Missing Data and Variable Selection Methods for Cure Models in Cancer Research

by

Lauren J Beesley

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2018

Doctoral Committee:

Professor Jeremy M G Taylor, Chair
Professor Thomas M Braun
Professor Roderick J Little
Associate Professor Laura M Rozek
Research Associate Professor Matthew J Schipper

Lauren J Beesley

lbeesley@umich.edu

ORCID iD: 0000-0002-3788-5944

Dedication

To Woody, my partner in crime. All the love.

Acknowledgements

There are so many people who have provided me with guidance, insight, direction, support, and encouragement during my time as a PhD student, and I would like to express my genuine gratitude to these people.

First, I would like to thank my PhD advisor, Jeremy Taylor. I started working with Jeremy my first year as a PhD student, and I have met with him nearly every week since then. He has been generous with his time, and he has never failed to give insightful comments. He has also provided financial support throughout my time as a PhD student. I consider him a mentor in every sense of the word. He has helped me gain the knowledge and confidence to stand on my own as a researcher, and for that I will always be grateful.

I would also like to thank my dissertation committee. Rod Little has spent a great deal of time giving feedback on some missing data-related ideas in this dissertation, and his suggestions have helped me greatly improve both my ideas and the way I present them. His feedback has really helped me grow as a researcher. Tom Braun has been a strong source of guidance and support throughout my PhD. I have worked with Matt Schipper on several collaborative projects, and I have learned a lot from him. Laura Rozek has generously given her time to be part of my committee, and she has contributed some very helpful feedback. I would also like to thank Tim Johnson for always keeping his door open for my many drive-by questions about Bayesian methods.

My years as a PhD student have been some of the best of my life, and for that I want to thank my biostats family, the students and faculty that have made Ann Arbor feel like home. In particular, I want to thank Julie Strominger, Krithika Suresh, Anagha Tolpadi, Wenting Cheng, Allison Cullen Furgal, Marco Benedetti, Lauren Ziemba, Michelle Earley, Evan and Riley Reynolds, Nick Seewald, Emma Beyers-Carlson, Jed Carlson, Sarah Abraham, and Yilun Sun. I am so grateful for your friendship.

I also want to thank my parents, Kari and Darin Beesley. They have always been there for me, and I would not be where I am now without them. Through all the ups and downs, through the tears and the excitement and the joy, they have never wavered in their love and support. They are strong, kind, wise, and generous. Every day, I try to live up to their example. I want to also thank my sisters, Allison Beesley and Jennifer Hawkins.

Finally, I want to thank my friend, Jenny Zhong, and my partner, Woody VanDervort. Jenny has been a never-ending source of inspiration, hope, love, and vision. She has been my cheerleader and my confidante since middle school. Woody is my best friend and the love of my life. Every day, these two remarkable people remind me what really matters in life.

Table of Contents

Chapters

List of Tables

List of Figures

List of Appendices

Abstract

In survival analysis, a common assumption is that all subjects will eventually experience the event of interest given long enough follow-up time. However, there are many settings in which this assumption does not hold. For example, suppose we are interested in studying cancer recurrence. If the treatment eradicated the cancer for some patients, then there will be a subset of the population that will never experience a recurrence. We call these subjects "cured."

The Cox proportional hazards (CPH) mixture cure model and a generalization, the multistate cure model, can be used to model time-to-event outcomes in the cure setting. In this dissertation, we will address issues of missing data, variable selection, and parameter estimation for these models. We will also explore issues of missing covariate and outcome data for a more general class of models, of which cure models are a particular case.

In Chapter II, we propose several chained equations methods for imputing missing covariates under the CPH mixture cure model, and we compare the novel approaches with existing chained equations methods for imputing survival data without a cured fraction.

In Chapter III, we develop sequential imputation methods for a general class of models with latent and partially latent variables (of which cure models are an example). In particular, we consider the setting where covariate/outcome missingness depends on the latent variable, which is a missing not at random mechanism.

In Chapter IV, we develop an EM algorithm for fitting the multistate cure model. The existing method for fitting this model requires custom software and can be slow to converge. In contrast, the proposed method can be easily implemented using standard software and typically converges quickly. We further propose a Monte Carlo EM algorithm for fitting the multistate cure model in the presence of covariate missingness and/or unequal censoring of the outcomes.

In Chapter V, we propose a generalization of the multistate cure model to incorporate subjects with persistent disease. This model has many parameters, and variable selection/shrinkage methods are needed to aid in estimation. We compare the performance of existing variable selection/shrinkage methods in estimating model parameters for a study of head and neck cancer.

In Chapter VI, we develop Bayesian methods for performing variable selection when we have order restrictions for model parameters. In particular, we consider the setting in which we have interactions with one or more order-restricted variables. A simulation study demonstrates promising properties of the proposed selection method.

Chapter I

Introduction

One goal of cancer research is to identify patient characteristics (clinical, demographic, or molecular biomarkers) related to health outcomes such as time to death or time to disease recurrence. A clear understanding of the relationship between characteristics and outcomes can be used for prediction and inform medical decision-making. With the increasing availability of patient information (from past medical records, new diagnostics, genetic testing, etc), there is a strong need to develop statistical methods to handle the challenges presented.

One substantial challenge is that of missing data. Missingness may occur for a variety of reasons. For example, not all patients may undergo the same diagnostic testing, resulting in missingness in the test results for some patients. Data may be combined across multiple hospitals, and these hospitals may collect information differently. Data collected over time may have missingness due to missed doctors appointments or loss to follow-up. These types of missing data are particularly common in observational data, which are often used in cancer research. Missing data may also arise from the conceptual framework used to model the data. In the study of cancer recurrence, for example, we sometimes introduce a partially latent variable representing whether the subject was cured of their cancer by their initial treatment. When we have loss to follow-up, cure status is only known for subjects with observed recurrences, resulting in an induced source of missing data. Statistical methods are needed to account for the missing information appropriately.

Another challenge arising from increased data availability is that of variable selection. In the setting in which many predictors are available for each subject (large p), statistical methods are needed to determine which of the variables are important and should be included in the model. Inclusion of too many predictors can result in numerical issues and overfitting. Additionally, greater data availability opens the door for more complicated modeling strategies. For example, multistate models in survival analysis can incorporate information from multiple time-to-event outcomes and are incredibly useful for prediction and for identifying the impact of predictors on different parts of the disease process. With even a modest number of covariates, however, these models can quickly end up with an intractable number of model parameters, and variable selection or shrinkage methods are needed to produce good model inference.

My dissertation will broadly consist of five projects, each of which tackles an issue of missing data or variable selection arising in the study of cancer. The methods we develop, however, can be applied to other diseases and different scientific questions. In particular, we are interested in exploring issues of missing data and variable selection for cancer data when there is a cured fraction of the population. We suppose that we are interested in studying cancer recurrence after initial treatment. If the treatment eradicated the cancer for some patients, then there will be a subset of the population that will never experience a recurrence. We call these subjects "cured" of their primary cancer. Before introducing the statistical methods explored in this dissertation, we describe the dataset motivating the methodological development.

This dissertation is broadly motivated by data collected by the UM Head and Neck Cancer Specialized Program of Research Excellence (SPORE). After initial treatment for head and neck cancer, patients were followed for recurrence and death. Covariate information was also collected at baseline. It is been well-established that some head and neck cancer patients can be cured of their cancer through their primary treatment, and the data further support the hypothesis that some subjects were cured (Taylor, 1995; Grau et al., 1997; Cognetti et al., 2008). This cure setting also occurs for some other types of cancers such as breast cancer. Our general interest lies in studying the association between baseline covariates and the rate of recurrence, the rate of death after recurrence, and the probability of being cured by treatment.

Several existing frameworks are available for modeling recurrence time data with a cured fraction. The Cox proportional hazards mixture cure model is a common modeling strategy (Sy and Taylor, 2000), and recently Conlon et al. (2013) proposed a generalization of the mixture cure model called the multistate cure model that can also incorporate death information. When we apply existing estimation methods to the head and neck dataset, however, several problems arise. Firstly, HPV (human papillomavirus) status is unavailable for roughly 50% of the subjects, and a small amount of missingness was present in other study variables. Existing missing data methods for Cox proportional hazards mixture cure model often involve modeling the joint distribution of the covariates, which may not be easily done and may be restrictive. Additionally, these methods make MAR assumptions, which may not always hold in practice. No methods have been developed for dealing with missing covariates in the multistate cure model setting. Secondly, for many patients (about 60%), follow-up for recurrence was substantially shorter than follow-up for death. For some subjects, this results in a time interval in which death status is known but recurrence status is unknown. This creates missing data in the outcome information. Little work has been done to address this issue. Thirdly, even with a modest number of covariates, the number of parameters in the cure model and multistate cure model can become large, which motivates the development of variable selection methods. In tackling these issues, we must keep in mind that cure status is only known for subjects with observed recurrences, which presents a further source of missing information. In this dissertation, we propose statistical methodology to address these issues.

In **Chapter II**, we develop chained equations methods for imputing missing covariates for the Cox proportional hazards mixture cure model, and we will compare the novel approaches with existing chained equations methods for imputing survival data without a cured fraction. Simulations demonstrate improved performance of the proposed method (in terms of bias of cure model parameters) over existing methods.

In **Chapter III**, we explore sequential imputation methods for a more general class of models with latent and partially latent variables (of which cure models are a particular example). In particular, we consider the setting where covariate or outcome missingness depends on the latent variable, which is a missing not at random mechanism (Little and Rubin, 2002). The proposed methods represent the first thorough exploration into the implementation of sequential imputation methods for general latent-dependent missingness.

In **Chapter IV**, we develop an EM algorithm for fitting the multistate cure model. The existing method in the literature for fitting this model requires custom software and can be slow to converge. In contrast, the proposed method can be easily implemented using standard software and typically converges quickly. We further propose a Monte Carlo EM algorithm for fitting the multistate cure model in the presence of covariate missingness and/or unequal censoring of the outcomes.

In **Chapter V**, we propose a generalization of the multistate cure model to incorporate subjects with persistent disease. Like many multistate models, this model has many parameters, and variable selection/shrinkage methods are needed to aid in estimation. However, such methods have not previously been explored in the multistate modeling context. We compare the performance of existing variable selection/shrinkage methods in estimating model parameters for the head and neck cancer data.

In **Chapter VI**, we develop Bayesian methods for performing variable selection when we have order restrictions for model parameters. In particular, we consider the setting in which we have interactions with one or more order-restricted variables. A simulation study demonstrates promising properties of the proposed method.

Chapter II

Covariate Imputation for the CPH Cure Model

2.1 Introduction

In survival analysis, a common assumption is that all subjects will eventually experience the event of interest given long enough follow-up time. However, there are many settings in which this assumption does not hold. For example, suppose we are interested in studying cancer recurrence in patients treated for head and neck cancer. If the treatment completely eradicated the cancer in some individuals, then there will be a subset of the population that will never experience a recurrence. We call these subjects "cured" or "non-susceptible." We note that cure status is only known for subjects with observed recurrences.

One commonly used modeling approach for survival data with a cured fraction is a mixture model with two components. The first component is a model for the probability that a subject is not cured, which is usually modeled using logistic regression. The second component is a model for the failure time in the susceptible (non-cured) population. Parametric, semiparametric, and nonparametric formulations of the failure time model exist in the literature (Farewell, 1982; Yamaguchi, 1992; Lu and Ying, 2004; Kuk and Chen, 1992; Peng and Dear, 2000; Sy and Taylor, 2000; Zhuang et al., 2000). We consider a formulation of the mixture cure model where failure time in the susceptible population is modeled using a Cox proportional hazards regression model (Kuk and Chen, 1992; Sy and Taylor, 2000; Cox, 1972). It is important to note that subjects with observed events are known to be non-cured, but cure status is not known for censored subjects. Cure models are appealing because they enable enhanced interpretation and inference from data with a cure structure as cure models allow us to model both the probability that a subject is cured and the hazard of an event in the non-cured group separately.

A challenge that arises in the application of these cure models is that often one or more covariates are only partially observed. One simple approach is to ignore the missing data and analyze only the patients with complete covariate data. "Complete case" analysis is an undesirable approach since it does not use data from patients with missing covariate values and is therefore inefficient. Also, complete case analysis may be biased if the covariate missingness mechanism depends on the outcome. Other approaches in the literature for handing missing covariates in the cure setting often involve modeling the joint distribution of the missing covariates using general location models (Zhuang et al., 2000; Cho et al., 2001) or by specifying a series of conditional distributions (Chen and Ibrahim, 2002). Both approaches require us to explicitly specify the joint distribution of the covariates, which may not be easily done, and they are not easily implemented using standard software.

In this chapter, we explore multiple imputation as another approach for handling missing data in the cure model setting. When performing multiple imputation, it is important to include outcome information in the model for imputing partially observed covariates (Moons et al., 2006). In the cure setting, however, many aspects of the outcome (cure status and event times in the non-cured subjects) are not fully observed due to censoring. We are interested in comparing different methods for incorporating the observed outcome information to impute partially observed covariates when the primary outcome has a Cox proportional hazards cure structure. We will study covariate imputation approaches using fully conditional specification.

Fully conditional specification (FCS) is a multiple imputation approach in which we specify a conditional distribution for each partially observed covariate (Van Buuren et al., 2006; Raghunathan, 2001). We then use these conditional distributions to impute covariates as part of an iterative algorithm that cycles through the conditional distributions for all the partially observed covariates. This often involves specifying a regression model for each partially observed covariate and then using the regression models to impute the missing values. An attractive feature of FCS is that it does not require us to explicitly specify the joint distribution of the covariates.

Suppose X is a set of covariates and Y is an outcome variable. Also, suppose our ultimate goal is to fit a standard regression model for $Y|X$ (e.g. linear, logistic). Let $X^{(p)}$ denote the p^{th} covariate in X and $X^{(-p)}$ denote all covariates in X except $X^{(p)}$. We would like to use the distribution of $X^{(p)}|X^{(-p)}, Y$ to impute each partially observed $X^{(p)}$. If we have the distributions for $Y|X$ and $X^{(p)}|X^{(-p)}$, then we can derive the distribution for $X^{(p)}|X^{(-p)}, Y$ directly. When $X^{(p)}|X^{(-p)}$ and $Y|X$ are normally distributed with predictors incorporated in the mean structure, then the distribution of $X^{(p)}|X^{(-p)}, Y$ will also be normal and will correspond to a linear regression that can be readily used to impute $X^{(p)}$. When the true distribution of $X^{(p)}|X^{(-p)}, Y$ is unknown or difficult to sample from, we may attempt to approximate the distribution using a simpler and more computationally convenient standard regression model. For example, for normal $X^{(p)}$, we may specify the distribution of $X^{(p)}|X^{(-p)}, Y$ using some function of $X^{(-p)}$ and Y as predictors in a linear regression model.

In survival analysis, the primary outcome usually consists of the pair (Y, δ). If T is the underlying event time and C is the censoring time, then $Y = \min(T, C)$ and $\delta = I(T \leq C)$. The ultimate goal is usually to fit a model for $T|X$. Although T is the outcome of interest, it is not directly observed due to censoring. We can still derive the exact distribution of $X^{(p)}|X^{(-p)}, Y, \delta$ to impute each partially observed $X^{(p)}$. However, due to the complicated structure of survival data, the exact distribution of $X^{(p)}|X^{(-p)}, Y, \delta$ will often be inconvenient or computationally intensive to sample from (Bartlett et al., 2014).

