} + p_{23}) \times (p_{31} + p_{32}))$ and $OR_{g1} = (p_{11} \times (p_{22} + p_{23} + p_{32} + p_{33}))/((p_{12} + p_{13}) \times (p_{21} + p_{31}))$.

In the counterfactual framework, we propose a new measure, common associative proportion (CAP), based on the principal surrogacy concept

$$CAP = p_{22}/(p_{12} + p_{21} + p_{22} + p_{23} + p_{32}), \tag{4.3}$$

which quantifies the relationship between the causal effect on S and that on T. When $p_{12} = p_{21} = p_{23} = p_{32} = 0$, $F_{PS} = 1$; when $p_{22} = 0$ then $F_{PS} = 0$. This measure is usually smaller than the associative proportion (AP). One of the good properties of CAP and AP is that they always fall in the range $[0, 1]$.

To better understand these surrogacy measures and the underlying assumptions, we calculate the measures in several hypothetical scenarios and two examples are given below. In Example 1 in Table 4.8, when the causal treatment effect on T is the same across three principal strata, CAP and AP are relatively small indicating a small causal association between S and T; however, the large values in F_{WT}, OR_{g0} and OR_{g1} show that S is closely related to T in a conventional model setup.

In Example 2 in Table 4.9, all surrogacy measures indicate a close relationship between S and T in both the traditional model and counterfactual model framework.

By examining the formulas for these surrogacy measures, we find the connections among CAP, AP, OR_{g0}, OR_{g1} and F_{WT} are complex. In general, when p_{11} and p_{33} are relatively large compared with the off-diagonal probabilities in the same rows and columns, S is highly associated with T with respect to Z in a traditional model setup. When p_{22} is relative large compared with the off-diagonal probabilities in the same row and column, S is closely associated with T in a counterfactual model framework.

4.7 Missing True Endpoints

Previously, we focused on the situation when S, T and Z are completely observed. In this Section, we extend the method to accommodate partially missing T. A surrogate marker can serve as an auxiliary variable and could increase the efficiency of the estimate of the treatment effect on T when S is observed on more subjects than T. We assume T missing completely at random. In a counterfactual framework, we can easily incorporate S from the subjects whose Ts are not observed by modifying the I-steps in the data augmentation procedure. Let m_{zs} denote the number of subjects with $Z = z$, $S = s$ and T unobserved. Let $\omega_3 = p_{11} + p_{12} + p_{21} + p_{22} + p_{13} + p_{23}$, $\omega_4 = p_{31} + p_{32} + p_{33}$, $\omega_5 = p_{11} + p_{12} + p_{13}$ and $\omega_6 = p_{21} + p_{31} + p_{22} + p_{23} + p_{32} + p_{33}$. The P-step stays the same as before, and in the I-step, we add the contributions from m_{zs} to n_{11}, n_{12}, ..., and n_{33} in Steps 7, 8, and 9. Denote n_{11}^{m0l} as the draw of the count that contributes to n_{11} from m_{00} from the lth iteration, n_{21}^{m1l} as the draw of the count that contributes to n_{21} from m_{11} from the lth iteration and so on. In

the following, we calculate the contributions in the lth iteration:

$$(n_{11}^{m0l}, n_{12}^{m0l}, n_{21}^{m0l}, n_{22}^{m0l}, n_{13}^{m0l}, n_{23}^{m0l}) \sim \text{Multi}(m_{00}, \frac{p_{11}^l}{\omega_{3l}}, \frac{p_{12}^l}{\omega_{3l}}, \frac{p_{21}^l}{\omega_{3l}}, \frac{p_{22}^l}{\omega_{3l}}, \frac{p_{13}^l}{\omega_{3l}}, \frac{p_{23}^l}{\omega_{3l}})$$

$$(n_{31}^{m0l}, n_{32}^{m0l}, n_{33}^{m0l}) \sim \text{Multi}(m_{01}, \frac{p_{31}^l}{\omega_{4l}}, \frac{p_{32}^l}{\omega_{4l}}, \frac{p_{33}^l}{\omega_{4l}})$$

$$(n_{11}^{m1l}, n_{12}^{m1l}, n_{13}^{m1l}) \sim \text{Multi}(m_{10}, \frac{p_{11}^l}{\omega_{5l}}, \frac{p_{12}^l}{\omega_{5l}}, \frac{p_{13}^l}{\omega_{5l}})$$

$$(n_{21}^{m1l}, n_{31}^{m1l}, n_{22}^{m1l}, n_{23}^{m1l}, n_{32}^{m1l}, n_{33}^{m1l}) \sim \text{Multi}(m_{11}, \frac{p_{21}^l}{\omega_{6l}}, \frac{p_{31}^l}{\omega_{6l}}, \frac{p_{22}^l}{\omega_{6l}}, \frac{p_{23}^l}{\omega_{6l}}, \frac{p_{32}^l}{\omega_{6l}}, \frac{p_{33}^l}{\omega_{6l}})$$

Steps 7, 8, and 9 in the I-Step in Section 3.6 are modified as:

7. $n_{11}^l = n_{11}^{0l} + r_{100} + n_{11}^{m0l} + n_{11}^{m1l}$; $\quad n_{12}^l = n_{12}^{0l} + n_{12}^{1l} + n_{12}^{m0l} + n_{12}^{m1l}$;

   $n_{13}^l = n_{13}^{0l} + r_{101} - n_{12}^{1l} + n_{13}^{m0l} + n_{13}^{m1l}$;

8. $n_{21}^l = n_{21}^{0l} + n_{21}^{1l} + n_{21}^{m0l} + n_{21}^{m1l}$; $\quad n_{22}^l = n_{22}^{0l} + n_{22}^{1l} + n_{22}^{m0l} + n_{22}^{m1l}$;

   $n_{23}^l = r_{001} - n_{13}^{0l} + n_{23}^{1l} + n_{23}^{m0l} + n_{23}^{m1l}$;

9. $n_{31}^l = n_{31}^{0l} + r_{110} - n_{21}^{1l} + n_{31}^{m0l} + n_{31}^{m1l}$; $\quad n_{32}^l = r_{010} - n_{31}^{0l} + n_{32}^{1l} + n_{32}^{m0l} + n_{32}^{m1l}$;

   $n_{33}^l = r_{011} + n_{33}^{1l} + n_{33}^{m0l} + n_{33}^{m1l}$

We apply the method to the CIGTS data based on a total of 573 patients of whom we completely observe S but do not observe T among 345 patients. We use the same prior specifications as before. The summary statistics for the counterfactual probabilities from the posterior distributions based on data from the patients whose S and T are observed (complete cases) and the summary statistics based on data from all 573 patients (all cases) are both listed in Table 4.10. Only modest efficiency gain is obtained with the use of the surrogate marker in most of the parameters. Unless there is much higher association, it is likely that the potential gain from S is limited when S and T are both binary.

4.8 Extension to Multiple Trials

In this Section, we extend our method to a multiple-trial setting. When multiple trials are available and when it is valid to consider the trials exchangeable, we may take advantage of such exchangeability to estimate the degree of association between $(S(0), S(1))$ and $(T(0), T(1))$ by making distributional assumptions to warrant the sharing of information on the properties of the surrogate across trials. For trial $h = 1, ..., H$, the number of patients cross-classified by $(S(0), S(1))$ and $(T(0), T(1))$ is presented in Table 4.11. The complete-data likelihood is given by:

$$L_{com} = \prod_z \prod_h \prod_j \prod_k \frac{\exp(-\exp(\lambda_h + \lambda_{hjS} + \lambda_{hkT} + \lambda_{hjk}))(\exp(\lambda_h + \lambda_{hjS} + \lambda_{hkT} + \lambda_{hjk}))^{n^z_{hjk}}}{n^z_{hjk}!}$$

where, $(\lambda_{h2S} = \lambda_{h2T} = \lambda_{hj2} = \lambda_{h2k} = 0)$.

We assume nearly noninformative priors, gamma(a, b), for λ_h, λ_{hjS} and λ_{hkT} in trial h where $a = 0.01$, $b = 100$. We treat λ_{hjk} as the trial-specific parameter that centers around a population parameter u. The relationship between u and λ_{hjk} is specified using a hierarchical structure:

$$p(\lambda_{h11}) = \text{normal}(u, v^2), \;\; p(\lambda_{h13}) = \text{normal}(-u, v^2),$$
$$p(\lambda_{h31}) = \text{normal}(-u, v^2), \;\; p(\lambda_{h33}) = \text{normal}(u, v^2).$$

We assume independent priors for u and v^2:

$$p(u) = \text{normal}(\delta, \sigma^2), \quad p(v^{-2}) = \text{gamma}(\tau_a, \tau_b),$$

where δ, σ^2, τ_a, and τ_b are pre-specified hyper-parameters which we can use to reflect our belief of the ordering restriction. An example of the hyper-parameters can be: $\delta = 0.7$, $\sigma^2 = 10$, $\tau_a = 0.01$ and $\tau_b = 100$. We reflects our belief that the odds ratios are more likely to be greater than 1 than not through δ. Note that the information

that comes from the prior distributions is much weaker than what we have assumed in a single trial setting because we convey our belief through u and v^2 which have very "vague" hyperprior specifications. Here we assume normal priors with the same means and variances for λ_{h11}, $-\lambda_{h13}$, $-\lambda_{h31}$ and λ_{h33}. We have also tried normal priors with different means and variances; however, it appears that the data we apply this method to do not have enough information to estimate this more flexible model.

The data augmentation procedure can be readily extended to the multiple-trial setting where the I-step and the P-step are similar within each trial conditioning on u and v^2 to those in the single trial setting. We can obtain the posterior distributions of u and v^2 in the P-step as follows:

$$u \sim \text{normal}(\frac{B}{A}, \frac{1}{2A}),$$
$$v^{-2} \sim \text{gamma}(2H + \tau_a, \frac{1}{C}),$$

where, $A = \frac{2H}{v^2} + \frac{1}{2\sigma^2}$, $B = \frac{\sum_h \lambda_{h11} - \sum_h \lambda_{h13} - \sum_h \lambda_{h31} + \sum_h \lambda_{h33}}{2v^2} + \frac{\delta}{2\sigma^2}$ and $C = \frac{\sum_h (\lambda_{h11} - u)^2}{2} + \frac{\sum_h (\lambda_{h13} + u)^2}{2} + \frac{\sum_h (\lambda_{h31} + u)^2}{2} + \frac{\sum_h (\lambda_{h33} - u)^2}{2} + \frac{1}{\tau_b}$.

4.8.1 Data Analysis 1

The Data

To evaluate the causal surrogacy in a meta-analytic setting, we treat the centers in the CIGTS study as independent trials for illustration purposes. We use the IOP measure at month 12 as the surrogate marker for the IOP measure at month 96, which serves as the true endpoint. A preliminary analysis of these data shows that the estimate of the between-trial variances is non-positive definite. Borrowing the idea of Gail *et. al.* (2000), we rescale up the data size by simulating S_{ij} and T_{ij} from bivariate normal distributions for each center and treatment group with the

center-specific and treatment-specific means and variance-covariances from the real data. The CIGTS study includes 14 centers, from which we delete five centers (i.e., 5, 7, 12, 13, 14) either because they had too few observations or because of non-positive definite covariance matrices within center. We also deleted two outliers that are greater than 35 mmHg. For the centers included ($n = 9$), we increase the sample sizes to be 335, 176, 385, 264, 539, 368, 286, 528, and 319. The trial-specific and treatment-specific means and correlations for S and T are listed in Table 4.12. We define both S and T as 1 if IOP is less than 18mmHg and 0 if otherwise. Table 4.13 lists the number of patients in each combination of Z, S and T within each center.