One possible alternative is to obtain a more convenient approximation to the exact conditional distribution of $X^{(p)}|X^{(-p)}, Y, \delta$ for each partially observed covariate $X^{(p)}$. White and Royston (2009) derived an approximate conditional distribution for proportional hazards survival data that reduced to a regression model of $X^{(p)}$ with predictors $X^{(-p)}, \delta,$ and $\hat{H}_0(Y)$, where $\hat{H}_0(Y)$ is the estimated cumulative baseline hazard function. One adaptation of this would be to using log(Y) in place of $\hat{H}_0(Y)$ (Van Buuren et al., 1999). Another adaptation would be to use a regression model for $X^{(p)}$ with predictors $X^{(-p)}, \delta f_1(Y),$ and $(1-\delta) f_2(Y)$, where $f_1(Y)$ and $f_2(Y)$ are functions of Y specified using splines or step functions.

Additionally, since $Y = \min(T, C)$ is a mixture of a censoring time and the event time of interest, it may not be appealing to include Y in the imputation regression models, and we may instead wish to incorporate T directly. We can treat T as another partially observed variable and impute the value of T from the distribution of $T|T > C, X$ for censored subjects. Assuming C is uninformative for $X^{(p)}$, we can then try to impute each partially observed $X^{(p)}$ by specifying the exact conditional distribution $X^{(p)}|X^{(-p)}, T$ or by approximating the exact distribution with a regression model using T.

When the ultimate goal is to fit a mixture cure model, the form for the distribution of $T|X$ is more complicated. The most convenient estimation method introduces a partially observed variable, G, which indicates cure status. Either an imputed value or the expectation of G is used in the mixture cure model estimation algorithm (Sy and Taylor, 2000). When we have partially observed covariates, we can impute each partially observed $X^{(p)}$ from the corresponding distribution of $X^{(p)}|X^{(-p)}, Y, \delta, G$. Using assumptions for the distribution of $X^{(p)}|X^{(-p)}$, we can derive the exact conditional distribution from which to impute. We can also impute using approximations to the exact conditional distribution that are more computationally convenient. Alternatively, we can impute the event time T for censored individuals and then impute each partially observed $X^{(p)}$ using the approximated conditional distribution of $X^{(p)}|X^{(-p)}, T, G$.

In this chapter, we derive the exact conditional distribution and suggest a sampling scheme for imputing partially observed covariates in the Cox proportional hazards mixture cure model setting. Additionally, we propose several approximations to the exact distribution that are more convenient to use for imputation. We compare the performance of our proposed imputation approaches to methods for survival data without a cure fraction.

In **Section 2.2**, we present details about the Cox proportional hazards cure model. In **Section 2.3**, we present possible approaches for imputing partially observed covariates in the cure setting. In **Section 2.4**, we report results from a set of simulations and compare the performance of the imputation algorithms. In **Section 2.5**, we apply two imputation approaches to a study of cancer recurrence in head and neck cancer patients, and in **Section 2.6** we present a discussion.[1]

[1] A version of this chapter has been previously published in Beesley et al. (2016).

2.2 Cox Proportional Hazards Cure Model

We consider the setting where the primary outcome is a censored event time and there is an underlying subset of the study population that will never experience the event of interest. We call individuals that will never experience the event "cured." The Cox proportional hazards (CPH) cure model is a mixture model with two components: 1) a model for the probability that an individual is not cured and 2) a Cox proportional hazards model for the hazard of an event for non-cured subjects (Kuk and Chen, 1992).

Let $Y_i = \min(T_i, C_i)$ be the observed event/censoring time for individual i where T_i is the underlying event time (defined as infinity if a subject is cured) and C_i is the censoring time. Let $\delta_i = I(T_i \leq C_i)$. We define the cure status of individual i, G_i, as 1 when the individual is not cured and 0 when the individual is cured. G_i is 1 when $\delta_i = 1$ and is unknown when $\delta_i = 0$. We assume censoring is independent of G and T given covariates. We model the data as follows:

$$\text{Logistic Model of Cure Status: } \operatorname{logit}(P(G_i = 1|X_i)) = \alpha_0 + \alpha^T X_i \qquad i = 1, ..., n$$

$$\text{CPH Model of Failure Time: } h(t|X_i, G_i = 1) = h_0(t)e^{\beta^T X_i} \qquad i = 1, ..., n$$

where $h_0(t)$ is the baseline hazard of having an event in the non-cured group. For simplicity, we assume that we have the same set of covariates in both parts of the mixture model. Estimation of model parameters can be done using an EM algorithm (Peng and Dear, 2000; Sy and Taylor, 2000).

We consider the complete data partial log-likelihood corresponding to the CPH cure model assuming that G_i is observed. The EM algorithm iterates between two steps. In the E-step for a given iteration, we replace G_i in the complete data log-likelihood with

$$w_i = E(G_i|\delta_i, Y_i, X_i) = \delta_i + (1 - \delta_i)\frac{p_i S(Y_i|X_i, G_i = 1)}{1 - p_i + p_i S(Y_i|X_i, G_i = 1)} \tag{2.1}$$

Here, $p_i = P(G_i = 1|X_i) = \operatorname{expit}(\alpha_0 + \alpha^T X_i)$ and $S(Y_i|X_i, G_i = 1) = e^{-H_0(Y_i)e^{\beta^T X_i}}$ using the estimates of α_0, α, and β from the previous iteration and an estimate of $H_0(t)$ obtained using a Breslow estimator weighted by w_i (Breslow, 1972). To improve the stability of the EM algorithm (model parameters are nearly unidentifiable), we define censored individuals with very late censoring times as cured with $w_i = 0$ (Sy and Taylor, 2000). The M-step involves taking the complete data partial log-likelihood with w_i substituted for G_i and maximizing it with respect to α_0, α, and β. The EM algorithm allows us to handle the fact that cure status is only partially observed. Variances of model parameter estimates can be estimated via bootstrap.

2.3 Multiple Imputation of Missing Covariates

In this section, we discuss imputation by fully conditional specification in more detail. Then, we derive the exact conditional distribution to impute partially observed covariates in the cure setting. We also present several approximations to the exact distribution that are more convenient to use for imputation. We include several covariate imputation models for survival data without a cured fraction.

2.3.1 Fully Conditional Specification Approach

Fully conditional specification (FCS) or "chained equations" is a multiple imputation approach in which we specify the conditional distribution for each partially observed variable and then use these distributions to impute variables one-by-one as part of an iterative procedure (Van Buuren et al., 2006; Raghunathan, 2001). When imputing variable V, we first specify the full conditional distribution for V (with parameter v) given all the other variables. This may be an approximation of the "exact" conditional distribution. We then impute V by 1) drawing v from its posterior distribution and then 2) drawing V using its full conditional distribution at the drawn value of v. We then iterate between univariate imputation steps for the variables with missingness.

Suppose we are interested in fitting a model to outcome O with partially observed covariates $W = (X^{(1)}, \dots, X^{(d)})$ and fully observed covariates $Z = (X^{(d+1)}, \dots, X^{(s)})$. Let $X = (W, Z)$. Recall that $X^{(p)}$ denotes the p^{th} covariate in X and $X^{(-p)}$ denotes all covariates in X except $X^{(p)}$. For each partially observed $X^{(p)}$, we specify the conditional distribution $f(X^{(p)}|X^{(-p)}, O; \phi^p)$ where ϕ^p is a set of parameters. Let $f(\phi^p|X, O)$ denote the posterior distribution of ϕ^p and let $X^{(p,miss)}$ and $X^{(p,obs)}$ denote the missing and observed portions of $X^{(p)}$. To impute missing values for $X^{(1)} \dots X^{(d)}$, we perform the following iterative chained equations algorithm. At iteration k, we obtain updated imputed values by drawing

$$\begin{aligned}
&\phi^1_{(k)} \sim f(\phi^1 | X^{(1,obs)}_{(k-1)}, \ldots, X^{(d)}_{(k-1)}, Z, O)\\
&X^{(1,miss)}_{(k)} \sim f(X^{(1)} | X^{(2)}_{(k-1)}, \ldots, X^{(d)}_{(k-1)}, Z, O; \phi^1_{(k)})\\
&\phi^2_{(k)} \sim f(\phi^2 | X^{(1)}_{(k)}, X^{(2,obs)}_{(k-1)}, \ldots, X^{(d)}_{(k-1)}, Z, O)\\
&X^{(2,miss)}_{(k)} \sim f(X^{(2)} | X^{(1)}_{(k)}, X^{(3)}_{(k-1)}, \ldots, X^{(d)}_{(k-1)}, Z, O; \phi^2_{(k)})\\
&\ldots\\
&\phi^d_{(k)} \sim f(\phi^d | X^{(1)}_{(k)}, \ldots, X^{(d-1)}_{(k)}, X^{(d,obs)}_{(k-1)}, Z, O)\\
&X^{(d,miss)}_{(k)} \sim f(X^{(d)} | X^{(1)}_{(k)}, \ldots, X^{(d-1)}_{(k)}, Z, O; \phi^d_{(k)})
\end{aligned}$$

We iterate until convergence. When we have missingness in only one variable, no iteration is required, and the algorithm reduces to standard parametric multiple imputation.

In our cure setting, we want to use the conditional distribution $f(X^{(p)}|X^{(-p)}, Y, \delta, G; \phi^p)$ to impute each partially observed covariate $X^{(p)}$. In practice, however, $f(X^{(p)}|X^{(-p)}, Y, \delta, G; \phi^p)$ may be difficult to use for imputation, and we may use an approximation, $\tilde{f}(X^{(p)}|X^{(-p)}, Y, \delta, G; \tilde{\phi}^p)$. If the distribution used for imputation explicitly depends on G, we treat G as another partially observed variable and impute G as part of the chained equations algorithm. If we also impute the true event time T for censored subjects, we could impute partially observed $X^{(p)}$ using $f(X^{(p)}|X^{(-p)}, T, G; \phi^p)$ or a corresponding approximation. We will assume that the covariates are missing at random (MAR).

For many of the imputation approaches we consider, drawing $\tilde{\phi}^p$ and missing $X^{(p)}$ values (assuming a flat prior for ϕ^p) will reduce to fitting a regression model for $X^{(p)}$ using some function of $X^{(-p)}, G, Y, \delta$, and maybe T as predictors. As in standard FCS, we fit this regression model only for subjects with observed $X^{(p)}$. We then draw the parameter $\tilde{\phi}^p$ from a multivariate normal with mean and variance obtained using the regression model fit and then use the drawn $\tilde{\phi}^p$ and the conditional distribution implied by the regression model to draw each missing value of $X^{(p)}$. We will call this regression model the imputation model for $X^{(p)}$. Alternatively, we can obtain a draw of $\tilde{\phi}^p$ by fitting the imputation model to a bootstrap sample of the data (Little and Rubin, 2002). Multiple imputation using standard regression models can be implemented using the package MICE in R (Van Buuren and Groothuis-Oudshoorn, 2011). For imputing covariates assumed to be normally distributed, we use predictive mean matching as implemented in MICE.

The chained equations (FCS) algorithm will result in a single imputed dataset. We repeat the algorithm to create several imputed datasets. Suppose our goal is to make inference from a particular model fit (in our case, the CPH cure model). We fit this model to each imputed dataset, and then we use Rubin's Rules to produce a final estimate of the parameters and their variances from which we can make the desired inference (Rubin, 1987).

2.3.2 Imputation using the Exact Conditional Distribution

We can use the complete data likelihood from the CPH cure model and an assumption about the distribution of $X^{(p)}|X^{(-p)}$ to derive the imputation $X^{(p)}|X^{(-p)}, \delta, G,$ and Y for each partially observed $X^{(p)}$ up to proportionality. Let $f(X_i^{(-p)};\gamma)$ be the joint distribution of $X_i^{(-p)}$. In practice, we will not need to explicitly specify this distribution. Let $f(X_i^{(p)}|X_i^{(-p)};\theta)$ be the distribution of $X^{(p)}$ given all the other covariates. We assume that censoring does not depend on $X^{(p)}$ but may depend on other covariates. Therefore, we do not need to specify a model for the censoring mechanism to derive the conditional distribution of $X^{(p)}$. We consider the complete data likelihood (assuming cure status is known) for the CPH cure model:

$$L(\alpha,\alpha_0,\beta,\theta,\theta_0,\gamma,\sigma^2) = \prod_{i=1}^{n}\left[h(Y_i|G_i=1,X_i;\beta)^{\delta_i}S(Y_i|G_i=1,X_i;\beta)\right]^{G_I}$$

$$\times\left[P(G_i=1|X_i;\alpha,\alpha_0)\right]^{G_i}\left[P(G_i=0|X_i;\alpha,\alpha_0)\right]^{1-G_i} f(X_i^{(p)}|X_i^{(-p)};\theta)f(X_i^{(-p)};\gamma)$$

$$\propto \prod_{i=1}^{n}\left\{\left(h_0(Y_i)e^{\beta^T X_i}\right)^{\delta_i} e^{-H_0(Y_i)e^{\beta^T X_i}}\frac{e^{\alpha^T X_i+\alpha_0}}{1+e^{\alpha^T X_i+\alpha_0}}\right\}^{G_i}\left\{\frac{1}{1+e^{\alpha^T X_i+\alpha_0}}\right\}^{1-G_i} f(X_i^{(p)}|X_i^{(-p)};\theta)$$

Using that the conditional distribution is proportional to L (with respect to $X^{(p)}$)

$$f(X_i^{(p)}|G_i,\delta_i,Y_i,X_i^{(-p)}) \propto \left\{e^{\delta_i\beta^T X_i}e^{-H_0(Y_i)e^{\beta^T X_i}}\frac{e^{\alpha^T X_i+\alpha_0}}{1+e^{\alpha^T X_i+\alpha_0}}\right\}^{G_i}$$

$$\times\left\{\frac{1}{1+e^{\alpha^T X_i+\alpha_0}}\right\}^{1-G_i} f(X_i^{(p)}|X_i^{(-p)};\theta) \qquad (2.2)$$

We can use this kernel (distribution known up to proportionality) to impute $X_i^{(p)}$ within the chained equations imputation procedure. This kernel depends on both G_i and $H_0(t)$, and it is parameterized by $\phi^p = (\alpha, \alpha_0, \beta, \theta)$. When $X_i^{(p)}$ is assumed to be normal, we can draw from (2.2) using an accept-reject algorithm as described below. When $X_i^{(p)}$ is categorical, the full form of the imputation distribution can be easily derived using (2.2).

In order to impute partially observed covariates using (2.2), we treat G as another partially observed variable and impute G within the chained equations algorithm. We also append a step at the start of each chained equations iteration in which we estimate $H_0(t)$. We can impute by iterating the following steps:

Step 1: Estimating $H_0(t)$

We can estimate $H_0(t)$ several different ways. Firstly, we can estimate $H_0(t)$ using a weighted Breslow estimator (Breslow, 1972). Suppose we have event times $t_1, ..., t_J$ and let R_j be the risk set at time t_j. Using the imputed X from the most recent iteration, we estimate $H_0(t)$ at the k^{th} iteration of the imputation algorithm as the step function

$$\hat{H}_0^{(k)}(t) = \sum_{t_j \leq t}^{J} \frac{\# \text{ events at time } t_j}{\sum_{i \in R_j} e^{\left[\beta^{(k-1)}\right]^T X_i} w_i^{(k)}}$$

where $w_i^{(k)}$ is the conditional probability that a person is not cured at iteration k as expressed in equation (2.1) evaluated at $\beta^{(k-1)}$, the draw of β from the previous iteration (Sy and Taylor, 2000). We use this approach to estimate $H_0(t)$ in our simulations.

We can also obtain a parametric estimate of $H_0(t)$ by fitting a CPH cure model with a parametric baseline hazard such as Weibull. If the baseline hazard of an event in the non-cured subjects is truly Weibull, then fitting a Weibull cure model rather than a semi-parametric CPH cure model may produce extra efficiency in estimating β. However, if the baseline hazard in the non-cured group is not believed to be Weibull, using this approach is not advised. Alternatively, $H_0(t)$ can be estimated using only the subset of the data such that $G_i = 1$ (non-cured) as imputed at iteration $k - 1$. This can be estimated by fitting a Cox model and using a traditional Breslow estimator applied to the $G_i = 1$ subset of the data or by assuming a parametric form for the event hazard in the $G_i = 1$ group.

Step 2: Imputing Cure Status

To produce proper imputations using the FCS algorithm, we first draw the parameters from their posterior distributions. Assuming flat priors, this can be done (approximately) by either 1) fitting a cure model to a bootstrap sample of the data or by 2) fitting a logistic model for $G|X$ and a CPH regression model of $(Y, \delta)|X$ on the $G = 1$ subset using bootstrap samples of the most recent imputed data (Little and Rubin, 2002).

Using the complete data likelihood for the CPH cure model, we can show that $\text{logit}(P(G_i = 1|X_i, \delta_i = 0, Y_i)) = -\hat{H}_0(Y_i)e^{\beta^T X_i} + \alpha^T X_i + \alpha_0$. We can draw imputed values of G_i using this probability relation. We note that if $\delta_i = 1$, then G_i is known to be 1, so we will not need to impute. Also, we define censored individuals with late censoring times (after some cut-point c) as cured. Therefore, G is treated as missing only if $\delta = 0$ and $Y \leq c$, so we can view missingness in G as MAR conditional on δ and Y.

Step 3: Imputing the Missing Covariates

We specify the distribution $f(X_i^{(p)}|G_i, \delta_i, Y_i, X_i^{(-p)}; \phi^p)$ for each covariate $X^{(p)}$ with missing values. As described in **Section 2.3.1**, we 1) draw ϕ^p from its posterior distribution and 2) impute missing values of $X^{(p)}$ from its conditional distribution using the drawn ϕ^p. If only one covariate has missingness, we perform 1) and 2) a single time for that covariate. If we have missingness in many covariates, we perform 1) and 2) sequentially for each covariate with missingness using the most recent imputations of the other variables.

Suppose first that $X^{(p)}$ is Bernoulli such that $f(X^{(p)}|X^{(-p)}; \theta)$ is a logistic regression model with $X^{(p)}$ as the outcome and $X^{(-p)}$ as covariates. We can impute missing values of $X_i^{(p)}$ from a Bernoulli(π_i) distribution using $\pi_i = P(X^{(p)} = 1|X^{(-p)}; \theta)$ obtained from (2.2).

Suppose instead that $X_i^{(p)} \sim N(\theta_0 + \theta_1 X^{(-p)}, \sigma^2)$. In this case, the form of the full conditional distribution implied by (2.2) is known only up to proportionality. We can draw $(\theta_0, \theta, \sigma^2)$ under the Bayesian linear regression model with $X^{(p)}$ as the outcome and with $X^{(-p)}$ as the predictors using the most recent imputed values. This model is described by Rubin (1987) and used in MICE (Van Buuren and Groothuis-Oudshoorn, 2011). We then want to impute each missing value $X_i^{(p)}$ by taking draws from the full conditional distribution knowing only the kernel in (2.2). Many methods exist to draw from a distribution using only the kernel. To obtain an imputed value for $X_i^{(p)}$ at a given iteration, we perform a Metropolis-Hastings draw from (2.2) using a normal random walk proposal distribution centered at the imputed value from the previous iteration (Hastings, 1970; Metropolis et al., 1953). The variance of this proposal distribution is a tuning parameter that must be determined to ensure good mixing properties and a reasonable acceptance rate (Sherlock et al., 2010). Due to this accept-reject sampling, we may need to perform many iterations of the chained equations fitting algorithm to reach convergence. Rejection sampling methods can also be used (Bartlett et al., 2014).

This "Exact Cure" approach imputes each partially observed $X^{(p)}$ using its conditional distribution implied by the CPH cure model and the model for $X^{(p)}|X^{(-p)}$. However, for some specifications of $f(X^{(p)}|X^{(-p)};\theta)$, we must use an accept-reject algorithm to impute each missing $X_i^{(p)}$ using (2.2), and this can quickly result in a large computational burden. This burden is amplified when we have missingness in multiple covariates. To impute multiple partially observed covariates, we must specify the model for $X^{(p)}|X^{(-p)}$ for each partially observed $X^{(p)}$, which increases the number of parameters that must be drawn. Additionally, we must derive the form of $f(X_i^{(p)}|G_i, \delta_i, Y_i, X_i^{(-p)})$ separately for different forms of the model for each $X^{(p)}|X^{(-p)}$ (e.g. Gamma, Poisson, etc). Due to this, we do not apply the Exact Cure approach to the head and neck cancer example later on, which has missingness in many variables.

2.3.3 Approximations using Regression Models

In this section, we consider approximations to the “exact” conditional distributions (derived in **Section 2.3.2**) that do not require accept-reject sampling and can more easily by implemented with existing software. We are interested in approximations that correspond to standard regression models.

We start by describing two simple covariate imputation approaches for survival data without a cure fraction. We then describe an approach in the literature for imputing survival data without a cure fraction that is motivated directly by the *standard* Cox proportional hazards model. Then, we propose an approximate distribution that incorporates the cure structure of the data and is motivated by the CPH cure model formulation. Finally, we consider a modification to these approaches in which event time T is imputed for censored subjects.

logY Imputation for survival data without a cure fraction

One approach in the literature for imputing covariates for survival data without a cure fraction is to use $X^{(-p)}$, δ, and $\log(Y)$ as predictors in the imputation model for $X^{(p)}$ used in the chained equations algorithm (Van Buuren et al., 1999). Unlike the method in **Section 2.3.2**, this approach does not require us to impute cure status or estimate $H_0(t)$, so we do not require iteration of the chained equations algorithm when we have missingness in only one covariate. We can impute using MICE in R by specifying regression models with predictors $X^{(-p)}$, δ, and $\log(Y)$ for imputing each partially-observed $X^{(p)}$ (Van Buuren and Groothuis-Oudshoorn, 2011).

Outcome Binning Imputation for survival data without a cure fraction

One adaptation of existing approaches for imputing covariates in the non-cure setting would be to use a regression model for imputing each partially observed $X^{(p)}$ with predictors $X^{(-p)}, \delta f_1(Y),$ and $(1-\delta)f_2(Y)$ where $f_1(Y)$ and $f_2(Y)$ are some functions of Y. We propose using f_1 and f_2 in the form of step functions with step height determined by the data. This allows for a very flexible association between the outcome and the partially observed covariate. Additionally, this approach does not require us to impute cure status or estimate $H_0(t)$ explicitly.

We call this approach "Outcome Binning" because it involves binning individuals based on the composite outcome, (Y, δ). We first separate subjects into a $\delta = 1$ and $\delta = 0$ group. We then define bins of Y within each δ group using summary statistic-based cutoffs or by other methods. For convenience, we define the bins using quartiles of Y within each of the $\delta_i = 1$ and $\delta_i = 0$ groups. We define a set of dummy indicator variables, $M_1, \ldots, M_m$, which identify the bin membership of each individual ($M_k = 1$ if the subject is in bin k). We then impute each partially observed covariate within the chained equations procedure using a regression model for each $X^{(p)}$ with $X^{(-p)}$ and binary indicators $M_2, \ldots, M_m$ as predictors. After determining $M_1, \ldots, M_m$, we can perform the chained equations imputation using MICE in R (Van Buuren and Groothuis-Oudshoorn, 2011). With missingness in only one covariate, we can perform a single iteration of the chained equations algorithm.

White and Royston Imputation for the CPH model without a cure fraction

Based on algebraic derivation involving Taylor series approximations, White and Royston (2009) suggests using $X^{(-p)}, \delta,$ and $H_0(Y)$ as predictors in the imputation model for each partially observed $X^{(p)}$ in the standard CPH model setting without a cure fraction. This is quite similar to the method in Van Buuren et al. (1999) but replacing $\log(Y)$ with $H_0(Y)$. This requires us to obtain an estimate of $H_0(t)$ but does not require us to impute cure status.

We note that $H_0(t)$ is the cumulative baseline hazard of an event in the entire study population. This is not the same as the cumulative baseline hazard in the non-cured population, as the cured subjects cannot experience the event of interest. When applied to survival data with a cure fraction, $H_0(t)$ is the cumulative baseline hazard of an event in the (assumed to be misspecified) survival model without a cure fraction based on the entire study population.

White and Royston (2009) ultimately recommends using the Nelson-Aalen estimator of $H(t)$ to estimate $H_0(t)$ before imputation. However, they also investigated an approach in which they add a step to the imputation algorithm and re-estimate $H_0(t)$ at each iteration. We estimate $H_0(t)$ after each iteration of the chained equations algorithm by fitting a Cox model to all subjects using the most recent imputed data, drawing the Cox model parameter using a multivariate normal distribution with mean and covariance matrix from the Cox model fit, and then using a Breslow estimator. We can also draw parameter values by fitting the models to a bootstrap sample of the data (Little and Rubin, 2002). Alternatively, we can fit a Weibull regression model to all subjects and estimate the cumulative baseline hazard in the total population as a parametric function.

As we estimate $H_0(t)$ at the end of each iteration, we iterate the chained equations algorithm even when we only have missingness in a single covariate. We can impute using MICE in R by iterating the following steps: 1) Estimate $H_0(t)$ 2) Impute each partially observed covariate $X^{(p)}$ sequentially using an appropriate elementary imputation method in MICE (e.g. mice.impute.logreg() for binary covariates) with predictors $X^{(-p)}, \delta,$ and $\hat{H}_0(Y)$ (Van Buuren and Groothuis-Oudshoorn, 2011).

Approximated Imputation for the CPH cure model

We use a similar approach to White and Royston (2009) to derive approximate imputation models for normal and binary covariates in the CPH cure model setting. We will call this imputation approach the "Approximate Cure" approach. Although not shown here, we can derive approximate imputation models for covariates with other distributions in a similar fashion. Suppose we have the same set of covariates in both parts of the mixture cure model and that the set contains s covariates. Therefore, α and β both have dimension s. Again, we suppose that a partially observed $X^{(p)} \sim N(\theta_1^T X^{(-p)} + \theta_0, \sigma^2)$. Taking the logarithm of kernel (2.2), we have that

$$\begin{aligned}\log\left(f(X_i^{(p)}|G_i, \delta_i, Y_i, X_i^{(-p)})\right) = \frac{-1}{2\sigma^2}\left(X_i^{(p)} - \theta_1^T X_i^{(-p)} - \theta_0\right)^2 + G_i\alpha^T X_i \\ - \log\left(1 + e^{\alpha^T X_i + \alpha_0}\right) + G_i\delta_i\beta^T X_i \\ - G_i H_0(Y_i) e^{\beta^T X_i} + \text{constant}\end{aligned}$$

We treat terms that do not depend on $X_i^{(p)}$ as constant. We note that log(1+z) $\approx$ log(1+c) + (z-c)/(1+c) if z is near c and $e^{aX+bY} \approx e^{a\bar{X}+b\bar{Y}}\left[1 + a(X - \bar{X}) + b(Y - \bar{Y})\right]$ if Var(aX + bY) is small. Assuming $\text{Var}(\alpha^T X_i)$ and $\text{Var}(\beta^T X_i)$ are small and using first and zeroth order Taylor series approximations, we can approximate the above by:

$$\log\left(f(X_i^{(p)}|G_i, \delta_i, Y_i, X_i^{(-p)})\right) \approx \frac{-1}{2\sigma^2}\left(X_i^{(p)} - \theta_1^T X_i^{(-p)} - \theta_0\right)^2 + G_i\alpha^T X_i$$

$$
\begin{aligned}
&- \frac{e^{\alpha^T \bar{X}+\alpha_0}}{1+e^{\alpha^T \bar{X}+\alpha_0}} \left[1+\alpha_p(X_i^{(p)}-\bar{X}^{(p)})+\sum_{j\neq p}^{s}\alpha_j(X_i^{(j)}-\bar{X}^{(j)})\right]+G_i\delta_i\beta^T X_i \\
&- G_i H_0(Y_i)e^{\beta^T\bar{X}}\left[1+\beta_p(X_i^{(p)}-\bar{X}^{(p)})+\sum_{j\neq p}^{s}\beta_j(X_i^{(j)}-\bar{X}^{(j)})\right]+\text{constant} \\
&= \frac{-1}{2\sigma^2}\left(X_i^{(p)}-\theta_1^T X_i^{(-p)}-\theta_0\right)^2+\left[G_i\alpha_p-\frac{e^{\alpha^T\bar{X}+\alpha_0}}{1+e^{\alpha^T\bar{X}+\alpha_0}}\alpha_p\right. \qquad (2.3) \\
&\left. +G_i\delta_i\beta_p-G_iH_0(Y_i)e^{\beta^T\bar{X}}\beta_p\right]X_i^{(p)}+\text{constant}
\end{aligned}
$$

where $\bar{X}^{(p)}$ is treated as a constant because it only very weakly depends on $X_i^{(p)}$. If we complete the square on (2.3), we see that the mean of this normal distribution will be a linear combination of $X_i^{(-p)}$, G_i, $G_i \times \delta_i$ and $G_i \times H_0(Y_i)$. A second order Taylor series approximation of $e^{\alpha^T X_i}$ and $e^{\beta^T X_i}$ will also give the interaction $G_i \times H_0(Y_i) \times X_i^{(-p)}$. This suggests that when $X^{(p)}$ is normal and the assumptions are satisfied, we can approximate the exact distribution $f(X_i^{(p)}|G_i, \delta_i, Y_i, X_i^{(-p)})$ using a linear regression model with $X_i^{(-p)}$, G_i, $G_i \times \delta_i$, $G_i \times H_0(Y_i)$, and perhaps $G_i \times H_0(Y_i) \times X_i^{(-p)}$ as predictors.