Estimation and Results

We run MCMC simulations for $100,000$ iterations with a burn-in period of $50,000$ and save $2,000$ simulations for every $25th$ iteration to form the posterior distributions. The convergence is examined in the same way as that in a single trial setting and is deemed adequate. We notice that the convergence is generally much faster than that in a single trial setting. Table 4.14 lists the medians and their 95% credible intervals of the posterior predictive distributions of the probabilities when we choose $\delta = 0.7$, $\sigma^2 = 10$, $\tau_a = 0.01$ and $\tau_b = 100$. The summary statistics of each probability are based on the average of the trial-specific probabilities across the nine trials (i.e., $p_{11} = \sum_{h=1}^{H} p_{h11}/H$) at each iteration. We can treat the averages as the estimates of the population-level probabilities. We report the medians in Table 4.14 because some posterior distributions are skewed. In a single trial setting, because of the lack of information on the odds ratios from the data, it is not feasible to take a less informative Bayesian approach; hence, we adopt rather informative priors to achieve convergence. However, in a multiple trial setting, more information is provided from the data on these ORs than that in a single trial setting, since we assume the log

ORs from different trials are from the same underlying distributions. As such, the priors do not have as much impact on the posteriors and rather "vague" priors for the hyper-parameters are sufficient for achieving convergence. The associative proportion and the causal effect on T explained by causal effect on S are fairly small, indicating that the surrogate marker only captures a relatively small part of how Z affects T. For the four odds ratios, OR_4 has much less variability than the other odds ratios, showing that more information about OR_4 is provided from the data.

Figure 4.3 presents the histograms of the posterior distributions of u and v^2. The fact that u has the mean of 0.6 and 95% CI being $(0.16, 1.12)$ agrees with what we believe about the positive associations between $(S(0), S(1))$ and $(T(0), T(1))$, that is, the log odds ratios are more likely to be positive than not. The estimated v^2 (95%CI)is $1.75(0.79, 3.99)$ and the CI is fairly tight and informative about the variability of u.

Figure 4.4 plots the posterior medians and standard deviations of the center-specific treatment effects estimated using our method against the observed center-specific treatment effects and their standard errors. The observed treatment effect and its standard error from each center are based on the observable quantity $(p(T = 1|Z = 1) - p(T = 1|Z = 0))$. The treatment effect estimated using our method is based on the MCMC simulations of the posterior distribution of $p_{12} + p_{22} + p_{32}$. We find that our MCMC estimates agree with the observable treatment effects; the posterior standard deviations are similar to the standard errors of the treatment effects except for center 1, 7 and 9 in which the posterior standard deviations are relatively smaller due to the shrinkage effect in the hierarchial model.

Figure 4.5 plots the densities of a few selected quantities of interest from a few centers and the density of the averages of the center-specific quantities. The model

allows the densities from one center to be quite different from that from another center. Since the averages are less variable, their density curves appear sharper than the center-specific densities.

We also examined the sensitivity to the hyperparameters. Assuming $\sigma^2 = 10$, $\tau_a = 0.01$ and $\tau_b = 100$, we vary δ from 0, to 0.7, then to 1.4. The posterior median and SD for u vary from 0.6(0.24), to 0.6(0.24), then to 0.6(0.23), those for v^2 vary from 1.69(0.76), to 1.75(0.81), then to 1.67(0.80); and those for p_{22} vary from 0.051(0.0069), to 0.0486(0.00653), then to 0.05(0.0068). With $\delta = 0.7$, $\sigma^2 = 100$, $\tau_a = 0.01$, and $\tau_b = 100$, the median and 95% CI for u is 0.61(0.24), those for v^2 is 1.72(0.76) and those for p_{22} is 0.0498(0.0639). With drastic smaller values for τ_a and τ_b, we generally have smaller posterior standard deviations for the quantities of interest. For example, with the same $\delta = 0.7$, $\sigma^2 = 100$, $\tau_a = 4$ and $\tau_b = 1/3$, we have 0.56(0.17) for u, 0.92(0.35) for v^2 and 0.0508(0.00567) for p_{22}. Posterior distributions of the counterfactual probabilities and those of u and v^2, nevertheless, are not overly sensitive to the values of the hyper-parameters.

4.8.2 Data Analysis 2

The Data

We apply our method to evaluate causal surrogacy in a meta-analytic study for advanced colorectal cancer trials. There are a total of 28 cancer clinical trials conducted between 1990 and 1996 by the Meta-Analysis Group in Cancer. The objective is to examine the effect improvement of several experimental treatments over a standard treatment. The standard treatment is fluoropyrimidines (5FU) given as a bolus intravenous injection. The experimental treatments are slight modifications of the standard treatment regimen: either 5FU with leucovorin, 5FU with methotrexate, 5FU given in continuous infusion, or hepatic arterial infusion of 5FU for patients with

metastasis confined to the liver. The data summaries based on 27 trials are presented in the book *The Evaluation of Surrogate Endpoints* (Molenberghs, Burzykowski and Buyse, 2004) which include a plot of Kaplan-Meier survival curves by treatment types (Figure 12.1) and two tables that contain the complete statistics on tumor responses, and the median survival and hazard ratios for each trial by treatment type (Tables 12.1 and 12.2). For these data the surrogate is the tumor response and we reconstructed the individual data on time to death from an accelerated failure time model with square-root-transformed survival time. We select final data sets from over 1000 simulations which match most closely with Figure 12.1 and Tables 12.1 and 12.2. Further details are given in the appendix. For our analysis, there is no censoring. The surrogate marker is the tumor response ($S = 1$ for complete or partial response and $S = 0$ for stable or progressive disease) and the true endpoint is the survival status at 1.75 years after the treatment (e.g., $T = 1$ for being alive and $T = 0$ for being dead). The data tables cross-classified by trial, treatment and survival status are given in Tables 4.21 and 4.22 in the Appendix section.

Estimation and Results

We run the MCMC simulations for $80,000$ iterations with a burn-in period of $40,000$ and save 2000 simulations for every $25th$ iteration to form the posterior distributions. The convergence is examined in the same way as that in a single trial setting and is deemed adequate. Table 4.15 lists the medians and their 95% credible intervals of the counterfactual probabilities from their posterior distributions when we choose $\delta = 0.7$, $\sigma^2 = 100$, $\tau_a = 0.01$ and $\tau_b = 100$. The summary statistics of each probability are based on the average of the trial-specific probabilities across all trials (i.e., $p_{11} = \sum_{h=1}^{H} p_{h11}/H$) at each iteration. We can treat the averages as the estimates of the population-level probabilities. We report the medians in Table 4.15

because some posterior distributions are skewed. In a single trial setting, because of the lack of information on the odds ratios from the data, it is not feasible to take a less informative Bayesian approach. Hence, we adopt rather informative priors to achieve convergence. However, in a multiple trial setting, more information is provided from the data on these ORs than that in a single trial setting since we assume the log ORs from different trials are from the same underlying distributions. As such, the priors do not have as much impact on the posteriors and rather "vague" priors for the hyper-parameters are sufficient for achieving convergence. The associative proportion and the causal effect on T explained by the causal effect on S are fairly small, indicating that the surrogate marker only captures a relatively small part of how Z affects T. Four odds ratios have similar distributions.

Figure 4.6 plots the posterior medians and standard deviations of the trial-specific treatment effect on T estimated using our method against the observed trial-specific treatment effects on T and their standard errors. The observed treatment effect and its standard error from each trial are based on the observable quantity $(p(T = 1|Z = 1) - p(T = 1|Z = 0))$. The treatment effect estimated using our method is based on the MCMC simulations of the posterior distribution of $p_{12} + p_{22} + p_{32}$. Our estimated treatment effects are all positive because our model assumes no negative individual-level effects and shrinks the negative effects towards 0. The posterior standard deviations from our hierarchical model are generally smaller than the observed standard errors of the treatment effects because of the shrinkage effects from the exchangeability assumption. Usually, the smaller the sample size the trial has, the more the shrinkage occurs; and the outliers on the plots are from trials with very few patients including the trial conducted in the City of Hope in 1996.

Figure 4.7 plots the densities of the quantities of interest from a few selected trials

and the averages of these quantities across all trials. The model allows the densities from one trial to be quite different from that from another trial. Since the averages are less variable, their density curves appear sharper than the trial-specific densities.

Figure 4.8 presents the histograms of the posterior distributions of u and v^2. The fact that u has the median of 0.588 and 95% CI being $(0.499, 0.691)$ agrees with what we believe about the positive associations between $(S(0), S(1))$ and $(T(0), T(1))$, that is, the log odds ratios are more likely to be positive than not. The median and 95% CI of v^2 is estimated as $0.052(0.0052, 0.259)$; the CI is tight and informative about the variability of u.

We also examined the sensitivity towards the hyperparameters. In Figures 4.8, 4.9 and 4.10, we present a few selected estimates of interest under "vague" priors (specifically, $\sigma^2 = 100$, $\tau_a = 0.01$ and $\tau_b = 100$) and informative priors (e.g., $\sigma^2 = 1.96$, $\tau_a = 2$ and $\tau_b = 1/7$). The prior mean of the log odds ratios u is not sensitive towards the change of the prior distributions, although v^2 is more sensitive. The identifiable quantities such as p_{11} and the causal effect are less variable than the other quantities. Overall, the posterior distributions of the counterfactual probabilities and those of u and v^2, nevertheless, are not overly sensitive to the hyper-parameters.

4.9 Discussion

This manuscript considers a potential outcomes approach to study the causal relationship among Z, S and T. It examines the association between the effect of Z on S and the effect of Z on T, respectively, as if we had observed both outcomes of S and T corresponding to two treatment options for every patient in the study. Previous surrogacy measures used to study the treatment-adjusted association between S and T often do not have causal interpretations because the model used $(T|S, Z)$ adjusts

for a post-treatment variable S which cuts off the effect of Z on T by conditioning on S. They quantify the association between $(S(0)|Z = 0)$ and $(T(0)|Z = 0)$ and that between $(S(1)|Z = 1)$ and $(T(1)|Z = 1)$. On the other hand, with the potential outcomes framework setup, we consider the association between $(S(0), S(1))$ and $(T(0), T(1))$ which is made on a common set of subjects and their values can not be changed by the treatment assignment. Hence, the association measures always have causal interpretations. The causal framework is similar in spirit to that used in the compliance literature (Holland, 1986; Imbens and Rubin, 1997) where the main interest is to estimate the causal effect of a treatment within the set of patients who would comply with their treatment assignment. In our setting, S and T are binary and we are interested in all the probabilities that completely describe the likelihood of each possible reaction in T to the two treatment options within different sets of patients stratified by the possible combinations of potential outcomes of S under two treatment options.

We compared the surrogacy measures using the conventional models with those using the counterfactual models. With a traditional model setup, very large surrogacy measures indicating the high correlation between S and T do not necessarily imply that the causal effect of Z on T mostly agrees with that of Z on S, vice versa. In an extreme situation, when the causal effect on S is equal to the causal effect on T, that is, S is a perfect surrogate in a causal sense, S is not a perfect surrogate in a conventional framework and does not satisfy the criteria for perfect surrogacy defined by Prentice's criteria. It illustrates the difficulty with drawing the conclusion on the causal relationship between S and T without the counterfactual framework. The traditional models ignore the fact that the effect of Z on T may occur to the patients who are inherently never-responsive or always-responsive in S regardless of

the treatment received, however, the counterfactual model teases out the effect of Z on T in each subgroup of subjects defined by their responsiveness to the treatment received. We find that when the diagonal probabilities, p_{11}, p_{22} and p_{33}, are all relatively large, S and T is closely associated in both the causal inference framework and the traditional model framework.

Frangakis and Rubin (2002) laid out a counterfactual framework to make it possible to discuss the assumptions under which the causal interpretation may or may not be plausible. We have extended this idea by proposing a Bayesian estimation method to make it possible to estimate the probabilities that measure the causal associations between S and T with respect to Z. We use the log-linear model to directly model the association between the potential outcomes of S and T through the odds ratios of $(S(0), S(1))$ and $(T(0), T(1))$. We believe that there is an ordering in the sequence of the potential outcomes of $(0, 0)$, $(0, 1)$ and $(1, 1)$. We also believe that $(S(0), S(1))$ and $(T(0), T(1))$ are closely associated. We incorporate these scientific assumptions conveniently through prior distributions for the odds ratios, for which there is little information from the observed data, and hence deal with the non-identifiable problems resulted from the over-parameterization using the counterfactual model. The proposed estimation method can be readily extended to the settings when T is partially missing or when there are multiple trials. Besides the log-linear model, we also fit a multinomial model with the Dirichlet prior distributions. Although it is easier computationally, the model is less flexible and the impact of the priors on the estimable quantities such as the treatment effect on T is much larger than the log-linear model. Although the Poisson log-linear model has one more parameter for the sample size than the multinomial model, there is a one-to-one relationship between the two model parameters conditioning on the sample size and independent prior

specification for the parameter denoting the sample size (Birch, 1963 and Forster, 1996).

It is important to realize that the probabilities in the counterfactual model are association measures instead of causation measures. If S is in the causal pathway between Z and T, p_{22} is very large. On the other hand, a very high p_{22} only shows that the causal effect of Z on S is highly associated with the causal effect of Z on T and it does not necessarily imply that Z affects T by affecting S. It is likely that there is an unmeasured variable denoted by U which can be a post-treatment confounder that can affect both S and T. Consequently, adjusting for S may induce the false association between Z and T. S is defined as "collider" in the economic literature. This problem can be directly addressed by another causal inference framework proposed by Greenland and Robins (1992) which allows one to manipulate S. The framework defines additional probabilities to describe the likelihood of how T changes by intervening S, which can measure the degree to which Z affect T through affecting S. This model has been used by Chen, Geng and Jia (2007) to study the surrogacy consistency.

One of the key assumptions in our method is the monotonicity assumption that requires that if a patient gets better if received $Z = 0$, she or he would not get worse if received $Z = 1$. It is essential to make this assumption to reduce the number of parameters to have a more identifiable counterfactual model. It is not usually contradicted by the data where on average patients do not become worse off when they received $Z = 1$ compared to those received $Z = 0$. If this assumption is correctly specified, we expect our estimates for the quantities of interest will be more efficient and less biased than the conventional model if the quantities are comparable. However, this assumption requires that every single patient would have done at least

as well as that when she or he receives $Z = 1$ relative to that when she or he receives $Z = 0$. It is perhaps true for most of the patients but not usually obviously satisfied for all patients, for example, in the CIGTS study where we compare the effect of medicine with that of surgery, it is conceivable that some patients may be better if they received medicine instead of surgery, even though the average effect of surgery is consistently better. On the other hand, for the colon cancer study, the assumption is more likely to hold for every individual because the experimental treatments are only slight modifications of the standard treatment and are intended to improve its effectiveness. Assessing the impact of the violations of the monotonicity assumption is an important extension in the current work.

Here, we assumed that missingness is ignorable, but it would be useful to investigate what assumption is necessary for the approach to be valid and conduct sensitivity analysis to examine the impact of different assumptions about missingness. In particular, when we have missingness in T or there is non-compliance in either a single trial setting or a multiple trial setting, it may be possible to add a third or fourth pair of potential outcomes to study alternative missingness assumptions or the impact of compliance status in the surrogate marker context. In this area, Frangakis and Rubin (1999) considers a weaker assumption of latent ignorability which allows missingness to depend on the principal strata. It will also be useful to calculate the non-parametric bounds to quantify the range of the counterfactual probabilities in our context (Balke and Pearl, 1997). Extensions to other data types are possible. For example, when both S and T are continuous, we can model the joint distribution of the potential outcomes of S and T using a parametric form. The association measures can be correlations instead of probability measures used in our manuscript.

S and T Observed			
		S	
	T	0	1
$Z=0$	0	28	14
	1	29	55
$Z=1$	0	11	10
	1	8	83
S Observed and T Missing			
$Z=0$		69	97
$Z=1$		35	144

Table 4.1: Data Summary from the Collaborative Initial Glaucoma Treatment Study

	$(T(0),T(1))$			
$(s(0),s(1))$	$(0,0)$	$(0,1)$	$(1,1)$	$(1,0)$
$(0,0)$	p_{11}	p_{12}	p_{13}	p_{14}
$(0,1)$	p_{21}	p_{22}	p_{23}	p_{24}
$(1,1)$	p_{31}	p_{32}	p_{33}	p_{34}
$(1,0)$	p_{41}	p_{42}	p_{43}	p_{44}

Table 4.2: Causal Probabilities from the Counterfactual Model

	$(T(0),T(1))$		
$(s(0),s(1))$	$(0,0)$	$(0,1)$	$(1,1)$
$(0,0)$	p_{11}	p_{12}	p_{13}
$(0,1)$	p_{21}	p_{22}	p_{23}
$(1,1)$	p_{31}	p_{32}	p_{33}

Table 4.3: Causal Probabilities from the Counterfactual Model with Monotonicity Assumption

		T	
	S	0	1
$Z=0$	0	$p_{11}+p_{12}+p_{21}+p_{22}$	$p_{13}+p_{23}$
	1	$p_{31}+p_{32}$	p_{33}
$Z=1$	0	p_{11}	$p_{12}+p_{13}$
	1	$p_{21}+p_{31}$	$p_{22}+p_{23}+p_{32}+p_{33}$

Table 4.4: Probabilities Associated with Observed Counts Using Counterfactual Parameters

	$(T(0), T(1))$		
$(s(0), s(1))$	$(0,0)$	$(0,1)$	$(1,1)$
	Median (95%CI)	Median (95%CI)	Median (95%CI)
$(0,0)$	$0.11(0.060, 0.17)$	$0.022(0.002, 0.072)$	$0.055(0.012, 0.12)$
$(0,1)$	$0.049(0.011, 0.108)$	$0.027(0.0028, 0.092)$	$0.17(0.080, 0.26)$
$(1,1)$	$0.051(0.0078, 0.11)$	$0.050(0.0036, 0.12)$	$0.45(0.36, 0.53)$

Table 4.5: Medians and 95% Credible Intervals of the Posterior Distributions for the Counterfactual Probabilities for CIGTS data

	Median(SD)	% Change in Median (% Change in Posterior SD)			
Quantities	$u = 0.7$	$u = 0$	$u = 1.4$	$u = 0.7$	$u = 0.7$
of Interest	$v^2 = 1.96$	$v^2 = 1.96$	$v^2 = 1.96$	$v^2 = 1$	$v^2 = 4$
p_{11}	0.11(0.028)	0%(1%)	−2%(2%)	−4%(−1%)	0%(3%)
p_{12}	0.022(0.019)	−5%(−4%)	10%(2%)	2%(−9%)	−2%(20%)
p_{22}	0.027(0.024)	−9%(−3%)	14%(6%)	8%(−3%)	−7%(6%)
p_{32}	0.050(0.032)	−3%(−4%)	10%(1%)	3%(−7%)	0.6%(7%)
Causal Effect	0.111(0.056)	−4%(−4%)	10%(3%)	1%(0.5%)	3%(−1%)
OR_1	2.99(8.04)	0.6%(−27%)	6%(−37%)	−10%(−60%)	12%(125%)
OR_2	2.45(10.09)	−5%(−6%)	435%(−96%)	−3%(−62%)	22%(2057%)
OR_4	1.56(1.87)	−4%(0%)	−3%(−9%)	1%(−41%)	−13%(73%)

Table 4.6: Prior Sensitivity on Posterior Distributions

Parameters	Truth	Bias	SD($\overline{Est}$)	$\overline{SD}$	Coverage
p_{11}	0.166	0.00383	0.0214	0.0223	95.5%
p_{12}	0.136	-0.00368	0.0210	0.0276	98.5%
p_{13}	0.0304	0.00453	0.00845	0.0205	100%
p_{21}	0.0869	0.00369	0.0148	0.0331	100%
p_{22}	0.117	-0.00975	0.0223	0.0353	100%
p_{23}	0.0582	-0.00429	0.0117	0.0218	98%
p_{31}	0.071	-0.00358	0.0144	0.0322	100%
p_{32}	0.158	0.00519	0.0221	0.0361	100%
p_{33}	0.175	0.00406	0.0227	0.0227	95%
OR_1	1.649	-0.00774	0.564	1.764	100%
OR_2	2.226	-0.283	0.555	13.086	100%
OR_3	1.649	0.448	0.774	13.554	100%
OR_4	2.226	0.0420	0.852	2.559	100%
Causal Effect	0.412	-0.0082	0.0394	0.0384	96%

Table 4.7: Bias, Standard Deviation of Posterior Means, Mean of Posterior Standard Deviations and Coverage Rates from 200 Simulations

Potential Outcomes	$(T(0), T(1))$			
$(S(0), S(1))$	$(0,0)$	$(0,1)$	$(1,1)$	Marginal
$(0,0)$	0.267	0.066	0.001	0.334
$(0,1)$	0.133	0.066	0.133	0.332
$(1,1)$	0.001	0.066	0.267	0.334
Marginal	0.401	0.198	0.401	1

Table 4.8: Example 1: $AP = 1/3$; $CAP = 0.142$; $F_{WT} = 1.003$; $OR_{g0} = 16$; $OR_{g1} = 16$

Potential Outcomes	$(T(0), T(1))$			
$(S(0), S(1))$	$(0,0)$	$(0,1)$	$(1,1)$	Marginal
$(0,0)$	0.31	0.03	0.005	0.345
$(0,1)$	0.03	0.24	0.04	0.31
$(1,1)$	0.005	0.04	0.30	0.345
Marginal	0.345	0.31	0.345	1

Table 4.9: Example 2: $AP = 0.77$; $CAP = 0.63$; $F_{WT} = 0.80$; $OR_{g0} = 90$; $OR_{g1} = 157$

Parameter	Complete Cases		All Cases	
	Median	95% CI	Median	95% CI
p_{11}	0.11	$(0.060, 0.17)$	0.11	$(0.067, 0.16)$
p_{12}	0.026	$(0.002, 0.072)$	0.019	$(0.0015, 0.072)$
p_{13}	0.057	$(0.012, 0.12)$	0.057	$(0.015, 0.11)$
p_{21}	0.051	$(0.011, 0.11)$	0.045	$(0.0092, 0.097)$
p_{22}	0.033	$(0.0028, 0.092)$	0.022	$(0.0018, 0.0698)$
p_{23}	0.17	$(0.080, 0.26)$	0.16	$(0.093, 0.23)$
p_{31}	0.053	$(0.0078, 0.11)$	0.056	$(0.011, 0.11)$
p_{32}	0.054	$(0.0036, 0.12)$	0.051	$(0.0046, 0.12)$
p_{33}	0.46	$(0.36, 0.53)$	0.46	$(0.39, 0.53)$
Causal Effect	0.11	$(0.013, 0.23)$	0.10	$(0.013, 0.21)$
Associative Proportion	0.27	$(0.078, 0.54)$	0.23	$(0.057, 0.49)$

Table 4.10: $a = 0.01$, $b = 100$, $u = 0.7$, and $v^2 = 1.4^2$

	$(T(0), T(1))$			
$(s(0), s(1))$	$(0,0)$	$(0,1)$	$(1,1)$	
$(0,0)$	n_{h11}	n_{h12}	n_{h13}	n_{h1+}
$(0,1)$	n_{h21}	n_{h22}	n_{h23}	n_{h2+}
$(1,1)$	n_{h31}	n_{h32}	n_{h33}	n_{h3+}

Table 4.11: Number of Patients by Potential Outcomes of S and T in trial h

Center	Sample Size	Medicine	Surgery	**Individual-level Correlation**	
		(Means of S, T)	(Means of S, T)	Medicine	Surgery
1	670	(17.63, 16.52)	(13.76, 14.59)	0.367	0.608
2	352	(17.22, 16.42)	(14.63, 12.98)	-0.455	0.467
3	770	(19.27, 17.58)	(15.81, 16.17)	0.589	0.548
4	528	(17.17, 15.51)	(10.93, 12.88)	0.176	0.540
5	1078	(18.52, 18.67)	(14.99, 15.32)	0.435	0.407
6	736	(18.62, 18.89)	(15.13, 17.11)	-0.16	-0.0056
7	572	(18.35, 15.34)	(14.59, 14.53)	0.177	0.396
8	1056	(18.59, 16.16)	(13.60, 13.72)	0.31	0.95
9	638	(17.56, 16.82)	(14.19, 14.61)	0.042	0.756

Table 4.12: Description of Pseudodata: Treatment-Specific Means and Individual-Level Correlations for Each Center.