Suppose instead that $X^{(p)} \sim$ Bernoulli(t) where t = expit$(\theta_1^T X^{(-p)} + \theta_0)$. Using the complete data likelihood for the CPH cure model, we have

$$
\begin{aligned}
&\text{logit}\left(P(X_i^{(p)}=1|G_i,\delta_i,Y_i,X_i^{(-p)})\right)=\log\left(\frac{L(\alpha,\alpha_0,\beta,\theta_1,\theta_0,\gamma)|_{X_i^{(p)}=1}}{L(\alpha,\alpha_0,\beta,\theta_1,\theta_0,\gamma)|_{X_i^{(p)}=0}}\right) \\
&=\log\left(\left\{\left(e^{\beta_p}\right)^{\delta_i}e^{-H_0(Y_i)e^{\beta_p+\sum_{j\neq p}^{s}X_i^{(j)}\beta_j}}\frac{e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j+\alpha_p}}{1+e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j+\alpha_p}}\right\}^{G_i}\right. \\
&\left.\times\left\{\frac{1}{1+e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j+\alpha_p}}\right\}^{1-G_i}e^{\theta_1^T X^{(-p)}+\theta_0}\right) \\
&-\log\left(\left\{e^{-H_0(Y_i)e^{\sum_{j\neq p}^{s}X_i^{(j)}\beta_j}}\frac{e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j}}{1+e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j}}\right\}^{G_i}\left\{\frac{1}{1+e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j}}\right\}^{1-G_i}\right) \\
&=\theta_0+\theta_1^T X_i^{(-p)}+G_i\delta_i\beta_p-G_iH_0(Y_i)\left(e^{\beta_p}-1\right)e^{\sum_{j\neq p}^{s}X_i^{(j)}\beta_j}+\alpha_pG_i \\
&+\log\left(1+e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j}\right)-\log\left(1+e^{\alpha_0+\sum_{j\neq p}^{s}X_i^{(j)}\alpha_j+\alpha_p}\right) \qquad (2.4)
\end{aligned}
$$

This relation gives the form for the exact conditional distribution, which we can use to impute a partially observed, binary $X^{(p)}$. Now, we attempt to find a simpler approximated model. We use a similar approach as in the normal derivation. Assuming Var$(\alpha^T X_i)$ and

Var$(\beta^T X_i)$ are small, we approximate (2.4) by:

$$\begin{aligned}
&\text{logit}(P(X_i^{(p)} = 1|G_i, \delta_i, Y_i, X_i^{(-p)})) \approx \theta_0 + \theta_1^T X_i^{(-p)} + G_i\delta_i\beta_p + \alpha_p G_i + \text{constant} \\
&\quad - G_i H_0(Y_i)\left(e^{\beta_p} - 1\right) e^{\sum_{j\neq p}^{s} X_i^{(j)}\beta_j} + \frac{e^{\alpha_0 + \sum_{j\neq p}^{s} X_i^{(j)}\alpha_j}}{1 + e^{\alpha_0 + \sum_{j\neq p}^{s} \bar{X}^{(j)}\alpha_j}} - \frac{e^{\alpha_0 + \sum_{j\neq p}^{s} X_i^{(j)}\alpha_j + \alpha_p}}{1 + e^{\alpha_0 + \alpha_p + \sum_{j\neq p}^{s} \bar{X}^{(j)}\alpha_j}} \\
&\quad \approx \theta_0 + \theta_1^T X_i^{(-p)} + G_i\delta_i\beta_p - G_i H_0(Y_i)\left(e^{\beta_p} - 1\right) e^{\sum_{j\neq p}^{s} \bar{X}^{(j)}\beta_j}\left[1 + \sum_{j\neq p}^{s} \beta_j(X_i^{(j)} - \bar{X}^{(j)})\right] \\
&\quad + \alpha_p G_i + \frac{e^{\alpha_0 + \sum_{j\neq p}^{s} \bar{X}^{(j)}\alpha_j}}{1 + e^{\alpha_0 + \sum_{j\neq p}^{s} \bar{X}^{(j)}\alpha_j}}\left[1 + \sum_{j\neq p}^{s} \alpha_j(X_i^{(j)} - \bar{X}^{(j)})\right] \\
&\quad - \frac{e^{\alpha_0 + \sum_{j\neq p}^{s} \bar{X}^{(j)}\alpha_j + \alpha_p}}{1 + e^{\alpha_0 + \alpha_p + \sum_{j\neq p}^{s} \bar{X}^{(j)}\alpha_j}}\left[1 + \sum_{j\neq p}^{s} \alpha_j(X_i^{(j)} - \bar{X}^{(j)})\right] + \text{constant} \qquad (2.5)
\end{aligned}$$

This equation is a linear combination of $X_i^{(-p)}$, G_i, $G_i \times \delta_i$, $G_i \times H_0(Y_i)$, and $G_i \times H_0(Y_i) \times X^{(-p)}$. This suggests that we can impute $X_i^{(p)}$ using $X_i^{(-p)}$, G_i, $G_i \times \delta_i$, $G_i \times \hat{H}_0(Y_i)$, $G_i \times \hat{H}_0(Y_i) \times X^{(-p)}$ as predictors in a logistic regression model if we impute G_i for censored subjects and estimate $H_0(Y_i)$ as additional steps in the multiple imputation algorithm.

The approximate imputation models implied by (2.3) and (2.5) explicitly depend on $H_0(t)$ and G_i. To use the derived approximate distributions for covariate imputation, we estimate $H_0(t)$ and impute G_i as part of the chained equations algorithm as we did in **Section 2.3.2**. In contrast, the logY, Outcome Binning, and White and Royston imputation approaches discussed previously do not require us to impute G_i.

The final interaction term in the imputation models implied by (2.3) and (2.5) may have many parameters if X_i consists of many covariates, so that term may have to be dropped for settings with many covariates. Also, it may be that the imputed G_i and $G_i \times \delta_i$ are highly correlated, so one may need to only use G_i due to collinearity issues.

In order to impute partially observed covariates using these approximations, we can perform a modification of the Exact Cure algorithm proposed in **Section 2.3.2**. We can impute using MICE in R by iterating the following steps: Step 1) Estimate $H_0(t)$ as in **Section 2.3.2**, Step 2) Impute Cure Status as in **Section 2.3.2**, and Step 3) Impute each partially observed covariate $X^{(p)}$ sequentially using an appropriate elementary imputation method in MICE (e.g. mice.impute.logreg() for binary covariates) with predictors $X_i^{(-p)}$, G_i, $G_i \times \delta_i$, $G_i \times \hat{H}_0(Y_i)$, and perhaps $G_i \times \hat{H}_0(Y_i) \times X_i^{(-p)}$ (Van Buuren and Groothuis-

Oudshoorn, 2011).

A natural alternative to the proposed Approximate Cure approach is to first impute G and then impute covariates separately for the $G = 1$ and $G = 0$ groups. We could then apply imputation approaches for survival data without a cure fraction (such as the White and Royston method) for imputing covariates in the $G = 1$ group. In simulations (not shown), this approach resulted in similar bias and inflated variances compared to the Approximate Cure approach.

A Modification: Event Time Imputation

Since the observed event/censoring time $Y = \min(T, C)$ is a mixture of two underlying random variables, it may not be very intuitive to include Y as a predictor in standard regression models for imputing missing covariates. Instead, we may wish to include the true event time, T, which is not fully observed. We can treat T as another partially observed variable and impute values of T for censored individuals within the chained equations algorithm used to impute missing covariates. This modification can conceptually be applied to any of the imputation approaches we have discussed.

In the cure setting, T is defined as infinity for cured individuals and is an event time for non-cured individuals. Although cure status is not known for censored individuals, if we also impute G as part of the chained equations imputation algorithm, then we can impute values of T for the non-cured, censored subjects using an assumed truncated distribution $f(t|t > C, G = 1, X)$. We can modify the Exact and Approximate Cure imputation algorithms by adding a step to the chained equations imputation algorithm to impute T_i for censored individuals who have $G_i = 1$ at iteration k. Then, we replace (Y_i, δ_i) in the subsequent imputation models for the partially observed covariates with the imputed (T_i, G_i). In several simulations (not shown), however, T imputation does not appear to improve the performance of the Exact Cure and Approximate Cure imputation algorithms.

We are particularly interested to see how some simple covariate imputation approaches for survival data without a cure fraction are impacted by first imputing T and then substituting (Y, δ) by $(T, 1)$ in the covariate imputation models. We consider both the logY and Outcome Binning approaches. For the Outcome Binning approach, we use octiles to define bins of T among all subjects. In these two approaches, cure status is not known or imputed for censored individuals, and so we cannot impute censored T using the truncated distribution $f(t|t > C, X, G = 1)$. Instead, we impute the event time T using the truncated distribution $f(t|t > C, X)$, which we assume has a proportional hazards structure with a Weibull baseline.

We use a Cox proportional hazards model for the hazard of an event in the total study population. The survival function of the truncated distribution $f(t|t > C_i, X_i)$ of T_i is in the form $S_{TRUNC}(t|X_i) = e^{-[H_0(t)-H_0(C_i)]e^{\beta^T X_i}}, t > C_i$. To impute T_i for a censored individual, we can first generate U_i from a Uniform(0,1) distribution. We can then draw T_i using the relation $T_i = H_0^{-1}\left(-\log(U_i)e^{-\beta^T X_i} + H_0(C_i)\right)$. This requires us to draw β and estimate $H_0(t)$. If we assume the failure time is Weibull such that $S(t|X_i) = e^{-\lambda t^\eta e^{\beta^T X_i}}$, then we can generate T_i as $T_i = \left(\frac{-\log(U_i)e^{-\beta^T X_i}+\lambda C_i^\eta}{\lambda}\right)^{1/\eta}$ after drawing values for β, λ, and η. Within the chained equations algorithm, we generate a T_i value for all censored subjects at each iteration. We can obtain draws of β, λ, and η by first fitting a Weibull regression model to the entire study population using the most recent imputed X and then drawing β, λ, and η from a multivariate normal distribution with mean and covariance estimated by the Weibull fit.

We note that in the CPH cure model setting, the truncated distribution $f(t|t > C, X)$ is incorrectly specified, and it may seem unintuitive to use this misspecified model to impute event times. However, event time imputation has been used in the non-cure survival setting, and an analyst might naively try to apply the same approach to survival data with a cure fraction (Taylor et al., 2002). We want to see whether this approach improves or worsens the performance of imputation approaches for survival data without a cured fraction when applied in the cure setting.

2.4 Simulations

In this section, we present results from a simulation study to compare the imputation approaches in terms of bias, empirical variance, and coverage of cure model parameters across imputation methods. We also compare with complete case analysis and analysis of the full data without any covariate missingness.

2.4.1 Simulation 1: Missingness in a Single Covariate

We create 500 simulated datasets of 500 observations each. For each dataset, we simulate multivariate normal covariates $X = (X_1, X_2)$ with zero means, unit variances, and a correlation of 0.5. We then simulate cure status using the relation $\text{logit}\left(P(G_i = 1|X_{i,1}, X_{i,2})\right) = 0.5 + 0.5X_{i,1} + 0.5X_{i,2}$, leading to an average cure rate of 40%. For the non-cured group, we simulate a survival time T_i. We model the event hazard in the non-cured group as $h(t) = h_0(t)e^{0.5X_1+0.5X_2}$ with $h_0(t) = 0.002$. We then generate censoring times $C_i \sim U(250, 4500)$ and define $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$.

We impose ~50-55% missingness in X_2 using three models: **(1)** missing completely at random (MCAR) with $P(X_2 \text{ missing}|X_1, \delta, Y) = 0.5$, **(2)** missing at random (MAR) with $\text{logit}(P(X_2 \text{ missing}|X_1, \delta, Y)) = X_1$, and **(3)** MAR with $\text{logit}(P(X_2 \text{ missing}|X_1, \delta, Y)) = 0.3 - 0.4\delta - 0.5X_1\delta$. While this final missingness mechanism may seem implausible, it could be induced when missingness depends on an unobserved variable U that is independently related to T.

We note that we impose missingness in only a single covariate rather than many covariates (the typical setting where FCS is applied). However, we are mainly interested in investigating various strategies for modeling the univariate conditional distribution for one partially observed covariate. As such, we can compare the imputation approaches by imposing missingness in only one covariate. Similar results can be seen when we apply the imputation approaches with missingness in multiple covariates as shown in Simulation 2. We also consider the setting with many partially observed covariates in our head and neck cancer example.

We perform multiple imputation of X_2 using methods described in this chapter. For each simulation and method, we produce 10 imputed datasets. We then fit a CPH cure model to each imputed dataset (ignoring imputed cure status) and use Rubin's Rules to obtain a single set of estimates for each simulation (Rubin, 1987). We then compute bias, relative variance (compared to analyzing the full data with no covariate missingness), and coverage in estimating model parameters across 500 simulations for each method. Alternatively, for imputation approaches that result in imputed values for G, we could have performed our final analysis by fitting Cox and logistic regressions given the imputed G. In simulations (not shown), this approach resulted in a slight increase in efficiency for estimating the intercept for the logistic part of the model, but it also resulted in some increases in bias for the approaches using approximated distributions for imputation.

We use 100 iterations for each imputation algorithm except Exact Cure, for which we use 1500 due to the slower convergence of the Metropolis-Hastings algorithms. When fitting the cure models to each imputed dataset, we use 100 iterations of the EM algorithm and use 100 bootstrap samples of the imputed dataset to estimate variances.

Computational time is shortest for the Outcome Bins and logY approaches, followed closely by the T imputation methods. The Approximate Cure approach takes about four times as long as the Outcome Bins method to run and about two times as long as the White and Royston method. The Exact Cure approach takes at least ten times as long as the Approximate Cure approach to run.

Table 2.1 shows simulation results under three different missingness mechanisms for X_2. Under missingness models (1) and (2), complete case (CC) analysis is essentially unbiased or has little bias. However, in model (3), CC analysis results in biased estimates, particularly in estimating parameters for the logistic part of the mixture cure model. In all missingness settings shown, the imputation methods have little bias in estimating α_0, α_1, and β_1, the logistic model intercept and the parameters associated with X_1.

In all three missingness settings, the logY, White & Royston, Outcome Binning, T imputation, and Approximate Cure (w/o extra interaction) approaches result in similar or larger bias than CC analysis in estimating α_2, the logistic parameter for X_2. For all three missingness models, the imputation approaches using T imputation result in larger α_2 bias than their counterparts without T imputation. The Approximate Cure approach with the interaction term and the Exact Cure approach produce comparably low bias in estimating α_2.

All imputation methods except the Exact Cure approach result in biased estimates for β_2, the failure time model parameter associated with X_2. Among the biased imputation methods, however, the Approximate Cure approach including the extra interaction term consistently results in the smallest β_2 bias. The logT approach produces smaller β_2 bias than the logY approach. Outcome Binning results in similar β_2 bias with and without the T imputation.

All imputation methods result in smaller empirical variance (so larger relative variance) in estimating α_0, α_1, and β_1 compared to CC analysis in all three simulation settings. Some reduction in the variance in estimating β_2 can also be seen, suggesting that we can still gain some information about the effect of X_2 by including information from subjects with missing X_2. Coverage rates for α_0, α_1, and β_1 are similar for all imputation methods in all three simulation settings. CC coverage of 95% confidence intervals for α_0 and α_1 under missingness model (3) is far below 0.95%. Reductions in coverage for some imputation approaches can be seen for α_2 and β_2. Undercoverage is mainly due to increased bias. The Exact Cure approach and the Approximate Cure approach with the extra interaction term tend to produce higher coverage rates in estimating β_2 compared to the other imputation methods.

In all three sets of simulations, we see large reductions in the Approximate Cure approach's corresponding biases by adding the extra interaction term. Although not shown, we do not see corresponding decreases in bias by adding a $\hat{H}_0(Y_i) : X^{(-p)}$ interaction term to the White and Royston approach (White and Royston, 2009). We also see that the Exact Cure imputation approach far outperforms all other imputation algorithms in terms of bias, and among the biased imputation approaches, the Approximate Cure approach with the interaction term is generally the best performer. In all three sets of simulations, the non-cure imputation approaches that involve T imputation tend to have worse coverage or bias properties than the corresponding approaches without T imputation. Finally, we see that among the approaches that do not take the cure fraction into account (Outcome Binning, logY, White & Royston, and logT), Outcome Binning without T imputation tends to produce the smallest bias overall across the three simulation settings.

Table 2.1: Cure Model Estimates with Imputation of One Missing Covariate

	α_0	α_1	α_2	β_1	β_2
Method	Bias (RV) CI†	Bias (RV) CI	Bias (RV) CI	Bias (RV) CI	Bias (RV) CI
Full Data	-0.01 (1.00) 0.93	0.02 (1.00) 0.93	0.02 (1.00) 0.94	-0.01 (1.00) 0.95	0.00 (1.00) 0.95
Missingness Model 1: MCAR missingness in X_2					
Exact Cure	0.00 (0.83) 0.94	0.01 (0.75) 0.92	0.03 (0.48) 0.94	-0.01 (0.82) 0.94	0.00 (0.48) 0.95
Approximations					
Non-Cure w/ (Y, δ)					
logY	0.00 (0.79) 0.94	0.00 (0.74) 0.91	0.08 (0.47) 0.92	0.01 (0.85) 0.95	-0.14 (0.71) 0.78
White & Royston	0.00 (0.81) 0.94	0.00 (0.73) 0.93	0.07 (0.47) 0.92	0.00 (0.82) 0.95	-0.13 (0.76) 0.81
Binning by (Y, δ)	0.00 (0.80) 0.94	0.01 (0.75) 0.93	0.04 (0.48) 0.93	0.00 (0.83) 0.96	-0.11 (0.66) 0.87
Non-Cure w/ T					
logT	0.00 (0.80) 0.94	-0.02 (0.79) 0.93	0.14 (0.55) 0.89	0.01 (0.94) 0.96	-0.12 (0.90) 0.86
Binning by T	0.00 (0.81) 0.94	0.00 (0.78) 0.93	0.09 (0.53) 0.92	0.00 (0.86) 0.95	-0.10 (0.71) 0.89
Cure w/ (G, Y, δ)					
Approx Cure	0.00 (0.85) 0.94	0.01 (0.75) 0.93	0.05 (0.50) 0.94	0.00 (0.81) 0.95	-0.13 (0.82) 0.82
Approx + Int*	0.00 (0.85) 0.93	0.02 (0.78) 0.92	0.02 (0.47) 0.93	0.00 (0.91) 0.95	-0.07 (0.75) 0.93
Complete Case	-0.01 (0.48) 0.94	0.03 (0.52) 0.96	0.03 (0.49) 0.94	0.00 (0.52) 0.97	0.00 (0.46) 0.95
Missingness Model 2: MAR missingness in X_2 dependent on X_1					
Exact Cure	0.00 (0.84) 0.95	0.01 (0.81) 0.94	0.04 (0.47) 0.93	0.00 (0.79) 0.95	-0.01 (0.34) 0.92
Approximations					
Non-Cure w/ (Y, δ)					
logY	0.00 (0.82) 0.94	0.01 (0.82) 0.95	0.13 (0.47) 0.91	0.02 (0.80) 0.95	-0.20 (0.63) 0.62
White & Royston	0.00 (0.79) 0.95	-0.02 (0.79) 0.95	0.14 (0.46) 0.89	0.02 (0.77) 0.96	-0.19 (0.63) 0.65
Binning by (Y, δ)	0.00 (0.83) 0.95	0.00 (0.80) 0.94	0.10 (0.51) 0.92	0.01 (0.78) 0.95	-0.16 (0.54) 0.73
Non-Cure w/ T					
logT	0.01 (0.83) 0.94	-0.01 (0.85) 0.94	0.15 (0.61) 0.88	0.02 (0.85) 0.95	-0.16 (0.71) 0.73
Binning by T	0.00 (0.84) 0.94	0.00 (0.82) 0.95	0.11 (0.58) 0.93	0.01 (0.83) 0.95	-0.15 (0.53) 0.76
Cure w/ (G, Y, δ)					
Approx Cure	0.00 (0.84) 0.94	-0.01 (0.76) 0.94	0.12 (0.49) 0.90	0.02 (0.71) 0.94	-0.20 (0.67) 0.61
Approx + Int*	0.00 (0.89) 0.95	0.01 (0.82) 0.94	0.05 (0.48) 0.94	0.00 (0.78) 0.94	-0.12 (0.65) 0.86
Complete Case	0.00 (0.41) 0.95	0.04 (0.43) 0.94	0.05 (0.50) 0.95	-0.02 (0.31) 0.95	-0.02 (0.33) 0.91
Missingness Model 3: MAR missingness in X_2 dependent on X_1, δ					
Exact Cure	0.00 (0.86) 0.94	0.01 (0.77) 0.93	0.03 (0.44) 0.94	-0.01 (0.82) 0.95	0.00 (0.60) 0.95
Approximations					
Non-Cure w/ (Y, δ)					
logY	0.00 (0.81) 0.93	0.00 (0.77) 0.94	0.07 (0.42) 0.93	0.00 (0.88) 0.95	-0.11 (0.88) 0.89
White & Royston	0.00 (0.83) 0.94	0.00 (0.79) 0.93	0.06 (0.44) 0.94	0.00 (0.87) 0.96	-0.09 (0.87) 0.90
Binning by (Y, δ)	0.00 (0.86) 0.94	0.02 (0.79) 0.94	0.02 (0.44) 0.94	0.00 (0.81) 0.96	-0.08 (0.71) 0.92
Non-Cure w/ T					
logT	0.01 (0.83) 0.94	-0.03 (0.81) 0.92	0.16 (0.52) 0.88	0.01 (0.94) 0.96	-0.09 (0.99) 0.92
Binning by T	0.00 (0.84) 0.94	0.00 (0.80) 0.94	0.09 (0.47) 0.92	0.00 (0.86) 0.96	-0.07 (0.78) 0.94
Cure w/ (G, Y, δ)					
Approx Cure	0.00 (0.87) 0.93	0.02 (0.78) 0.94	0.02 (0.44) 0.95	-0.01 (0.81) 0.96	-0.08 (0.94) 0.92
Approx + Int*	0.00 (0.88) 0.93	0.02 (0.82) 0.94	0.03 (0.44) 0.94	0.00 (0.89) 0.96	-0.05 (0.93) 0.95
Complete Case	0.18 (0.39) 0.83	0.29 (0.41) 0.77	0.03 (0.43) 0.96	0.00 (0.54) 0.95	0.00 (0.57) 0.95

*Includes $\hat{H}_0(Y) : G : X_1$ interaction in imputation model

†CI indicates empirical coverage of 95% confidence intervals and RV indicates variance relative to analysis of the full data.

2.4.2 Simulation 2: Missingness in Two Covariates

We create 500 simulated datasets of 500 observations each. For each dataset, we simulate multivariate normal covariates $X = (X_1, X_2)$ with zero means, unit variances, and a correlation of 0.5. We then simulate a third, binary covariate X_3 such that $P(X_3 = 1|X_1, X_2) = \text{expit}(X_1)$. We simulate cure status using the relation $\text{logit}\,(P(G_i = 1|X_{i,1}, X_{i,2}, X_{i,3})) = -0.5 + 0.5X_{i,1} + 0.5X_{i,2} + 0.5X_{i,3}$, leading to an average cure rate of 55%. For the non-cured group, we simulate a survival time T_i. We model the event hazard in the non-cured group as $h(t) = h_0(t)e^{0.5X_1+0.5X_2+0.5X_3}$ with $h_0(t) = 0.075t^{0.5}$. We then generate censoring times $C_i \sim U(2, 45)$ and define $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$.

We impose ∼50% missingness in X_2 and X_3 using two models: **(1)** MCAR with $P(X_2 \text{ missing}|X_1, \delta, Y) = 0.5$ and **(2)** MAR with $\text{logit}(P(X_2 \text{ missing}|X_1, \delta, Y)) = X_1$. In both cases, we set X_3 to be missing if and only if X_2 is missing. We perform multiple imputation of X_2 and X_3 using methods described in this chapter. We compute bias, relative variance (compared to analyzing the full data with no covariate missingness), and coverage in estimating model parameters across 500 simulations for each method.

We use 150 iterations for each imputation algorithm except Exact Cure, for which we use 1500 due to the slower convergence of the Metropolis-Hastings algorithms. When fitting the cure models to each imputed dataset, we use 100 EM iterations and 100 bootstrap samples to estimate variances.

Table 2.2 shows simulation results under two different missingness mechanisms for X_2 and X_3. In both cases, complete case analysis is essentially unbiased. Simulation results in this setting are broadly similar to results with missingness in only one covariate (**Table 2.1**). All imputation approaches produce little bias in estimating α_0, and all but the logT approach result in little bias for α_1. Substantial bias in estimating $\alpha_2, \alpha_3, \beta_1, \beta_2$, and β_3 can be seen for many methods. The Exact Cure Approach is the only imputation approach considered which results in unbiased estimates for all parameters. Compared to the other biased imputation approaches, the Approximate Cure approach with the interaction term results in large reductions in bias for estimating many parameters. The Approximate Cure approach without the interaction term produces smaller bias than the White and Royston approach, and the Outcome Binning approach produces further reductions of bias for some model parameters. As in the **Table 2.1** simulations, we see that the imputation approaches using T do not result in a uniform reduction of bias compared to their counterparts without T imputation.

All imputation methods result in smaller empirical variance in estimating α_0, α_1, and β_1 compared to CC analysis. Some reduction in the variance in estimating β_2 and β_3 can also be seen. The logY and logT imputation approaches resulted in much smaller variances for β_2 and β_3 than analysis of the full data. Due to the large bias of these two approaches, analysis using these approaches may produce small confidence intervals centered far from the true value. As such, we would not recommend using the logT or logY approaches for imputation.

Coverage rates for $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta_1$, and β_3 are near the nominal 95% level for all but the logT approach. Reductions in coverage for some imputation approaches can be seen for β_2. The Exact Cure approach and the Approximate Cure approach with the extra interaction term produce higher coverage rates in estimating β_2 compared to the other imputation methods.

The Exact Cure imputation approach is the best performer in terms of bias. Among the biased imputation approaches, the Approximate Cure approach with the interaction term performs the best. Among the imputation approaches which do not take the cure fraction into account, Outcome Binning without T imputation tends to produce the smallest bias overall. We see some bias in estimating a parameter associated with a fully observed variable, but it is biased to a lesser extent then the parameters for the imputed variables. These simulations demonstrate that the proposed imputation approaches have good performance when imputing multiple variables with binary or normal distributions.

Table 2.2: Cure Model Estimates with Imputation of Two Missing Covariates

Method	α_0 Bias (RV) CI[†]	α_1 Bias (RV) CI	α_2 Bias (RV) CI	α_3 Bias (RV) CI	β_1 Bias (RV) CI	β_2 Bias (RV) CI	β_3 Bias (RV) CI
Full Data	-0.01 (1.00) 0.95	0.00 (1.00) 0.95	0.01 (1.00) 0.95	0.02 (1.00) 0.95	0.00 (1.00) 0.94	0.00 (1.00) 0.92	0.00 (1.00) 0.93
Missingness Model 1: MCAR missingness in X_2, X_3							
Exact Cure	0.00 (0.64) 0.95	0.00 (0.74) 0.95	0.01 (0.46) 0.94	0.01 (0.47) 0.95	0.01 (0.76) 0.93	0.00 (0.49) 0.93	0.01 (0.48) 0.93
Approximations							
Non-Cure w/ (Y, δ)							
logY	-0.03 (0.62) 0.93	-0.01 (0.72) 0.94	0.05 (0.48) 0.93	0.07 (0.47) 0.94	0.06 (0.88) 0.93	-0.18 (1.17) 0.77	-0.18 (1.15) 0.94
White & Royston	-0.03 (0.62) 0.94	-0.01 (0.74) 0.94	0.05 (0.48) 0.94	0.06 (0.46) 0.94	0.04 (0.82) 0.93	-0.17 (1.07) 0.81	-0.15 (1.02) 0.95
Binning by (Y, δ)	-0.02 (0.62) 0.95	0.00 (0.73) 0.95	0.02 (0.48) 0.94	0.04 (0.46) 0.95	0.02 (0.79) 0.94	-0.13 (0.84) 0.87	-0.11 (0.72) 0.95
Non-Cure w/ T							
logT	-0.10 (0.56) 0.91	-0.07 (0.70) 0.91	0.15 (0.50) 0.87	0.19 (0.43) 0.89	0.06 (0.96) 0.92	-0.15 (1.38) 0.84	-0.14 (1.50) 0.98
Binning by T	-0.03 (0.63) 0.95	-0.01 (0.76) 0.95	0.06 (0.53) 0.95	0.07 (0.51) 0.95	0.05 (0.85) 0.92	-0.14 (1.07) 0.86	-0.14 (0.93) 0.95
Cure w/ (G, Y, δ)							
Approx Cure	-0.01 (0.63) 0.95	0.00 (0.71) 0.95	0.03 (0.48) 0.95	0.03 (0.47) 0.96	0.02 (0.76) 0.92	-0.16 (1.03) 0.84	-0.09 (0.64) 0.94
Approx + Int*	-0.01 (0.64) 0.94	0.01 (0.77) 0.95	0.01 (0.50) 0.96	0.02 (0.49) 0.96	0.02 (0.80) 0.94	-0.08 (0.91) 0.93	-0.02 (0.54) 0.95
Complete Case	-0.02 (0.49) 0.97	0.00 (0.46) 0.96	0.02 (0.47) 0.95	0.03 (0.48) 0.97	0.00 (0.48) 0.95	0.00 (0.48) 0.95	0.00 (0.47) 0.95
Missingness Model 2: MAR missingness in X_2, X_3 dependent on X_1							
Exact Cure	0.00 (0.62) 0.96	0.01 (0.74) 0.94	0.02 (0.44) 0.94	0.00 (0.48) 0.95	0.02 (0.71) 0.94	0.00 (0.43) 0.94	0.01 (0.38) 0.93
Approximations							
Non-Cure w/ (Y, δ)							
logY	-0.05 (0.62) 0.95	-0.02 (0.71) 0.93	0.08 (0.44) 0.92	0.09 (0.47) 0.94	0.08 (0.84) 0.91	-0.25 (1.31) 0.59	-0.23 (1.09) 0.93
White & Royston	-0.05 (0.62) 0.95	-0.02 (0.72) 0.93	0.09 (0.46) 0.93	0.10 (0.47) 0.94	0.06 (0.79) 0.92	-0.23 (1.08) 0.65	-0.20 (0.93) 0.93
Binning by (Y, δ)	-0.03 (0.64) 0.96	0.00 (0.74) 0.95	0.06 (0.48) 0.95	0.06 (0.49) 0.96	0.04 (0.72) 0.93	-0.18 (0.72) 0.77	-0.15 (0.60) 0.93
Non-Cure w/ T							
logT	-0.09 (0.56) 0.92	-0.05 (0.69) 0.92	0.15 (0.47) 0.86	0.18 (0.42) 0.90	0.06 (0.84) 0.92	-0.19 (1.16) 0.73	-0.18 (1.46) 0.97
Binning by T	-0.03 (0.68) 0.96	0.00 (0.74) 0.94	0.06 (0.54) 0.94	0.06 (0.55) 0.96	0.06 (0.80) 0.93	-0.20 (0.93) 0.75	-0.17 (0.85) 0.94
Cure w/ (G, Y, δ)							
Approx Cure	-0.04 (0.61) 0.95	-0.01 (0.72) 0.94	0.08 (0.44) 0.95	0.08 (0.48) 0.95	0.05 (0.72) 0.92	-0.23 (1.14) 0.65	-0.15 (0.58) 0.91
Approx + Int*	-0.02 (0.62) 0.95	0.01 (0.74) 0.95	0.02 (0.46) 0.95	0.03 (0.46) 0.95	0.03 (0.76) 0.95	-0.13 (0.88) 0.88	-0.05 (0.46) 0.95
Complete Case	-0.02 (0.45) 0.95	0.02 (0.38) 0.95	0.02 (0.45) 0.97	0.02 (0.47) 0.96	0.00 (0.32) 0.97	0.00 (0.40) 0.96	0.01 (0.38) 0.94

*Includes $\hat{H}_0(Y) : G : X_1$ interaction in imputation model

†CI indicates empirical coverage of 95% confidence intervals and RV indicates variance relative to analysis of the full data.

2.5 Application to Head and Neck Cancer Data

We consider data from a cohort study of time to cancer recurrence in N=1226 patients with head and neck squamous cell carcinoma (HNSCC). This study was conducted by the University of Michigan's Head and Neck Specialized Program of Research Excellence (SPORE) and included consenting patients treated for HNSCC at the University of Michigan Cancer Center between November 2003 and July 2013. Details regarding the cohort study can be found in Duffy et al. (2008) and Virani et al. (2015). Data on newly-diagnosed patients were collected from the time of diagnosis, and patients were then followed for cancer recurrence after the start of treatment. A patient is considered to have recurred if cancer becomes detectable. Personal and disease-related characteristics including age, cancer stage, cancer site, comorbidities, cigarette use, alcohol use, gender, and BMI were collected at the time of diagnosis and are reported in **Table 2.3**.