		$Z=0$				$Z=1$			
		$S=0$		$S=1$		$S=0$		$S=1$	
Center	Sample Size	$T=0$	$T=1$	$T=0$	$T=1$	$T=0$	$T=1$	$T=0$	$T=1$
1	670	53	98	25	159	14	11	47	263
2	352	7	62	43	64	3	52	2	119
3	770	150	112	25	98	67	62	50	206
4	528	27	81	29	127	11	11	13	229
5	1078	231	88	97	123	64	100	74	301
6	736	116	90	105	57	33	61	94	180
7	572	43	105	32	106	28	45	37	176
8	1056	143	162	62	161	49	24	7	448
9	638	56	90	60	113	39	19	31	230

Table 4.13: Number of Subjects with Combinations of Z, S and T for Each Center.

	Median	95%CI		Median	95%CI		Median	95%CI
p_{11}	0.087	$(0.078, 0.096)$	p_{12}	0.038	$(0.024, 0.051)$	p_{13}	0.092	$(0.076, 0.11)$
p_{21}	0.052	$(0.037, 0.067)$	p_{22}	0.051	$(0.037, 0.065)$	p_{23}	0.19	$(0.17, 0.21)$
p_{31}	0.056	$(0.041, 0.070)$	p_{32}	0.095	$(0.078, 0.11)$	p_{33}	0.34	$(0.32, 0.35)$
CE^*	0.184	$(0.163, 0.204)$	AP^*	0.286	$(0.22, 0.35)$	CAP^*	0.11	$(0.076, 0.13)$
OR_1	7.56	$(3.92, 29.96)$	OR_2	3.64	$(1.43, 14.67)$	OR_3	3.39	$(1.44, 16.39)$
OR_4	1.34	$(0.78, 2.54)$	-	-	-	-	-	-

Table 4.14: Medians and 95% Credible Intervals for Counterfactual Probabilities. CE^*: Causal Treatment Effect on T; AP^*: Associative Proportion: $\frac{p_{22}}{p_{12}+p_{22}+p_{32}}$; CAP^*: Common Associative Proportion: $\frac{p_{22}}{p_{12}+p_{21}+p_{22}+p_{23}+p_{32}}$.

	Median	95%CI		Median	95%CI		Median	95%CI
p_{11}	0.623	$(0.601, 0.645)$	p_{12}	0.037	$(0.026, 0.050)$	p_{13}	0.076	$(0.066, 0.088)$
p_{21}	0.088	$(0.070, 0.107)$	p_{22}	0.011	$(0.0063, 0.018)$	p_{23}	0.0367	$(0.0264, 0.0499)$
p_{31}	0.048	$(0.039, 0.060)$	p_{32}	0.0096	$(0.0061, 0.016)$	p_{33}	0.069	$(0.056, 0.083)$
CE^*	0.058	$(0.041, 0.079)$	AP^*	0.172	$(0.136, 0.211)$	CAP^*	0.041	$(0.030, 0.056)$
OR_1	1.893	$(1.688, 2.350)$	OR_2	1.833	$(1.651, 2.184)$	OR_3	1.835	$(1.647, 2.178)$
OR_4	1.833	$(1.631, 2.195)$	-	-	-	-	-	-

Table 4.15: Medians and 95% Credible Intervals for Counterfactual Probabilities. CE^*: Causal Treatment Effect on T; AP^*: Associative Proportion: $\frac{p_{22}}{p_{12}+p_{22}+p_{32}}$ CAP^*: Common Associative Proportion: $\frac{p_{22}}{p_{12}+p_{21}+p_{22}+p_{23}+p_{32}}$

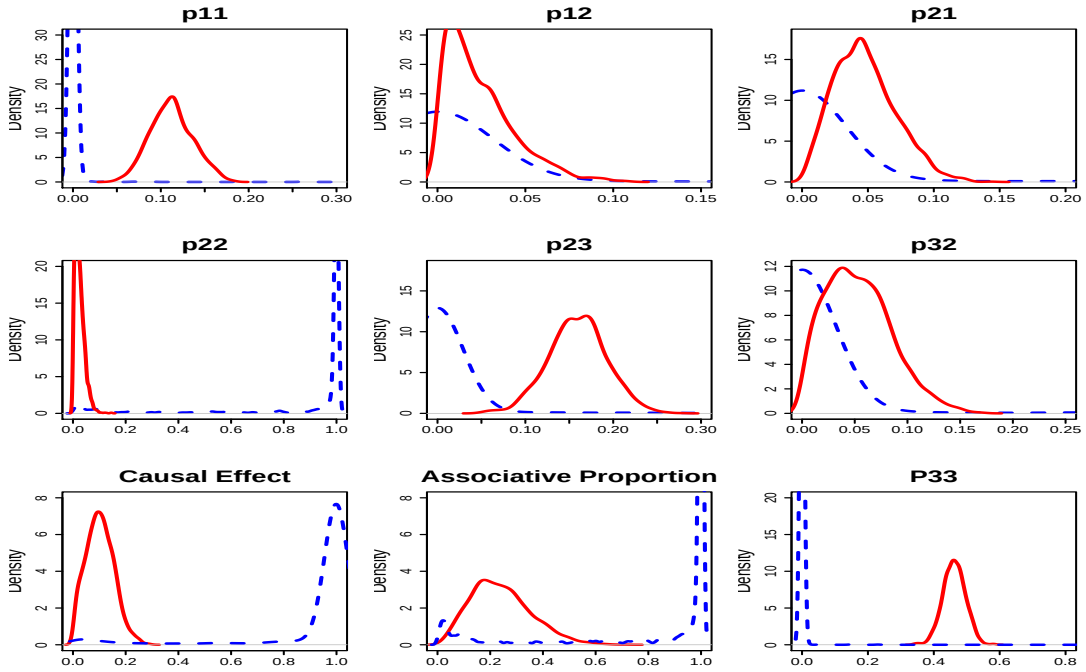

Figure 4.1: Prior and posterior distributions on selected quantities of interest. Dash lines for the prior distributions and solid lines for the posterior distributions.

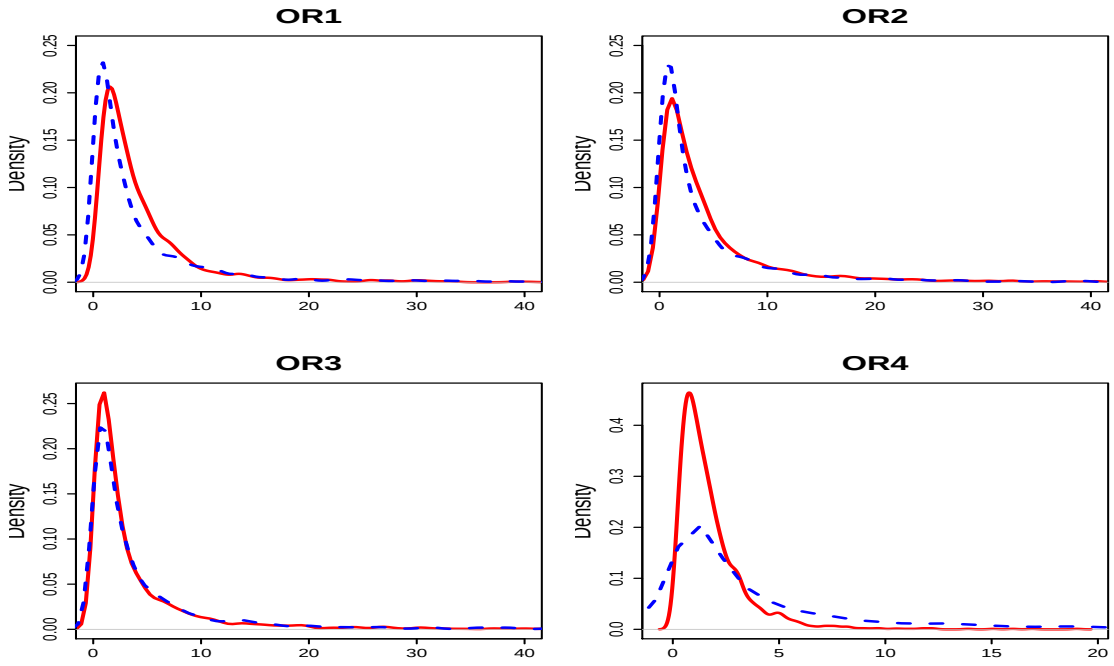

Figure 4.2: Prior and posterior distributions on four odds ratios. Dash lines for the prior distributions and solid lines for the posterior distributions.

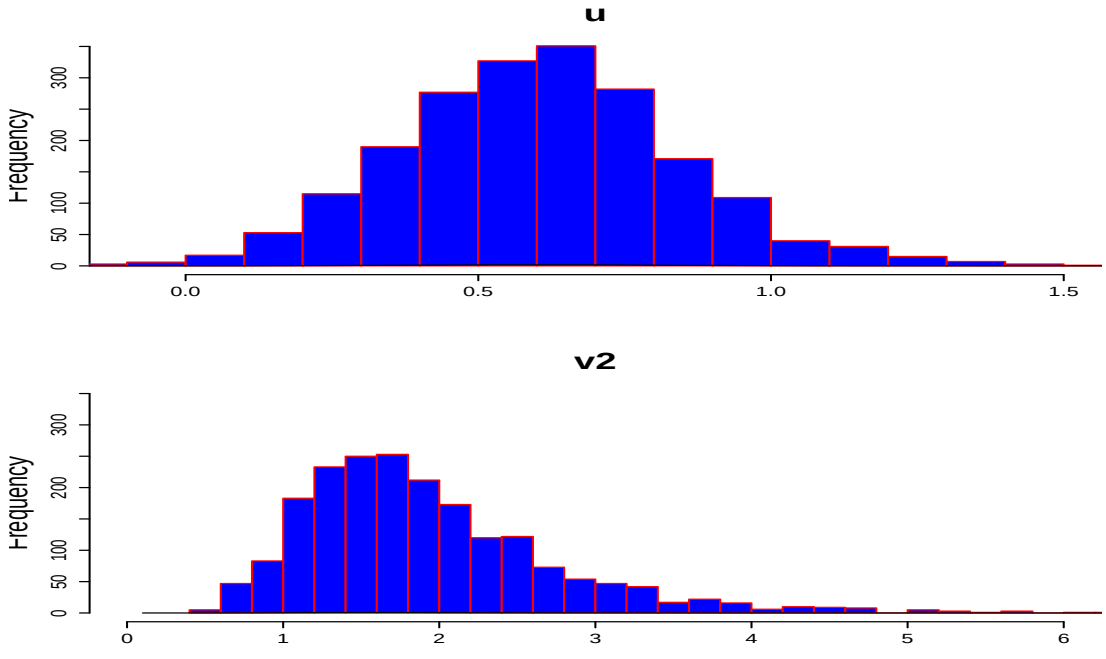

Figure 4.3: Histograms of 2000 MCMC Values from Posterior Distributions of u and v^2.

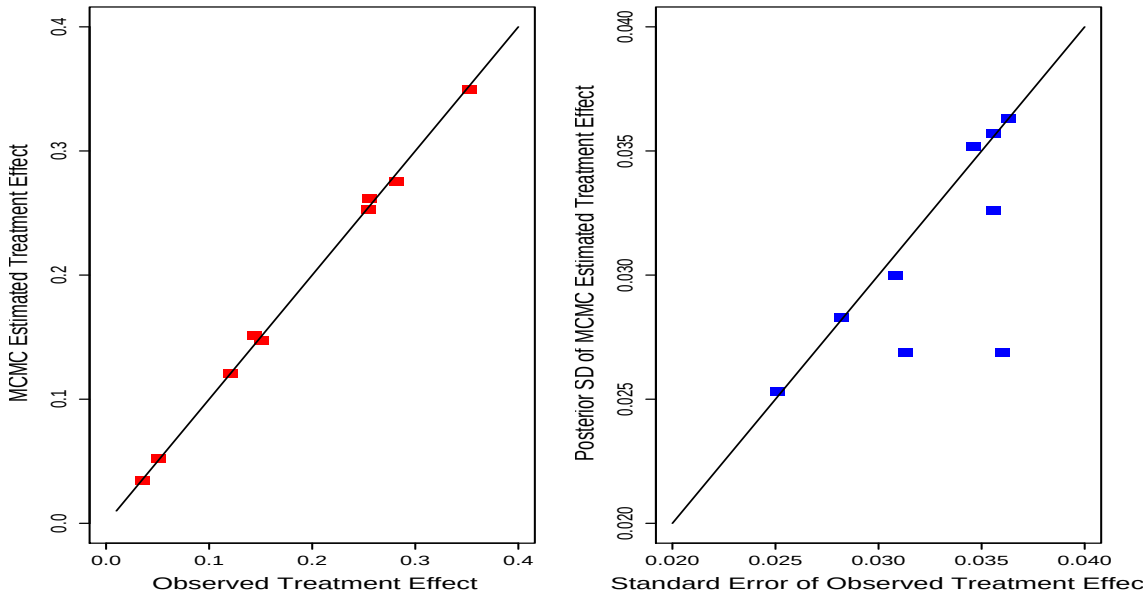

Figure 4.4: Observed Treatment Effect vs. MCMC Estimated Treatment Effect by Centers

4.10 Appendix

4.10.1 Data Reconstruction Description

The reconstructed data are based on an example of a meta-analysis study for advanced colorectal cancer discussed in *The Evaluation of Surrogate Endpoints* (Molenberghs *et al.*, 2004). The objective was to create data that matched a plot of Kaplan-Meier survival curves by treatment type ($Z = 1$ for experimental and $Z = 0$ for standard) and tumor response ($X = 1$ for complete response (CR), $X = 2$ for partial response (PR), $X = 3$ for stable disease (SD), and $X = 4$ for progressive disease (PD)) for four combined meta-analyses (Figure 4.5 from the book), provided with the median survival and HR estimated for each trial by treatment type (Tables 12.2 and 12.2 from the book). Several different survival models were considered, including proportional-hazards models using survival time simulated for a Weibull distribution

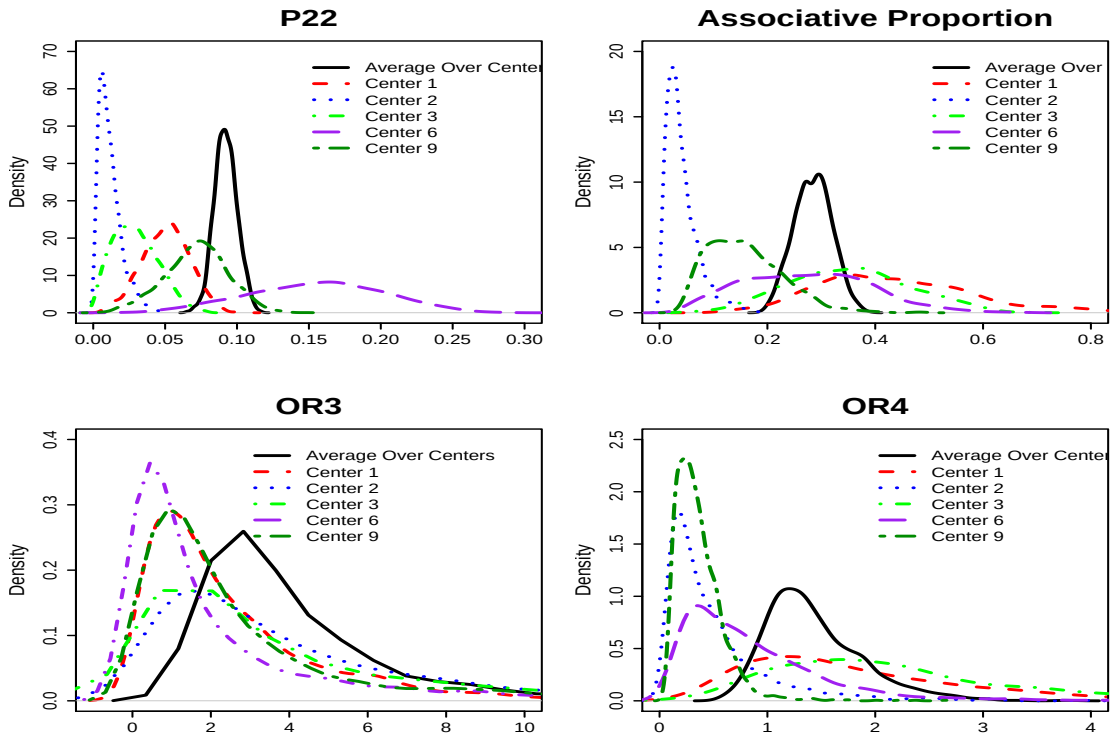

Figure 4.5: Posterior Distributions of Center-Specific Quantities and Their Averages by Centers

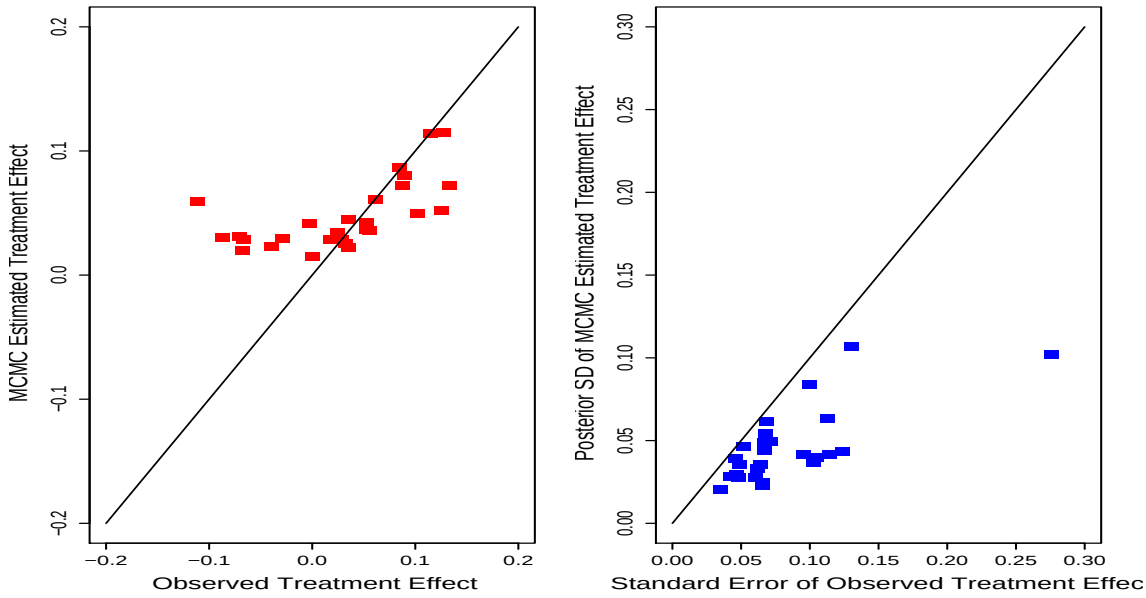

Figure 4.6: Observed Treatment Effect and its Standard Error vs. MCMC Estimated Treatment Effect and its Posterior Standard Deviation by Trial

and Accelerated Failure Time (AFT) models with log-transformed and square-root-transformed survival time. Ultimately, it appeared that simulated data from the AFT model with square-root-transformed survival time resulted in stratified survival curves that best resembled those provided in Figure 4.5 from the book. In particular, the final AFT model is given by:

$$(4.4)\sqrt{Y_{hzj}} = \eta_0 + \eta_{0hz} + \eta_1 I(X = 1) + \eta_2 I(X = 2) + \eta_3 I(X = 3) + \gamma_h Z + \varepsilon_j$$

for trial $h = 1, ..., 27$ and treatment $Z = 0, 1$ (experimental, standard), and subject $j = 1, ..., 4010$, where $\varepsilon_j \sim N(0, \sigma^2)$. The values of the parameters in this model were chosen to correspond to features of the data as summarized in Table 12.1, Table 12.2, and Figure 4.5 from the book. The initial values for each parameter in the AFT model are given in Table 4.16. In this table, $\widetilde{Y}_{hz}$ represents the median survival for trial h and treatment Z, and p_{xhz} is the proportion of subjects with tumor response level

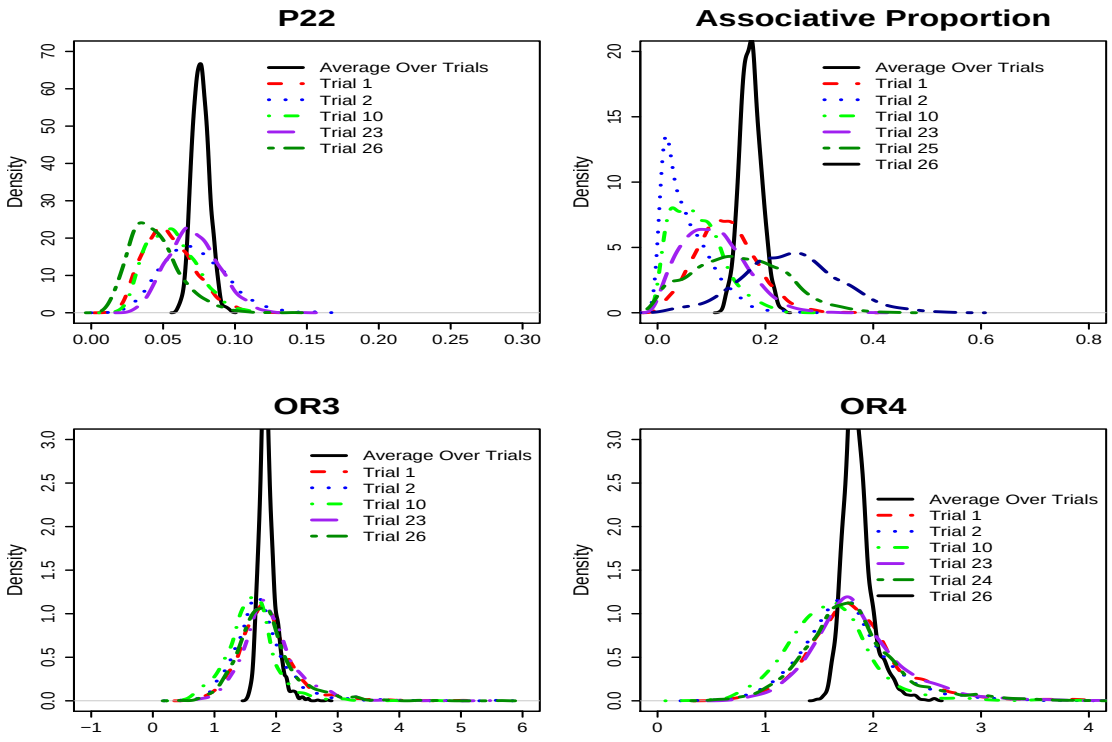

Figure 4.7: Posterior Distributions of Trial-Specific Quantities and Their Averages by Trials