Table 2.3: Patient Characteristics

Characteristic	N (%) or Mean (SD)	Missing N (%)	Characteristic	N (%) or Mean (SD)	Missing N (%)
		Model Variables			
Age at Diagnosis	59.5 (11.7)		ACE27 Comorbidities		1 (0.01)
Cancer Stage		0 (0)	None	343 (27.9)	
I/Cis	162 (13.2)		Mild	535 (43.6)	
II	123 (10.0)		Moderate	239 (19.4)	
III	181 (14.7)		Severe	108 (8.8)	
IV	760 (61.9)		Cancer Site		0 (0)
Cigarette Use		0 (0)	Larynx	245 (19.9)	
Never	285 (23.2)		Hypopharynx	53 (4.3)	
Current	559 (45.5)		Oral Cavity	413 (33.6)	
Former	382 (31.1)		Oropharynx	515 (42.0)	
HPV Status		685 (55.8)			
Negative	320 (26.1)				
Positive	221 (18.0)				
		Auxiliary Variables			
Gender		0 (0)	Enrollment Year		0 (0)
Female	315 (25.6)		2003-2008	559 (45.5)	
Male	911 (74.3)		2009-2011	363 (29.6)	
Alcohol use		1 (0.01)	2012-2013	304 (24.7)	
Never	115 (9.3)		No. Sexual Partners	16.8 (53.4)	765 (62.3)
Current	300 (24.3)		Body Mass Index (BMI)	26.9 (5.9)	6 (0.4)
Former	810 (66.0)				

Of the 1226 patients in the study, 374 (30.5%) experienced a cancer recurrence. Of these, 149 (39.8%) had detectable cancer toward the end of their planned treatment. These patients are called "persistent" and are given a recurrence time of 1 day as exact recurrence times are unavailable for these subjects. Patients were followed for a median time of 36.6 months. Of the observed recurrences, 360 (96.2%) occurred within 36 months. Few patients had recurrences after 36 months, and the estimated survival curve had a plateau in the later half of the study (∼36-60 months). For HNSCC, it is well established that patients can be cured (Taylor, 1995). This provides some evidence that these data may follow a cure structure.

Based on biological knowledge of HNSCC recurrence and empirical evidence in the data, we assume that a subset of the study cohort had been cured of disease by treatment, and we fit a mixture cure model. We assume a Cox proportional hazards model for the hazard of cancer recurrence in the non-cured group, and we model probability of being cured of the primary HNSCC after treatment using a logistic regression. In particular, the first component is a model for time until cancer becomes detectable in the non-cured group. We include persistent patients in our analysis as persistence was defined subjectively and roughly corresponded to whether there were early signs that the cancer was present. Because persistence is an outcome of the treatment that was unobserved at baseline, these patients were included in the analysis. We fit a Cox proportional hazards cure model to the complete case data using age at diagnosis, cancer stage, cigarette use, HPV status, comorbidities, and cancer site as predictors in both parts of the mixture cure model. Results of this model fit are shown in **Table 2.4**.

In the study of HNSCC, the association between HPV status and cancer recurrence is of particular interest. However, HPV status was only obtained for 541 (44.1%) of the patients. Investigation into the missingness of HPV status (not shown) suggests that HPV missingness is associated with diagnosis date and therefore censoring time. However, assuming censoring is independent of HPV status, we can still assume HPV status is missing at random (Rathouz, 2007). We want to impute HPV status using approaches discussed and then compare results from corresponding CPH cure model estimates between imputation approaches and to complete case analysis.

We performed multiple imputation of HPV status (55.8% missing) and comorbidities (0.01% missing) using both the Approximate Cure approach with the extra interaction term and the White & Royston approach. We did not use the Exact Cure approach as we have many partially observed covariates, and when we have many covariates to impute, the Exact Cure approach becomes increasingly computationally intensive. HPV status is known to be associated with factors such as gender, smoking, alcohol use, and number of sexual partners. HPV also has a much higher prevalence for oropharyngeal cancers compared to other types of head and neck cancer. We observe that HPV status is associated with calendar time and therefore year of study enrollment. As these variables are known to be associated with HPV status, they may help us to obtain better imputations of HPV. Therefore, we use all factors in **Table 2.3** as predictors for the various imputation models, requiring us to also impute BMI, number of sexual partners, and alcohol use as part of the chained equations algorithm. We note that sexual partners has a large amount of missingness (62.3%), but we include it in the imputation algorithm due to its strong association with HPV status. Number of sexual partners is observed for 198 (28.9%) of the subjects with missing HPV status. Year of study enrollment was categorized into three intervals reflecting different rates of HPV missingness. Greater effort was made to obtain HPV status for subjects enrolled after 2008, and some samples obtained in 2012 and 2013 have not yet been tested. Some of the **Table 2.3** variables are not included in the final cure model analysis as cure models become increasingly unstable with a large amount of predictors. We therefore implicitly assume that the predictors not included in the final model are not independent predictors of the outcome. In order to satisfy the assumptions made in the derivation of the Approximate Cure approach, we assume that censoring of recurrence time (including death from other causes) does not depend on the partially observed variables and in particular HPV status and number of sexual partners. We impute categorical covariates using polytomous regression in MICE (Van Buuren and Groothuis-Oudshoorn, 2011). Number of sexual partners is imputed using predictive mean matching on the log-scale. We produced 20 imputed datasets for each approach.

Table 2.4 shows the Cox proportional hazards cure model results for two imputation algorithms and complete case analysis. Point estimates and confidence intervals are very similar between the two imputation approaches. Based on the simulation results, we may expect the biggest difference between the two approaches to be the bias in estimating parameters for HPV status. For this dataset, however, the estimates for the parameters corresponding to HPV status are very similar between the two imputation approaches. When we apply other imputation approaches discussed in this chapter to these data (not shown), we see similar results.

Differences can be seen between the model fits from imputation and from complete case analysis. Confidence intervals tend to be narrower for the imputation approaches than for complete case analysis. Point estimates tend to be somewhat similar with some exceptions. The most notable difference between the imputation and complete case fits is in the estimates for the cigarette use variable. Point estimates from the imputation approaches suggest that cigarette use may be associated with a decrease in the probability of being cured, but it is not associated with the hazard of recurrence. In contrast, the complete case analysis suggests that cigarette use is associated with a decreased hazard of recurrence in the non-cured group, but it is not associated with cure status. Additionally, the confidence intervals for some cigarette use parameters from the imputation approaches do not include the complete case point estimates. The complete case fit shows some signs of model instability.

Point estimates for HPV status parameters are similar between the complete case and imputation approaches, but the confidence intervals are smaller in the imputation model fits. This suggests that some additional information about HPV status is obtained by including information from the patients with missing HPV status.

Table 2.4: Cox Proportional Hazards Cure Model Fits to Head and Neck Data

	Complete Case Analysis, N = 540		Approx Cure + Int*, N = 1226		White and Royston, N = 1226	
Patient Characteristic	Logistic OR, 95% CI	Failure Time HR, 95% CI	Logistic OR, 95% CI	Failure Time HR, 95% CI	Logistic OR, 95% CI	Failure Time HR, 95% CI
Age at Diagnosis						
10 Year ↑	1.07 (0.91, 1.26)	1.23 (1.02, 1.45)†	1.14 (1.00, 1.31)†	1.08 (0.98, 3.95)	1.14 (0.99, 1.31)	1.08 (0.98, 1.18)
Cancer Stage						
I/Cis (ref)						
II	0.94 (0.31, 2.88)	2.17 (0.51, 9.23)	1.25 (0.57, 2.74)	1.67 (0.70, 3.95)	1.26 (0.56, 2.84)	1.68 (0.66, 4.28)
III	2.25 (0.84, 6.00)	2.91 (0.72, 11.6)	2.36 (1.18, 4.72)†	2.42 (1.22, 4.79)†	2.31 (1.19, 4.47)†	2.42 (1.13, 5.19)†
IV	2.42 (1.11, 5.31)†	2.77 (0.68, 11.1)	3.32 (1.74, 6.33)†	2.76 (1.48, 5.16)†	3.25 (1.84, 5.75)†	2.78 (1.39, 5.59)†
Cigarette Use						
Never (ref)						
Current	1.03 (0.57, 1.89)	0.63 (0.34, 1.14)	1.46 (0.97, 2.18)	0.98 (0.70, 1.38)	1.49 (1.00, 2.21)†	0.99 (0.72, 1.35)
Former	1.09 (0.63, 1.87)	0.56 (0.35, 0.90)†	1.27 (0.85, 1.90)	0.94 (0.66, 1.33)	1.28 (0.84, 1.93)	0.95 (0.69, 1.32)
HPV Status						
Negative (ref)						
Positive	0.43 (0.21, 0.87)†	0.80 (0.35, 1.82)	0.34 (0.19, 0.58)†	0.91 (0.55, 1.48)	0.38 (0.19, 0.76)†	0.82 (0.52, 1.28)
Comorbidities						
None (ref)						
Mild	1.14 (0.66, 1.97)	0.93 (0.48, 1.81)	1.14 (0.77, 1.69)	0.89 (0.65, 1.23)	1.14 (0.80, 1.62)	0.89 (0.65, 1.21)
Moderate	1.32 (0.65, 2.68)	1.47 (0.72, 2.98)	1.66 (1.08, 2.56)†	1.10 (0.75, 1.61)	1.68 (1.12, 2.53)†	1.09 (0.74, 1.60)
Severe	1.70 (0.73, 3.92)	0.79 (0.24, 2.60)	1.94 (1.10, 3.43)†	1.07 (0.63, 1.80)	1.96 (1.09, 3.52)†	1.05 (0.62, 1.80)
Cancer Site						
Larynx (ref)						
Hypopharynx	7.90 (0.00, Inf.)	2.42 (0.88, 6.64)	1.93 (0.88, 4.22)	1.43 (0.77, 2.67)	1.91 (0.88, 4.16)	1.46 (0.76, 2.80)
Oral Cavity	1.58 (0.83, 3.00)	1.33 (0.61, 2.89)	1.24 (0.81, 1.90)	1.33 (0.90, 1.97)	1.24 (0.81, 1.90)	1.34 (0.92, 1.95)
Oropharynx	1.51 (0.66, 3.44)	0.93 (0.39, 2.18)	1.68 (0.94, 3.02)	1.02 (0.62, 1.68)	1.57 (0.84, 2.94)	1.11 (0.69, 1.78)

*Includes $\hat{H}_0(Y) : G : X^{(-p)}$ interaction in imputation model † Significant at p = 0.05

2.6 Discussion

In this chapter, we have explored approaches for imputing missing covariates in the Cox proportional hazards cure model setting. We considered multiple imputation using fully conditional specification, an approach in which we impute partially observed covariates by drawing from their conditional distributions.

We derived the "exact" conditional distribution and suggested a sampling scheme for imputing normal and Bernoulli covariates in the CPH cure model setting. We also proposed several approximations to the exact distribution that are simpler and more convenient to use for imputation. Our approach can be generalized to impute covariates with different distributions. We compared the performance of our proposed imputation approaches to existing imputation methods for survival data without a cure fraction.

A simulation study demonstrates that all imputation methods considered can substantially increase precision in estimating many CPH cure model parameters compared to complete case analysis. Imputation can produce smaller variances for estimating parameters corresponding to fully observed variables compared to complete case analysis. Some variance reduction may also be seen in estimating parameters associated with the imputed variables. The Exact Cure imputation approach outperformed all other imputation approaches in terms of bias. In our simulations, all other imputation approaches tended to have some bias in estimating at least one of the parameters associated with the imputed variable/s. Among the biased imputation approaches, the Approximate Cure approach with the interaction term was the best performer. Among the approaches that do not account for the cure fraction, Outcome Binning tended to have the best performance across the three simulation settings. The approaches in which the event time is imputed without accounting for the cure structure of the data did not perform well in the cure setting and are not recommended. In the head and neck cancer example, little difference could be seen between the imputation approaches, but many differences were present between imputation and complete case analysis.

While imputation using the exact conditional distribution is a clear frontrunner in terms of bias, it is typically more difficult to implement and takes much longer to run than other methods due to the many required Metropolis-Hastings draws. These issues become even more pronounced when there is missingness in multiple covariates. If one is willing to allow some bias in estimating some model parameters (particularly those associated with the imputed variables), then the Approximate Cure imputation approach with the interaction term may be preferred. For example, if we are only adjusting for an imputed variable as a possible confounder, then adding some bias in estimating its parameters in exchange for computational simplicity may be acceptable. If we desire an even simpler imputation scheme and do not want to impute cure status, we may still be able to obtain some bias reduction by using Outcome Binning without the event time imputation rather than other existing imputation approaches for survival data without a cure fraction.

We compare imputation approaches in terms of performance in estimating CPH cure model parameters, and most of the imputation approaches proposed are compatible with and directly motivated by the final modeling strategy. If we change the modeling strategy (for example, if we want to fit an accelerated failure time model with a cure fraction), then the imputation approach may need to be adapted and the comparative performance of the approaches may change. Additionally, although simulations suggest there is a difference between imputation approaches, there may not always be a large practical difference when applied to particular datasets as seen with the head and neck cancer data. The presented simulations are limited to a setting with normal and binary covariates with linear covariate effects in the logistic and failure time models. When imputing covariates with other distributions (e.g. ordered categorical), the comparative performance of the imputation approaches may be different. Also, if the failure time or logistic models include interactions/non-linear effects of the partially observed covariates, the difference between the Exact Cure method and the approximated methods would be expected to be even more pronounced than in the linear effects case considered here (Bartlett et al., 2014).

We note that $H_0(t)$ in the CPH model is really an infinite-dimensional parameter, and we do not directly incorporate this uncertainty into the estimation procedure. Additionally, we only consider multiple imputation using fully conditional specification. Fully conditional specification is convenient to use for imputation as it does not require us to explicitly specify the joint distribution of the covariates. However, in the case of multiple imputed variables, the assumed distributions for each partially observed $X^{(p)}|X^{(-p)}$ are not guaranteed to be compatible and form a valid joint distribution. In some cases, this could lead to problems (e.g. bias) when estimating parameters in the final model fitting (Bartlett et al., 2014). Several authors have provided conditions in which FCS is equivalent to joint model imputation and converges to the desired sampling distribution (Hughes et al., 2014; Liu et al., 2013).

Chapter III

Sequential Imputation for Models with Latent Variables

3.1 Introduction

Models that involve latent or partially latent variables in addition to an outcome variable and covariates are frequently the target for estimation and inference. For example, in the Cox proportional hazards mixture cure model, partially latent cure status describes whether individuals are at risk for the event of interest. Cure status is only partially latent because subjects with observed events are known to be non-cured. Another popular model with latent variables is the linear mixed model, where fully latent random effects account for correlation within clusters.

Additional considerations arise when dealing with missing covariates and/or outcomes in the presence of latent variables. Many authors have explored the issue of missing data for models with latent variables under assumptions that missingness is independent of the latent variable given the observed data. We do this in **Chapter II** of this dissertation. In this chapter, we explore a generalization of this missingness mechanism that allows covariate/outcome missingness to depend on the latent variable, which is a "missing not at random" (MNAR) mechanism (Little and Rubin, 2002). Previous examples of such mechanisms are called "latent ignorable" or "latent missing at random" (LMAR) missingness (Frangakis and Rubin, 1999; Harel, 2003; Harel and Schafer, 2009). For example, suppose we model a longitudinal outcome using a mixed model. One common LMAR scenario in the literature relates dropout to the random effect, which can be viewed as a measure of an individual's propensity to drop out.

In general, the underlying missingness mechanism can never be determined from the data alone, and inference under MNAR may be sensitive to unverifiable assumptions about the missingness mechanism. Additionally, inference under MNAR is susceptible to underidentification or weak identification of the model parameters (Little, 1995; Molenberghs et al., 2008). In this chapter, we consider a particular MNAR missingness mechanism (LMAR) in which missingness depends on unknown information *only* through the latent variable, which by assumption has a structured relationship with the observed variables. We may view LMAR missingness as a somewhat mild departure from MAR. Still, we must keep these issues in mind when handling missing data under LMAR.

One approach for handling missing data is to analyze only the fully observed subset of the data (complete case analysis). When missingness is LMAR, this approach will generally produce biased results (Little and Rubin, 2002). Several authors have discussed likelihood-based approaches for linear mixed models with missingness dependent on the random effect (e.g. Little, 1995; Wu and Carroll, 1988). These methods often involve an EM algorithm or a likelihood that has integrated out the latent variable.

Multiple imputation is a common general approach for dealing with missing data. One approach to multiple imputation requires one to specify a joint distribution for all the variables and use that joint distribution for imputation, usually in a Gibbs sampling-type algorithm. Each variable with missing values can be sequentially imputed using its conditional distribution, which is determined by the joint distribution. The distribution of the sampled parameters can then be used for inference. An alternative approach to inference is to extract m completed datasets (each consisting of the observed data plus imputed values), analyze each completed dataset using the desired model, and combine the results using Rubin's rules for inference from multiply imputed datasets (Rubin, 1987).

Several authors have proposed joint modeling approaches for handling latent ignorable missingness in specific modeling settings (Jung, 2007; Yang et al., 2008; Lu et al., 2011). Harel (2003) proposes a non-iterative imputation approach for dealing with general latent-dependent missingness under a joint model, but literature is sparse on the implementation of imputation based on joint models under general latent ignorable missingness. The main drawback of the joint modeling approach to imputation is that specification of the joint distribution may be difficult or too restrictive.

Chained equations imputation is an alternative to joint modeling in which variables are imputed iteratively in a series of univariate imputation steps (Raghunathan, 2001;

Van Buuren et al., 2006). These steps are usually accomplished using standard regression models, and these regressions as a set usually do not correspond to a valid joint distribution. This approach is simple and flexible, but it is less coherent than joint modeling and may not incorporate assumptions about the outcome model directly. Most literature on chained equations assumes that missingness is independent of all unobserved information, called “missing at random” (MAR) (Little and Rubin, 2002), and some authors have explored particular MNAR settings (e.g. Van Buuren, 2007; Little et al., 2009; Giusti and Little, 2011). An alternative approach proposed in Bartlett et al. (2014) incorporates the outcome model into the chained equations procedure, leading to improved properties. Similar findings are given in White and Royston (2009) and **Chapter II**. In the context of models with latent variable models, however, we have not found any literature exploring chained equations under latent ignorable missingness in general.

In this chapter, we develop a sequential imputation algorithm that can handle MAR and LMAR covariate and outcome missingness for models with latent or partially latent variables. The proposed method imputes the latent variable as part of the missing data, allowing the latent variable to be directly used when imputing the missing covariate/outcome values. The proposed algorithm is very flexible and can accommodate either a Gibbs sampling-type approach under joint model or a chained equations-type approach to imputation. In the joint modeling setting, we describe how we can directly incorporate our assumptions about the outcome and missingness model structures into the imputation procedure and provide several results. We then use results under a joint model to inform a chained equations-type imputation approach without a joint model.

Many works have explored MAR-based imputation in settings with latent variables under a joint model (e.g. Schafer, 1997; Schafer and Yucel, 2002; Chung et al., 2006). However, a distinguishing feature of the proposed algorithm over existing methods is that it provides a substantive model compatible approach to imputation, where the form of the imputation distribution is directly motivated by our outcome modeling assumptions without requiring a fully-specified joint model. This approach to imputation has been previously explored in the context of covariate imputation in Bartlett et al. (2014), but a general imputation algorithm for handling missingness in multiple variables has not previously been considered. To our knowledge, no other work has provided a chained equations algorithm for performing imputation for general latent variable models under

MAR. Under LMAR, the proposed methods represent the first sequential imputation algorithm for general LMAR settings.

In **Section 3.2**, we define latent ignorability. In **Sections 3.3 and 3.4**, we describe the proposed imputation approach. In **Section 3.5**, we present simulations that evaluate the performance of our method under a variety of scenarios. In **Section 3.6**, we apply the proposed methods to a study of time to recurrence in patients with head and neck cancer. In **Section 3.7**, we present a discussion.

3.2 Latent Ignorability

Suppose that the goal is to make inference about a model for outcome Y given covariates X and a latent (or partially latent) mixing variable, L. For example, the outcome model may be a linear mixed model with a latent random intercept. We may also be interested in the model for $L|X$. We restrict our attention to situations in which, if all of the covariate and outcome information were observed, the outcome model would be fully identified, and estimation using likelihood-based methods would be possible and lead to consistent parameter estimates. We consider missingness in X and/or Y, and we allow missingness to be related to the latent variable, L.

Let vector $D_i = (X_i, Y_i)$ represent the (possibly incomplete) data for subject i. We assume D_i and L_i are independent across subjects. Let R_i^D be a vector corresponding to whether each element of D_i is observed and R_i^L be an indicator for whether L_i is known (can be 0 for all subjects). Define $R_i = (R_i^D, R_i^L)$. For any vector V_i, let $V_i^{(obs)}$ and $V_i^{(mis)}$ be the observed and missing elements of V_i.

We assume that missingness in D_i is independent of $D_i^{(mis)}$ and R_i^L such that

$$f(R_i^D|D_i, L_i, R_i^L; \phi^D) = f(R_i^D|D_i^{(obs)}, L_i; \phi^D) \tag{3.1}$$

We assume that ϕ^D is distinct from all other model parameters. We call assumption (3.1) the "latent missing at random" (LMAR) or "latent ignorability" assumption. This missingness mechanism was first studied in Frangakis and Rubin (1999) and is a special case of latent ignorability explored in Harel (2003) and Harel and Schafer (2009). In longitudinal data analysis, a similar mechanism relating missingness in Y to latent random effects in a linear mixed model has been explored by many authors including Wu and Carroll (1988), Follmann and Wu (1995), Little (1995), and McCulloch et al. (2016). Since L_i is latent or partially latent by definition, the mechanism in (3.1) is a type of MNAR, and when (3.1) does not depend on L_i, the mechanism reduces to MAR. We can view LMAR as a generalization of MAR with less restrictive assumptions.

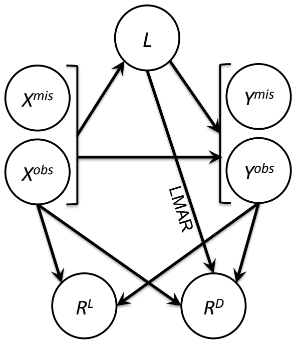
We now consider assumptions regarding missingness in L, which may be latent or partially latent. We make a subtle distinction between "partially latent" and "partially missing" variables. Latent variable L can be viewed as a modeling construct representing unobserved or perhaps unobservable quantities. The "observed" values of the partially latent L are usually just a function of the observed data, $D^{(obs)}$, and therefore contain no additional information. For example, known values of the partially latent cure status in a Cox proportional hazards cure model are entirely determined by the event indicator and the event/censoring time for each subject. In this way, partially latent variables are different from partially missing variables, which may contain additional information in their observed values. However, we will treat latent and partially latent variables as if they were missing data for the purposes of this method.

When L_i is fully latent, we can view missingness in L_i as missing completely at random (MCAR) with probability of missingness equal to 1. When L_i is partially latent, we allow missingness in L_i to depend on $D_i^{(obs)}$ (so L is MAR) such that

$$f(R_i^L | D_i, L_i, R_i^D; \phi^L) = f(R_i^L | D_i^{(obs)}; \phi^L) \tag{3.2}$$

Figure 3.1 shows the assumed relationships between variables. The arrows represent dependence. For example, R^L may depend on $X^{(obs)}$ and $Y^{(obs)}$.

Figure 3.1: Variable Relationships under Latent Ignorability



Example 1, Linear Mixed Model with a Random Intercept: Suppose our model for multivariate outcome Y_i is a linear mixed model with a latent random intercept, b_i, and covariates X_i. This model is commonly used for longitudinal data, where the outcome is measured within individuals over time. In such a setting, outcome missingness is particularly common due to dropout. Many authors have described scenarios in which dropout may be related to the random effects (Wu and Carroll, 1988; Little, 1995; Yang et al., 2008). In this example, b_i represents an individual's propensity to drop out. This is a LMAR mechanism with $L_i = b_i$. Covariate missingness may also be LMAR.

Example 2, Cox Proportional Hazards Mixture Cure Model: The Cox proportional hazards (CPH) mixture cure model is used in event time analysis when some ("cured") subjects are unable to experience the event of interest (Sy and Taylor, 2000). For subjects with events, cure status is known, and it is unknown for censored subjects. Therefore, cure status is partially latent. Missingness in cure status is entirely determined by observed information, so its missingness can be viewed as MAR. Suppose we have covariate missingness. We can imagine scenarios in which covariate missingness may depend on the underlying cure status. For example, suppose covariate information is collected through a patient survey. Cured subjects may be more or less likely to answer certain survey questions, resulting in an association between missingness and cure status. Additionally, cure status may be related to an unmeasured confounder that is related to missingness. This will induce a dependence between missingness and cure status. We consider a similar LMAR mechanism in our data application.

Example 3, Mixture of Generalized Linear Models: Suppose our outcome Y is generated from a mixture of K generalized linear models (GLMs). Let C_i be a fully latent mixing variable indicating which element of the mixture distribution generated the observation for subject i. Missingness in C_i can be viewed as MCAR with probability 1. If covariate or outcome missingness is related to C, missingness is LMAR. For example, suppose our data are collected using K different populations. For example, we may collect data and multiple different locations and not record the location. The covariate/outcome missingness rates may vary by population, resulting in LMAR missingness.

3.3 Imputation of Missing Data

We first propose a sequential imputation algorithm for dealing with ignorable and latent ignorable covariate and outcome missingness under a joint model for all the variables. We treat the latent variable as part of the missing data, and we use the form of the joint model to determine how each variable with missing values should be imputed. In particular, we determine which variables need to be included as predictors for each imputation model and describe the components of the joint model (e.g. outcome model, missingness model, covariate model) that are used for imputing each variable. We then use these results to guide our choice of sequential imputation models when a joint model is not specified.

3.3.1 Joint Modeling Approach

Suppose that the data are directly modeled using a fully-specified joint model as follows:

$$f(D, L, R; \nu) = \prod_{i=1}^{n} f(R_i|Y_i, X_i, L_i; \phi) f(Y_i|X_i, L_i; \theta) f(L_i|X_i; \omega) f(X_i; \psi) \tag{3.3}$$

where $\nu = (\phi, \theta, \omega, \psi)$ is the set of all model parameters. We assume a flat prior for ν such that ϕ, θ, ω, and ψ are all a priori independent (so they are distinct). The factorization (3.3) is a form of shared parameter model, where the latent variable is related both to missingness and to the distribution for Y_i (Little and Rubin, 2002).

We can impute missing values of D and L by iteratively drawing the missing values from their posterior predictive distributions, $D^{(mis)} \sim f(D^{(mis)}|D^{(obs)}, L, R)$ and $L^{(mis)} \sim f(L^{(mis)}|D, L^{(obs)}, R)$. This leads to draws from the joint posterior predictive distribution, $f(D^{(mis)}, L^{(mis)}|D^{(obs)}, L^{(obs)}, R)$ (Little and Rubin, 2002). Define $\rho = (\theta, \omega, \psi)$. We note the following properties of the (conditional) posterior predictive distributions:

Lemma 1: *Under MAR and LMAR, we can ignore $R = (R^D, R^L)$ when imputing D.*

The missingness mechanism is ignorable for imputing $D^{(mis)}$ if $f(D^{(mis)}|D^{(obs)}, L, R) =$

$f(D^{(mis)}|D^{(obs)},L)$. Using assumptions (3.1) and (3.2) and assuming ϕ and ρ are distinct,

$$\begin{aligned}f(D^{(mis)}|D^{(obs)},L,R) =&\frac{1}{f(D^{(obs)},L,R)}\int\int f(D,L,R,\nu)d\rho d\phi\\ =&\frac{1}{f(D^{(obs)},L,R)}\int f(R|D,L;\phi)\left[\int f(D,L;\rho)f(\rho)d\rho\right]f(\phi)d\phi\\ =&\frac{f(D^{(mis)}|D^{(obs)},L)f(D^{(obs)},L)}{f(D^{(obs)},L,R)}\int f(R|D^{(obs)},L;\phi)f(\phi)d\phi\\ =&f(D^{(mis)}|D^{(obs)},L)\end{aligned}$$

Therefore, the missingness mechanism is ignorable for imputing D. A similar result for a related latent ignorable missingness setting was shown in Harel (2003). We note that in practice, draws from the posterior predictive distribution are obtained by first drawing the model parameter ρ from its posterior distribution and then drawing $D^{(mis)}$ from $f(D^{(mis)}|D^{(obs)},L,R;\rho)$. We can perform both of these draws ignoring R.

Lemma 2: *Under MAR (but not under LMAR), we can ignore $R = (R^D, R^L)$ when imputing L.*

The missingness mechanism is ignorable for imputing $L^{(mis)}$ if $f(L^{(mis)}|L^{(obs)},D,R) = f(L^{(mis)}|L^{(obs)},D)$. Again using assumptions (3.1) and (3.2) and assuming ϕ and ρ are distinct,

$$\begin{aligned}f(L^{(mis)}|L^{(obs)},D,R) =&\frac{1}{f(L^{(obs)},D,R)}\int\int f(D,L,R,\nu)d\rho d\phi\\ =&\frac{1}{f(L^{(obs)},D,R)}\int f(R|D,L;\phi)\left[\int f(D,L;\rho)f(\rho)\right]d\rho f(\phi)d\phi\\ =&\frac{f(L^{(mis)}|L^{(obs)},D)f(L^{(obs)},D)}{f(L^{(obs)},D,R)}\int f(R|D^{(obs)},L;\phi)f(\phi)d\phi\end{aligned}$$

Suppose first that missingness is MAR. Then, $f(R|D^{(obs)},L;\phi) = f(R|D^{(obs)},L^{(obs)};\phi)$ and $f(L^{(mis)}|L^{(obs)},D,R) = f(L^{(mis)}|L^{(obs)},D)$. Therefore, R is ignorable. Under LMAR, however, the term $\int f(R|D^{(obs)},L;\phi)f(\phi)d\phi$ depends on $L^{(mis)}$, so R is not ignorable.

Lemma 3: *Suppose that missingness in subset S of $\{D,L\}$ is MAR. We can ignore the corresponding subset of R when imputing L provided a distinctness property holds.*

Let R^S denote the set of missingness indicators for S and R^{-S} denote the missingness indicators for the remaining variables in $\{D, L\}$. Note by assumption (3.2), $L \subset S$. Let $f(R_i^{-S}|D_i, L_i; \phi) = f(R_i^{-S}|D_i^{(obs)}, L_i; \phi^{-S})$ and $f(R_i^S|D_i, L_i, R_i^{-S}; \phi) = f(R_i^S|D_i^{(obs)}, L_i^{(obs)}; \phi^S)$. Assume also that ϕ^{-S} and ϕ^S are distinct (a priori independent). Then we have

$$\begin{aligned} f(R|D^{(obs)}, L; \phi) =& f(R^S|D^{(obs)}, L^{(obs)}; \phi^S) f(R^{-S}|D^{(obs)}, L; \phi^{-S}) \implies \\ f(L^{(mis)}|L^{(obs)}, D, R) \propto& f(L^{(mis)}|L^{(obs)}, D) \int f(R^{-S}|D^{(obs)}, L; \phi^{-S}) f(\phi^{-S}) d\phi^{-S} \end{aligned}$$

The contribution of R^S drops out of the posterior predictive distribution, so R^S is ignorable. A similar result, called “ignorability for submodels”, was shown in Harel (2003). For an example of submodel ignorability, see our data application in **Section 3.6**.