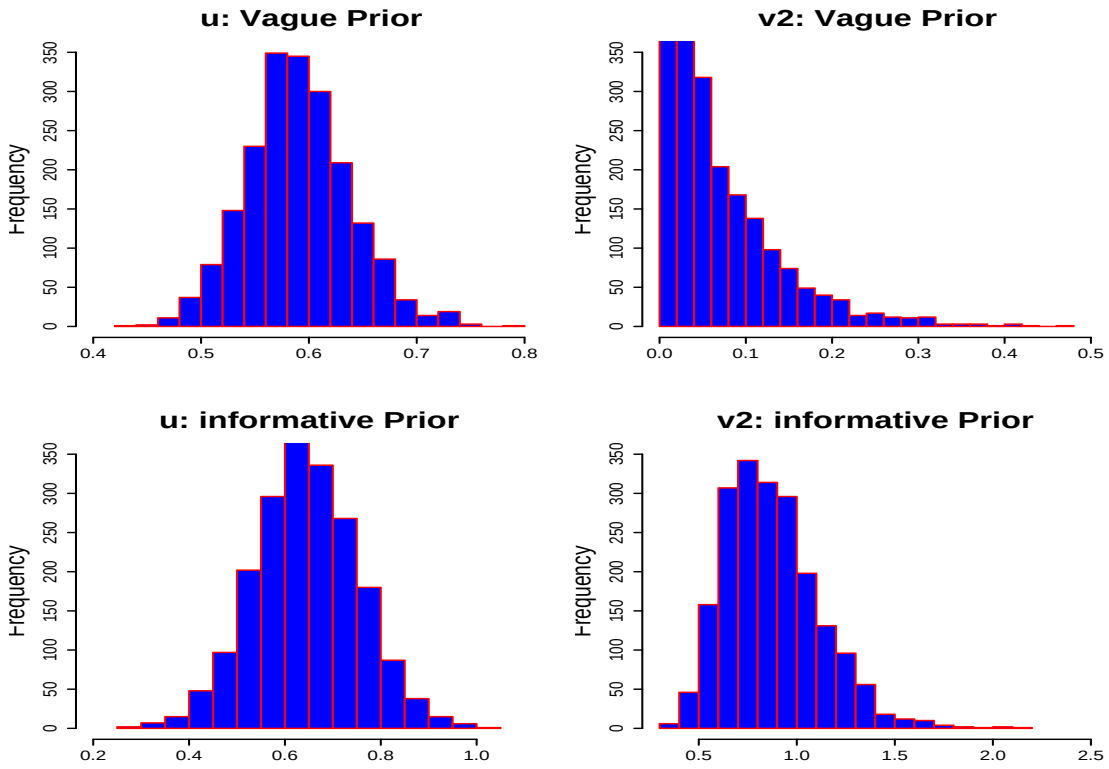

Figure 4.8: Histograms of 2000 MCMC Values from Posterior Distributions of u and v^2. Left Panel: Based on Vague Priors; Right Panel: Based on Informative Priors.

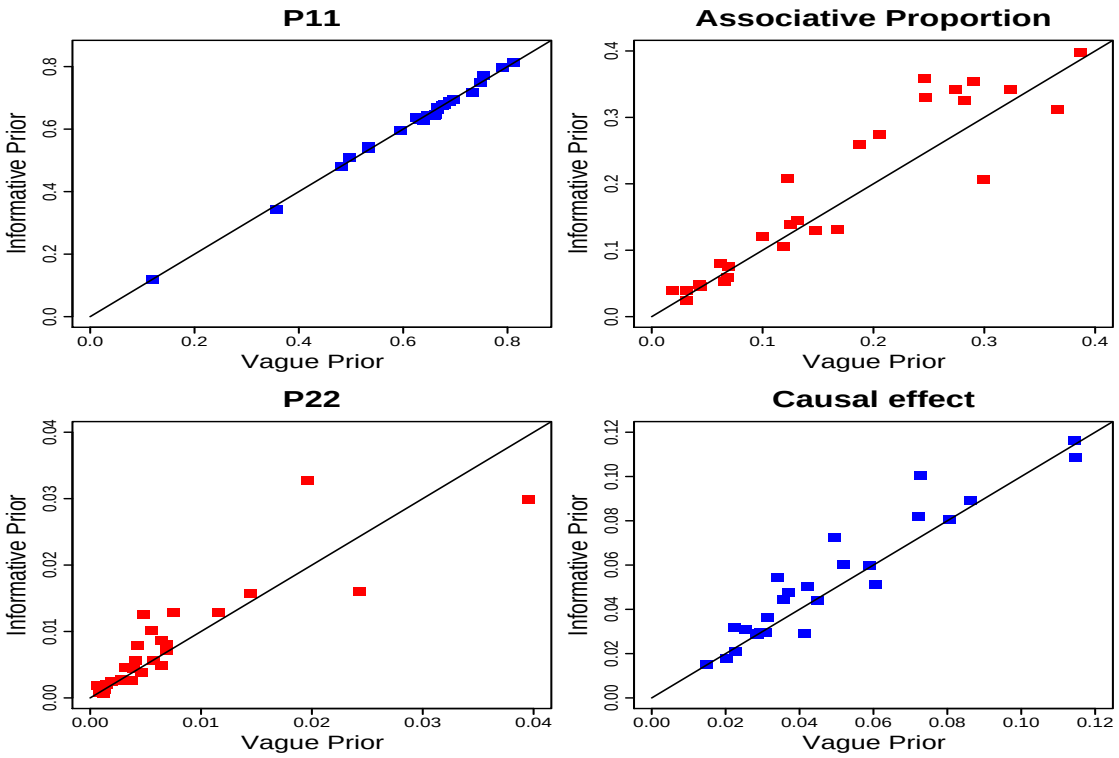

Figure 4.9: Posterior Medians of Trial-Specific p_{11}, p_{22}, Associative Proportions and Causal Treatment Effect Based on Informative Priors against Those Based on Vague Priors.

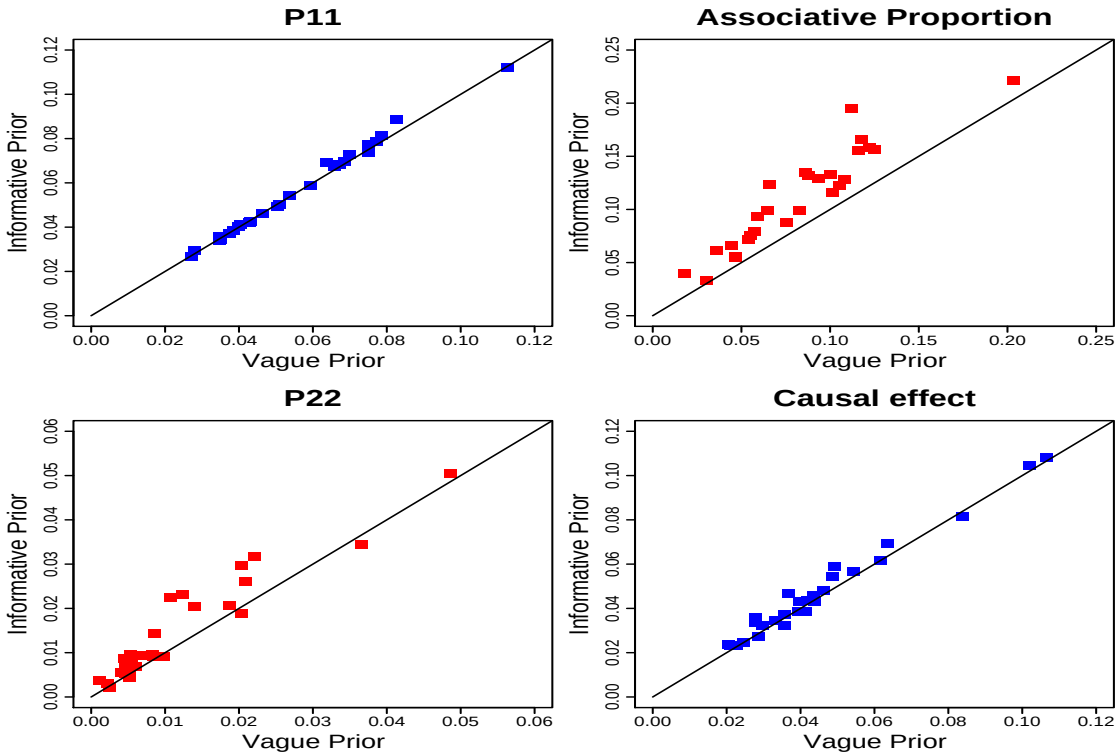

Figure 4.10: Posterior Standard Deviations (SD) of Trial-Specific p_{11}, p_{22}, Associative Proportions and Causal Treatment Effects Based on Informative Priors against Those Based on Vague Priors.

Parameter	Chosen Value
η_{0hz}	$\sqrt{\widetilde{Y}_{hz}/12} - \eta_0 - \eta_1 p_{1hz} - \eta_2 p_{2hz} - \eta_3 p_{3hz} - \gamma_h Z$
η_0	$\sqrt{2.5}$
η_1	$\sqrt{1.5} - \sqrt{2.5}$
η_2	$\sqrt{1.1} - \sqrt{2.5}$
η_3	$\sqrt{0.6} - \sqrt{2.5}$
γ_h	$-\frac{(\sqrt{\widetilde{Y}_{h0}} - \sqrt{\widetilde{Y}_{h1}/12}) - (\sqrt{11.8/12} - \sqrt{10.8/12})}{4}$
σ	0.37

Table 4.16: Description of initial parameter values for AFT model

x as provided in Table 12.1 from the book. In the definition of η_0, η_1, η_2 and η_3 the values 0.6, 1.1, 1.5,and 2.5 represent the approximate median survival times in years for CR, PR, SD, and PD based on Figure 4.5 in the book. In the definition of γ_h, 11.8 and 10.8 were found to be closer estimates to the "total" estimated median survival in months for experimental and standard treatment, respectively, than the values of 9.8 and 8.9 provided in Table 12.1 from the book. Setting σ to be 0.37 appeared to provide survival curves that resembled Figure 4.5 from the book.

Using the AFT model with square-root of survival time, histograms of the HRs by study from 3000 simulations were produced and the mode of the distribution was compared to the actual HR provided in Table 12.2 from the book. To ensure that actual HR was not an extreme observation from the modeled data, the values of γ_h were manually adjusted so that the actual HR was close to the center of the distribution of simulated HRs. All histograms were unimodal and looked nearly normal. It appeared that only eight of the 27 hazard ratio distributions could use shifted by a small amount. Table 4.17 summarizes which studies were adjusted and the value of the γ_h correction that was applied to center the HR distribution near the appropriate value. The final aspect of the data reconstruction involved piecing together subsets of data by trial over 2000 simulations with estimated median survival, $\hat{y}_{mhz}$, where $m = 1, ..., 1000$ for each treatment type closest to the value

Trial	γ_h Correlation
NCOG	+0.10
GOIRC	+0.07
GISCAD	+0.08
RPCI	+0.12
Spain	+0.29
NCIC	−0.08
France	−0.07
MAOP	−0.08

Table 4.17: Summary of the trials that needed manual adjustment to ensure that the HR given in Table 12.2 was near the center of the distribution of simulated HRs

provided in Table 12.1 from the book ($\widetilde{y}_{mhz}$), simultaneously, while having HR that matched the value from Table 12.2 from the book. That is, the quantity $\Delta_{mh} = |\hat{y}_{mh0} - \widetilde{y}_{h0}| + |\hat{y}_{mh1} - \widetilde{y}_{h1}|$ was calculated for each simulation, and the data for the trial corresponding to the smallest Δ_{mh} were combined for each $h = 1, ..., 27$. Recreated versions of Table 12.1 (median survival only) and Table 12.2 (HRs only) from the book are provided by Tables 4.18, 4.19 and 4.20 respectively. The recreated version of Figure 4.5 from the book is pictured in Figure 4.11. The tables and the figure match extremely well, if not identically.

4.10.2 Summary Statistics Used for Data Analysis in Section 8.2

Trial	Treatment	Median Survival
GITSG	5FU+L	11.3
	ST	10.9
NCOG	5FU+L	10.6
	ST	11.2
GOIRC	5FU+L	12.3
	ST	14.6
GISCAD	5FU+L	12.9
	ST	13.1
Genova	5FU+L	11
	ST	10.9
Toronto	5FU+L	12.1
	ST	9.4
City of Hope	5FU+L	14.2
	ST	13.1
RPCI	5FU+L	10
	ST	10.7
Bologna	5FU+L	10.4
	ST	7.4
EORTC	5FU+M	12.1
	ST	9
RPCI	5FU+M	10.3
	ST	11.2
NGTAG	5FU+M+L	8.1
	ST	5.9
AIO	5FU+M+L	10.2
	ST	13.4
NCOG	5FU+M+L	12.8
	ST	11.5
GOCS	5FU+M+L	11.5
	ST	8.9
Mar del Plata	5FU+M+L	0.7
	ST	1
Spain	5FU+M+L	13.4
	ST	11.2

Table 4.18: Recreation of Table 12.1 in *The Evaluation of Surrogate Endpoints* by Molenberghs *et al.* on Meta-Analyses in Advanced Colorectal Cancer: summary Results for 27 trials from reconstructed data from AFT model (no censoring).

Trial	Treatment	Median Survival
MSKCC	HAI	18.3
	ST	14.1
NCCTG	HAI	12.9
	ST	10.7
NCI	HAI	16.7
	ST	11.1
City of Hope	HAI	24.5
	ST	23.4
SWOG	CII	15
	ST	13.8
ECOG	CII	13.1
	ST	10.5
NCIC	CII	10
	ST	9
France	CII	8.4
	ST	9.1
MAOP	CII	10.7
	ST	11
Jerusalem	CII	9.5
	ST	12.4
Total	EX	11.6
	ST	10.8

Table 4.19: Recreation of Table 12.1 in *The Evaluation of Surrogate Endpoints* by Molenberghs *et al.* on Meta-Analyses in Advanced Colorectal Cancer: summary Results for 27 trials from reconstructed data from AFT model (no censoring).

Trial	Hazard Ratio	95% CI
GITSG	0.88	(0.70, 1.10)
NCOG	1.22	(0.88, 1.69)
GOIRC	1.23	(0.92, 1.65)
GISCAD	1.09	(0.81, 1.46)
Genova	0.9	(0.65, 1.25)
Toronto	0.78	(0.55, 1.11)
City of Hope	0.78	(0.49, 1.23)
RPCI	1.13	(0.65, 1.96)
Bologna	0.74	(0.44, 1.23)
EORTC	0.79	(0.63, 0.98)
RPCI	1.28	(0.71, 2.31)
NGTAG	0.76	(0.59, 0.97)
AIO	1.03	(0.75, 1.40)
NCOG	0.89	(0.64, 1.24)
GOCS	0.78	(0.54, 1.11)
Mar del Plata	0.98	(0.59, 1.64)
Spain	1.17	(0.69, 1.97)
MSKCC	0.77	(0.50, 1.17)
NCCTG	0.95	(0.60, 1.51)
NCI	0.81	(0.49, 1.35)
City of Hope	0.91	(0.31, 2.69)
SWOG	0.93	(0.75, 1.14)
ECOG	0.89	(0.71, 1.11)
NCIC	0.8	(0.60, 1.07)
France	0.86	(0.62, 1.19)
MAOP	0.83	(0.61, 1.13)
Jerusalem	1.29	(0.58, 2.86)
Overall	0.91	(0.86, 0.97)

Table 4.20: Recreation of Table 12.2 in *The Evaluation of Surrogate Endpoints* by Molenberghs *et al.* on Meta-Analyses in Advanced Colorectal Cancer: summary Results for binary tumor response and survival for 27 analyzed trials from reconstructed data from AFT model (no censoring).

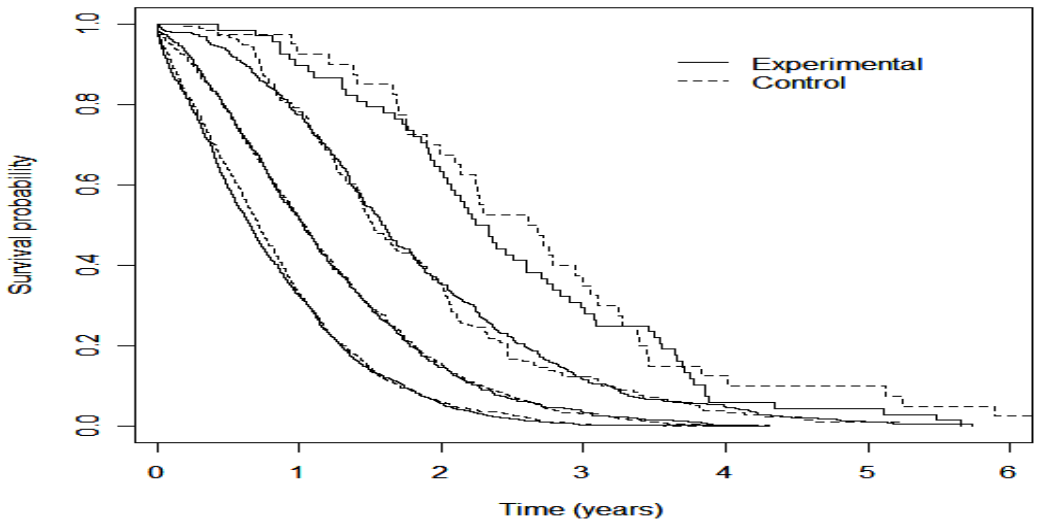

Figure 4.11: Recreation of Figure 4.5 in *The Evaluation of Surrogate Endpoints* by Molenberghs *et al.* on Meta-analysis in advanced colorectal cancer: overall survival curves by tumor responses for the four meta-analyses (advanced colorectal cancer meta-analysis project 1992, 1994, Meta-Analysis Group in Cancer 1996, 1998) using reconstructed data from AFT model.

Trial	Treatment	Bad Tumor Response Survival Status		Good Tumor Response Survival Status	
		Dead	Alive	Dead	Alive
Advanced Colorectal Cancer Meta-Analysis Project (1992)					
GITSG	5FU + L	91	10	7	5
	ST	187	24	32	26
NCOG	5FU + L	37	8	5	5
	ST	78	10	11	8
GOIRC	5FU + L	61	15	7	7
	ST	64	15	5	7
GISCAD	5FU + L	62	18	2	7
	ST	60	12	9	10
Genova	5FU + L	55	12	4	2
	ST	47	12	11	5
Toronto	5FU + L	54	6	2	2
	ST	39	6	10	11
City of Hope	5FU + L	28	7	1	4
	ST	22	2	5	10
RPCI	5FU + L	16	5	0	2
	ST	15	3	8	4
Bologna	5FU + L	28	1	1	0
	ST	21	4	9	0
Advanced Colorectal Cancer Meta-Analysis Project (1994)					
EROCTC	5FU + M	115	21	12	6
	ST	100	25	16	11
RPCI	5FU + M	18	3	0	2
	ST	19	3	1	0
NGTAG	5FU + M + L	119	5	1	2
	ST	97	5	14	6
AIO	5FU + M + L	56	9	8	5
	ST	60	6	9	11
NCOG	5FU + M + L	41	4	6	4
	ST	71	13	8	11
GOCS	5FU + M + L	47	7	5	2
	ST	42	5	6	11
Mar del Plata	5FU + M + L	33	0	0	0
	ST	23	0	5	0
Spain	5FU + M + L	23	5	4	1
	ST	18	2	5	1

Table 4.21: The Number of Patients by Treatment, Tumor Response and Survival Status at 1.75 Years after Treatment for 27 analyzed trials in a Meta-analysis in advanced colorectal cancer. ST - control treatment (bolus 5FU/FUDR); EX - experimental treatment (M - methotrexate; L - leucovorin; HAI - FUDR by hepatic arterial infusion; CII - 5FU by continuous intravenous infusion).

Trial	Treatment	Bad Tumor Response Survival Status		Good Tumor Response Survival Status	
		Dead	Alive	Dead	Alive
Meta-Analysis Group in Cancer (1996)					
MSKCC	HAI	30	10	4	4
	ST	14	8	11	10
NCCTG	HAI	23	6	2	4
	ST	21	2	8	8
NCI	HAI	25	2	0	5
	ST	16	3	5	8
City of Hope	HAI	1	2	1	2
	ST	2	0	2	5
Meta-Analysis Group in Cancer (1998)					
SWOG	CII	121	38	10	13
	ST	118	33	3	20
ECOG	CII	117	17	13	15
	ST	102	15	19	26
NCIC	CII	78	6	3	3
	ST	70	14	7	4
France	CII	65	3	7	3
	ST	52	5	15	5
MAOP	CII	65	12	5	3
	ST	58	4	12	14
Jerusalem	CII	13	1	1	0
	ST	8	2	1	0

Table 4.22: The Number of Patients by Treatment, Tumor Response and Survival Status at 1.75 Years after Treatment for 27 analyzed trials in a Meta-analysis in advanced colorectal cancer. ST - control treatment (bolus 5FU/FUDR); EX - experimental treatment (M - methotrexate; L - leucovorin; HAI - FUDR by hepatic arterial infusion; CII - 5FU by continuous intravenous infusion).

4.11 References

Alonso, A., Molenberghs, G. (2003). Surrogate Marker Evaluation from an Information Theory Perspective. Biometrics. **63**, 180-186.

Angrist, J., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*. **91**, 444-472.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*. **92**, 1171–1176.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal Royal Statistical Society B*. **25**, 220-233.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2004) The Evaluation of surrogate endpoints. Chapter 4. Springer.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. **1**, 49–68.

Chen, H., Geng, Z., and Jia, J. (2007). Criteria for surrogate endpoints. *Journal Royal Statistical Society B*. **69**, 919-932.

Fleming T.R., DeMets D.L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*. **125**, 605-613.

Forster, J.J. (2004) Bayesian inference for poisson and multinomial log-linear models. *Working paper.*.

Frangakis, C.E. and Rubin, D.B. (2002). Principal stratification in casual inference.

Biometrics. **58**, 21–29.

Frangakis, C.E., Rubin, D.B. and Zhou, X.H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advanced directive forms. *Biostatistics*. **3**, 147–164.

Freedman, L.S., Graubard, B.I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine*. **11**, 167–178.

Gail, M., Pfeiffer, R., Houwelingen, H.C.V., and Carrol, R.J (2000). On Meta-analytic assessment of surrogate outcomes. *Biostatistics*. **1**, 231–246.

Garret, E.S., Zeger, S.L (2000). Latent class model diagnosis. *Biometrics*. **56**, 1055–1067.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. New York: Chapman and Hall.

Gilbert, P.B. and Hudgens, M.G. (2008). Evaluating causal effect predictiveness of candidate surrogate endpoints. *Biometrics*, in Press.

Hirano, K., Imbens G.W., Rubin, D.B., and Zhou, X.H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*. **1**, 69–88.

Imbens, G.W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*. **25**, 305–327.

Lindley, D.V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*. **35**, 1622–1643.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. 2nd Edition. Wiley: New York.

Musch D.C., Lichter P.R., Guire K.E., Standardi C.L., CIGTS Investigators (1999): The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology*. **106**: 653–62.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials, definition and operational criteria. *Statistics in Medicine*. **8**, 431–440.

Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology*. **3**, 143-155.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *The Journal of the Royal Statistical Society, Series A*. **147**, 656–666.

Rubin, D.B. (1978). Bayesian-inference for causal effects - role of randomization. *Annals of Statistics*, **6**, 34–58.

Rubin, D.B. (1980). Randomization analysis of experimental-data - the Fisher randomziation test - comment. *Journal of American Statistical Association*, **75**, 591–593.

Taylor, J.M.G., Wang, Y., and Thiébaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*. **61**, 1102–1111.

Wang Y., Taylor J.M.G. (2003): A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*. **58**, 803-812.

CHAPTER V

Summary and Future Work

In Chapters II and III, we considered the use of a surrogate marker as an auxiliary variable in estimating the treatment effect in clinical trials. In Chapter II, we examined the factors that impact the degree of efficiency gain from S in estimating the treatment effect in both the single- and multiple-trial settings. While previous research results have mixed opinions on the value of surrogate markers (Murray and Tsiatis, 1996; Venkatraman and Begg, 1999; Fleming and DeMets, 1996), we have identified the scenarios that a surrogate marker can be very useful in increasing the precision of the treatment effect estimate. In a single-trial setting, the efficiency gain is small unless S and T are very highly correlated and the amount of missingness is substantial. In a multiple-trial setting, higher efficiency gain is associated with higher trial-level correlation but not individual-level correlation when only S, but not T is measured in a new trial; but, the amount of information recovery from S is negligible. However, when T is partially observed in the new trial and the individual-level correlation is relatively high, there is substantial efficiency gain by using S and one can extract most of the information on the treatment effect. In our study, both S and T are continuous; in reality, however, it is more common that both S and T are time to event. For example, time-to-recurrence and time-to-disease-progression as

surrogate markers, and survival time as the primary endpoint in studies with colon cancer patients or head-and-neck cancer patients. Since it could take too long or cost much to collect the information on time to death, it would be an important extension of our current work to investigate the extent of information recovery from S in terms of the effect of the treatment on improving patients' longevity.

In Chapter II, we proposed a fully Bayesian estimation method for the setting where Z and S are completely observed but T is not completely observed in a new trial; the proposed method was in a meta-analytic framework. The required computation of the Bayesian estimation is too intensive to permit extensive evaluation of the properties of the proposed method. However, based on some limited simulations on a small number of trials, we found that the coverage rates of the credible intervals tend to be less than the nominal level when we assume noninformative priors for the fixed effects and diffused inverse Wishart distributions as the prior for the between-trial and within-trial variances. While the fully Bayesian method incorporates all uncertainty associated with estimating every single parameter (Louis and Zelterman (1994)), it seems that in our research setting, the results are sensitive to the diffused inverse Wishart distributions. It will be a useful extension to find appropriate priors that lead to estimates with nominal-level coverage rates and unbiasedness.

In Chapter III, we proposed a generalized ridge regression method to incorporate information from an auxiliary variable, S, to estimate the treatment effect in a randomized trial setting when S and T are continuous. The method avoids the need to know the correct surrogacy assumption and allows for the uncertainty in the models that could describe such an assumption. The proposed method can be seen as striking a balance between bias reduction and efficiency gain, depending on the nature of the relationship between S and T. We intend to extend the method in several

directions. First, we can adapt the method to different data types. When S and T are binary data, we have observed the same phenomena that when S satisfies the perfect surrogacy assumption, we can obtain substantial efficiency gain by utilizing S in estimating Q (results not shown). A generalized ridge regression method could be developed a similar fashion to that when S and T are continuous. The setting where S and T are time-to-event data presents a more complicated challenge, as we need to consider censoring for both S and T. We consider both S and T are censored event times and that S always occurs before T. In advanced colorectal cancer, for example, S could be time to recurrence and T could be survival time. Cook and Lawless (2001) have shown that great efficiency gains can be obtained by using S under a three-state model, where an intermediate disease state will be entered prior to death. For individual i, the models for the hazard functions of $S|Z$ and $T-S|S,Z$ at time t can be specified as:

$$\lambda_S(S) = \lambda_{0S}(s)\exp(\alpha_1 z)$$

$$\lambda_{ST}(T-S) = \lambda_{0T}(t-s)\exp(\beta_1 s + \beta_2 z), \quad t > S.$$

The corresponding probability density functions of $S|Z$ and $T-S|S,Z$ are given by:

$$g(S|Z) = \lambda_{0S}\exp(\alpha_1 z)\exp(-\Lambda_{0S}(s)\exp(\alpha_1 z)),$$

$$h(T-S|S,Z) = \lambda_{0T}(t-s)\exp(\beta_1 s + \beta_2 z)\exp(-\Lambda_{0T}(t-s)\exp(\beta_1 s + \beta_2 z)),$$

where $\Lambda_{0S}(s) = \int_0^s \lambda_{0S}(u)du$ and $\Lambda_{0T}(t-s) = \int_0^{t-s} \lambda_{0T}(u)du$. Then the probability density function of $T|Z$ is:

$$f(T|Z) = \int_{s=0}^{t} h(t-s|s,z)g(s|z)ds.$$

We define the treatment effect on T, $Q(T_\nu)$, being the difference in the truncated life expectancy between the two treatment groups where ν is the maximum observed

survival time for T. The truncated life expectancy is denoted by $E(T_\nu|Z)$. Then we have:

$$E(T_\nu|Z) = \int_{t=0}^{\nu} t f(t|z)dt$$
$$Q(T_\nu) = E(T_\nu|Z=1) - E(T_\nu|Z=0)$$

We will investigate the shrinkage effect on β_2 when S can capture most of the treatment effect on the residual time $t - S$ after S has occurred. The method could be illustrated on data from clinical trials in advanced colorectal cancer. We could apply a normal prior in a same fashion as we have done to implement the ridge regression method. We could also examine different prior distributions (such as a mixture prior with point mass at $\beta_2 = 0$ or a Dirichlet process prior) can be used instead of the normal prior in order to allow for more probabilities for the parameter to be exactly 0.

We could also extend the ridge regression method to a general missing data problem setting. It is essentially a missing data problem to incorporate S in predicting the effect of treatment on T when Z and S are completely observed but T is not. We have found that the ridge regression has the data-adaptive and robust features; it also has superior properties in terms of MSE, bias and coverage rates when S and T is closely associated, compared with competing methods. We could adapt the ridge regression to more general missing data problems and impute draws to fill in the missing data. We could compare the use of the ridge regression with the commonly used regression method that assumes multivariate normality (Rubin, 1987), or a sequential regression imputation method (Raghunathan *et al.*, 2001). It is likely that the ridge regression maintains the same advantages we have found in the surrogate marker setting. It is also possible that a hybrid method that can perform both model

selection and shrinkage such as the least absolute shrinkage and selection operator method (LASSO) (Tibshirani, 1996), would have the similar advantages as ridge, in addition to the features that would go along with its model selection capacity. We have compared the ridge regression method with the inverse probability weighted method and have shown that the ridge method is more data-adaptive and has better MSE property when S and T is closely associated. The IPW method we used for comparison is proposed by Zhao and Lipsitz (1992). It would also be interesting to compare ridge with the improved versions of the IPW method which were proposed and studied by Robins *et al.* (1994) and Scharfstein *et al.* (1999) and have the double robustness features.

The second aspect of my dissertation involves modeling the association between S and T in a causal inference framework. Previous surrogacy measures require one to fit models for the distribution of T given S and Z, which does not have a causal interpretations because S is a post-randomization variable. We proposed a Bayesian estimation which incorporates assumptions that are plausible in the surrogate context by using prior distributions to reduce the nonidentifiability problem and possibly increase precision in both the single- and multiple-trial settings. We can extend this work in many possible ways. One of the extensions is to go beyond binary S and binary T. For example, when both S and T are continuous, we can model the joint distribution of the potential outcomes of S and T using a parametric form. The association measures can be correlations instead of the probability measures used in our current work. Another important extension is to incorporate covariates that are predictive of the potential outcomes of S and T and generalize our work to observational data. We have noticed that the causal surrogacy measures obtained using our method in either a single trial or a multiple trial setting have wide credible

intervals. One would think that the precision of these quantities can be increased with the incorporation of predictive covariates. While the parametric modeling in our work allows us to maximize the efficiency, it can also be useful to calculate the non-parametric bounds to quantify the range of the counterfactual probabilities in our context (Balke and Pearl, 1997).

One of the key assumptions in our method is the monotonicity assumption that requires that if a patient gets better if received $Z = 0$, she or he would not get worse if received $Z = 1$. It is essential to make this assumption to reduce the number of parameters to have a more identifiable counterfactual model. It is not usually contradicted by the data where on average patients do not become worse off when they received $Z = 1$ compared to those received $Z = 0$. If this assumption is correctly specified, we expect our estimates for the quantities of interest will be more efficient and less biased than the conventional model if the quantities are comparable. However, this assumption requires that every single patient would have done at least as well as that when she or he receives $Z = 1$ relative to that when she or he receives $Z = 0$. It is perhaps true for most of the patients but not usually obviously satisfied for all patients, for example, in the CIGTS study where we compare the effect of medicine with that of surgery, it is conceivable that some patients may be better if they received medicine instead of surgery, even though the average effect of surgery is consistently better. On the other hand, for the colon cancer study, the assumption is more likely to hand for every individual because the experimental treatments are only slight modifications of the standard treatment and are intended to improve its effectiveness. Assessing the impact of the violations of the monotonicity assumption is an important extension in the current work.

It is also important to realize that the probabilities in the counterfactual model

are association measures instead of causation measures. If S is in the causal pathway between Z and T, p_{22} in Table 4.3 is very large. On the other hand, a very high p_{22} only shows that the causal effect of Z on S is highly associated with the causal effect of Z on T and it does not necessarily imply that Z affects T by affecting S. It is likely that there is an unmeasured variable denoted by U which can be a post-treatment confounder that can affect both S and T. Consequently, adjusting for S may induce the false association between Z and T. S is defined as "collider" in the economic literature. This problem can be directly addressed by another causal inference framework proposed by Robins and Greenland (1992) which allows one to manipulate S, which has been used by Chen, Geng and Jia (2007) to study the surrogacy consistency. The framework defines additional probabilities to describe the likelihood of how T changes by intervening S, which can measure the degree to which Z affect T through affecting S. In this framework, the effect of Z affecting T through S is called indirect effect, while the effect of Z affecting T not through S is called direct effect. We could use a Bayesian estimation to incorporate appropriate prior distributions in this framework to reduce the identifiability problem and obtain the quantities of interest.

5.1 References

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92**, 1171–1176.

Chen, H., Geng, Z., and Jia, J. (2007). Criteria for surrogate endpoints. *Journal Royal Statistical Society B*. **69**, 919-932.

Cook R.J., Lawless J.F. (2001). Some comments on efficiency gains from auxiliary information for right-censored data. *Journal of Statistical Planning and Inference*. **96**, 191-202.

Fleming T.R., DeMets D.L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*. **125**, 605-613.

Louis, T.A., and Zelterman, D. (1994). Bayesian approaches to research synthesis. In H. Cooper, and L. V. Hedges (Eds.), The handbook of research synthesis. New York: Russell Sage Foundation.

Murray, S and Tsiatis, A. A. (1996), Nonparametric Survival Estimation Using Prognostic Longitudinal Covariates, *Biometrics*. **52**, 137-151.

Raghunathan, Lepkowski, Van Hoewyk and Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, June 2001.

Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology*. **3**, 143-155.

Robins, J.M. Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*. **89**, 846–866.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, New York: John Wiley and Sons, Inc.

Scharfstein D.O., Rotnitzky A, and Robins J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association*. **94**, 1096-1120.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*. **58**, 267-288.

Venkatraman E.S., Begg C.B. (1999). Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics*. **55**, 1171-1176.

Zhao L.P., and Lipsitz S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*. *11*. 769-82.