3.3.2 Sequential Imputation Algorithm under a Joint Model

Rather than drawing $D^{(mis)}$ and $L^{(mis)}$ from their posterior predictive distributions directly, we instead impute each variable with missingness sequentially through a series of univariate imputation steps. This will approximate draws of $D^{(mis)}$ and $L^{(mis)}$ from their posterior predictive distributions. At each step, missing values of a particular variable, V, are drawn from its posterior predictive distribution. In practice, we specify the full conditional distribution of V given all other variables (with parameter v) and obtain a draw from the posterior predictive distribution of V by 1) drawing v from its posterior distribution and 2) drawing missing values of V from its full conditional distribution at the drawn v.

Suppose we partition $\nu = (\phi, \rho)$, where ϕ represents the missingness model parameters and ρ represents all other model parameters. Define $D^{(p)}$ to be the p^{th} variable in D and $D^{(-p)}$ to be all variables in D except $D^{(p)}$. We sequentially impute the missing values of L and $D^{(1)}, \dots, D^{(d)}$ for D with d elements and repeat for many iterations until convergence. Just before the imputation step for each variable, we draw the parameters necessary for the imputation from a current estimate of the parameters' posterior distribution. Depending on the variable being imputed, this posterior may or may not condition on the most recent imputed values of the variable being imputed.

Assume we have independence of (D, L, R) across i. We show in *Lemma 1* that we can impute D ignoring R under both MAR and LMAR. We can use a similar argument to show that we can impute missing $D_i^{(p)}$ from $f(D_i^{(p)}|D_i^{(-p)}, L_i; \rho)$. From *Lemma 2*, we can impute missing L_i from $f(L_i|D_i, R_i; \nu) = f(L_i|D_i; \rho)$ under MAR. See **Appendix A** for details.

The **sequential imputation algorithm under MAR** proceeds as follows. In the imputation step for each variable, we treat the most recent imputations of the other variables as observed. At each iteration, we draw missing data and parameters from:

$$\begin{aligned}
\text{Impute } L: &\quad \rho \sim f(\rho|D, L^{(obs)}), \quad L_i^{(mis)} \sim f(L_i|D_i; \rho) \\
\text{Impute } D^{(1)}: &\quad \rho \sim f(\rho|D, L), \quad D_i^{(1,mis)} \sim f(D_i^{(1)}|D_i^{(-1)}, L_i; \rho) \\
\ldots\text{Impute } D^{(d)}: &\quad \rho \sim f(\rho|D, L), \quad D_i^{(d,mis)} \sim f(D_i^{(d)}|D_i^{(-d)}, L_i; \rho)
\end{aligned} \tag{3.4}$$

In the **sequential imputation algorithm under LMAR**, the steps for drawing missing values of D can proceed as in the MAR algorithm. In the step for imputing L under LMAR, we draw missing L and ν from:

$$\rho \sim f(\rho|D, L^{(obs)}), \quad \phi \sim f(\phi|D, L, R), \quad L_i^{(mis)} \sim f(L_i|D_i, R_i; \nu) \tag{3.5}$$

Iteration is required even if we have only one variable in D with missing values. We can ignore the imputation steps for each fully observed $D^{(p)}$. Details describing how the proposed algorithm can obtain an approximate draw of the missing data from the posterior predictive distributions and how to accomplish the various draws are in **Appendix A**. We initialize the missing values for each variable in D by drawing from the observed values with equal probability. We can initialize missing L using the distribution $f(L|X)$ obtained from a fit to the data with fully observed D (using methods that treat L as latent).

We perform the imputation procedure m times to construct m filled-in datasets (with m different initializations). We then estimate ρ by fitting our model of interest to each of the imputed datasets ignoring R. When we perform this analysis, we may choose to use only imputed D, only imputed L, or both. We can then use Rubin's combining rules to obtain a single set of parameter estimates and errors from which we make the desired inference (Rubin, 1987).

It is important to consider the impact of ignoring R for each one of these final analysis strategies. Harel and Schafer (2009) shows that when imputed L is included in the final analysis, we can ignore R. This result holds true under MAR and LMAR and whether or not imputed D is included in the final analysis. In *Lemmas 4-5*, we explore the ignorability of R when performing a final analysis using only the imputed D. We show that R is ignorable under MAR and that such an analysis ignoring R under LMAR is valid but not fully efficient. Even with a potential loss of efficiency, we may still choose to perform our final analysis ignoring the imputed L as this may provide improved numerical stability of the algorithm and more robustness to misspecification of the imputation models, and we may have little loss of efficiency in practice.

Lemma 4: *R is ignorable for ρ in a final analysis using only imputed D under MAR*

Suppose we perform our final analysis using the imputed values of D but ignoring the imputed L and again suppose that ϕ and ρ are distinct. In a Bayesian analysis, we want to make inference from the joint posterior of ϕ and ρ:

$$\begin{aligned}
f(\nu|D, L^{(obs)}, R) \propto & f(R|L^{(obs)}, D; \nu) f(L^{(obs)}, D; \rho) f(\nu) \\
\propto & \left[\int f(R|L, D; \phi) f(L^{(mis)}|L^{(obs)}, D; \rho) dL^{(mis)} \right] f(L^{(obs)}, D; \rho) f(\phi) f(\rho) \\
\propto & f(R|L^{(obs)}, D^{(obs}; \phi) f(\phi) f(L^{(obs)}, D; \rho) f(\rho) \text{ under MAR}
\end{aligned}$$

The posterior distributions of ϕ and ρ separate, and the posterior for ρ is independent of R. Therefore, we can ignore R for inference about ρ under MAR.

Lemma 5: *A final analysis for making inference about ρ using imputed D (but not imputed L) and ignoring R is valid but not fully efficient under LMAR.*

Under the setting of *Lemma 4* except assuming LMAR, we again have that

$$f(\nu|D, L^{(obs)}, R) \propto \left[\int f(R|L, D; \phi) f(L^{(mis)}|L^{(obs)}, D; \rho) dL^{(mis)} \right] f(\phi) f(L^{(obs)}, D; \rho) f(\rho)$$

Under LMAR, $f(R|L, D; \phi)$ depends on $L^{(mis)}$, so the contribution of R and ϕ does not factor out of the integral. Therefore, we cannot separate ϕ and ρ in the above equation. We rewrite the above equation as $f(\nu|D, L^{(obs)}, R) \propto h(\nu) f(\rho|D, L^{(obs)})$ where

$h(\nu) = \left[\int f(R|L, D; \phi) f(L^{(mis)}|L^{(obs)}, D; \rho) dL^{(mis)}\right] f(\phi)$. Clearly, ν and ρ are not distinct. However, L is MAR given imputed D. Under the ignorability conditions in Little and Rubin (2002) (pg. 119-120), inference ignoring the contribution of R (using $f(\rho|D, L^{(obs)})$) will be valid from a frequency perspective but may not be fully efficient. Intuitively, the loss of efficiency comes from a loss of information about the missing L that comes from ignoring R under LMAR. However, analysis is still valid since missing L is MAR given D.

3.3.3 Specifying Predictive Distributions under a Joint Model

Assuming a fully-specified joint model as in (3.3), we derive the full conditional distributions $f(D_i^{(p)}|D_i^{(-p)}, L_i; \rho)$ and $f(L_i|D_i, R_i; \nu)$ used to impute each of the variables with missingness using the property that the full conditional distributions are proportional to (3.3). This approach allows us to directly incorporate our modeling assumptions into the imputation.

Predictive Distribution of Latent Variable for Imputation

Define R^S and R^{-S} as in *Lemma 3* and assume the corresponding parameters ϕ^S and ϕ^{-S} are distinct. Then by *Lemma 3* we can ignore R^S when imputing L. Using assumptions (3.1)-(3.2) and joint model (3.3) and treating terms that do not depend on L_i as constants, we have

$$f(L_i|X_i, Y_i, R_i^{-S}; \nu) \propto f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, L_i; \phi^{-S}) f(Y_i|X_i, L_i; \theta) f(L_i|X_i; \omega) \tag{3.6}$$

Under MAR, (3.6) simplifies to

$$f(L_i|X_i, Y_i, R_i^{-S}; \nu) \propto f(Y_i|X_i, L_i; \theta) f(L_i|X_i; \omega) \propto f(L_i|X_i, Y_i; \rho)$$

When treated as a function of L_i, expression (3.6) is proportional to the desired full conditional distribution. We will call the distribution known up to proportionality the "kernel." The kernel in (3.6) involves the distribution of R_i^{-S} under LMAR but not under MAR. In order to impute L_i under LMAR using (3.6), we need to specify a model for R_i^{-S}.

In some particular settings (for example, when L_i is binary), we can use (3.6) to directly derive the full conditional distribution. When L_i is continuous, the distribution may only be known up to a proportionality constant. In this case, we may need to use more advanced techniques to impute L_i using (3.6). Many methods exists in the literature for drawing from a distribution knowing only the kernel. These include the Metropolis-Hastings algorithm and rejection sampling. For examples of such methods applied in the context of imputation, see Bartlett et al. (2014) and **Appendix D**.

Predictive Distributions of Covariates and Outcome for Imputation

In *Lemma 1*, we show that we can impute missing values of D ignoring the missingness mechanism under MAR and LMAR. We can similarly impute missing values of individual variables in D from their full conditional distributions without conditioning on R.

We first determine the full conditional distribution for imputing missing outcome values. We note that Y may be uni- or multivariate. Suppose that we are imputing the t^{th} element of Y_i, denoted $Y_i^{(t)}$. Let $Y_i^{(-t)}$ represent the terms in Y_i excluding $Y_i^{(t)}$. Using joint model (3.3), we can write the conditional distribution for imputing $Y_i^{(t)}$ under MAR and LMAR as

$$f(Y_i^{(t)}|Y_i^{(-t)}, X_i, L_i; \rho) \propto f(Y_i, X_i, L_i; \rho) \propto f(Y_i|X_i, L_i; \theta) \tag{3.7}$$

When $Y_i^{(t)} = Y_i$, the conditional distribution is equal to $f(Y_i|X_i, L_i; \theta)$.

Suppose that we are imputing the t^{th} covariate in X_i, denoted $X_i^{(t)}$. Let $f(X_i^{(t)}|X_i^{(-t)}; \psi)$ be the conditional distribution of $X_i^{(t)}$ given all other variables in X_i. Under joint model (3.3), we can write the conditional distribution for imputing $X_i^{(t)}$ under MAR and LMAR as

$$f(X_i^{(t)}|X_i^{(-t)}, Y_i, L_i; \rho) \propto f(Y_i|X_i, L_i; \theta) f(L_i|X_i; \omega) f(X_i^{(t)}|X_i^{(-t)}; \psi) \tag{3.8}$$

Equations (3.7) and (3.8) provide the kernels of the distributions we can use to impute outcomes and covariates in D. The kernels take the same form under MAR and LMAR, and they do not involve a model for R. As with the latent variable imputation, distributions (3.7) and (3.8) may only be known up to proportionality, requiring more advanced statistical methods to draw imputations. In **Appendix D**, we provide details regarding how we can perform each of the imputation steps for the examples discussed **Section 3.2**.

3.3.4 Sequential Imputation without Specifying a Joint Model

The imputation distributions derived in previously were developed assuming a fully-specified joint model as in (3.3), but often we will not want to specify such a joint model in practice. Specification of the joint model may be particularly difficult or restrictive in the setting with missingness in covariates of different types. Rather than specifying an explicit joint distribution as in (3.3), we propose following the imputation approach defined in (3.4) and (3.5) and imputing missing values using distributions (3.6) - (3.8) by specifying only the modeling components needed for each imputation or by approximating (3.6)-(3.8) using simpler imputation distributions. In practice, the resulting conditional distributions may not together correspond to a valid joint distribution for all the variables.

Imputation of missing values of Y using (3.7) requires a model for $Y|X, L$, and imputation of missing L using (3.6) further requires a model for $L|X$ and, under LMAR, a model for missingness. Imputation of missing covariate $X_i^{(t)}$ using (3.8) requires us to specify $f(X_i^{(t)}|X_i^{(-t)};\psi)$. In practice, we could specify explicit models for $Y|X, L$ and $L|X$ (and possibly missingness) but avoid specifying $f(X|\psi)$ by instead specifying $f(X_i^{(t)}|X_i^{(-t)};\psi)$ for covariates with missingness using simple regression models. A related approach (substantive model compatible fully conditional specification, SMC-FCS) introduced in Bartlett et al. (2014) incorporates outcome model assumptions to inform the structure of the imputation distributions without explicitly specifying the joint distribution. An additional appealing feature of SMC-FCS is that it has additional flexibility over joint modeling in terms of imputation model specification, and it also involves imputing with a model that is congenial with the final analysis model. By "uncongeniality," we mean that the imputation model and the final data analysis model are incompatible (Meng, 1994). Since SMC-FCS directly uses the final analysis model in the imputation procedure, it is attractive from a congeniality point of view.

Imputation using (3.6) - (3.8) may be difficult when the distributions are known only up to proportionality. An alternative, simpler sequential imputation approach involves using equations (3.6) - (3.8) solely to define what predictors are needed for each imputation. Specifically, equation (3.6) suggests that some function of Y, X, and possibly R (under LMAR) should be used as predictors when imputing L. Equation (3.7) suggests we need X, L, and $Y^{(-t)}$ when imputing $Y^{(t)}$, and equation (3.8) suggests we need Y, L, and $X^{(-t)}$ when imputing $X^{(t)}$. We can then perform imputation (by specifying a regression model for imputing each variable) using standard software for chained equations imputation (Raghunathan, 2001; Van Buuren et al., 2006). Such an approach would allow for increased flexibility in model specification (for example, by including quadratic or interaction terms) while still allowing L to be used in the imputation. We may view the *working* model actually used for imputation as an approximation to the "true" conditional model as in (3.6)-(3.8). We recommend imputing L using the kernel in (3.6) if possible.

We can modify the imputation algorithm in (3.4) and (3.5) for settings in which we cannot easily specify a joint model. Let $\tilde{f}(D^{(p)}|D^{(-p)}, L; \rho^p)$ be the *working* conditional distribution of $D^{(p)}$ *used for imputation* and $\tilde{f}(\rho^p|D, L)$ denote the posterior distribution of ρ^p. We can then replace the step in (3.4) for imputing $D^{(p)}$ with the following:

$$\text{Impute } D^{(p)}: \quad \rho^p \sim \tilde{f}(\rho^p|D, L) \qquad\qquad D_i^{(p,mis)} \sim \tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \rho^p)$$

Suppose we specify explicit models for $Y|X, L$ and $L|X$, and in the case of an imputed covariate, we specify a regression model form for $f(X^{(p)}|X^{(p)}; \psi)$. Under flat priors, we can obtain a (approximate) draw of ρ^p by fitting the corresponding models to a bootstrap sample of (D, L). Given a draw for ρ^p, we can then impute $D^{(p)}$ using equations (3.6) - (3.8) or using the regression model $f(D^{(p)}|D^{(-p)}, L; \rho^p)$.

The proposed imputation approach, therefore, can be easily modified to accommodate settings without a fully-specified joint distribution. Indeed, Gelman (2004) argues that "having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g. zero/nonzero features in income components, bounds, skip patterns, nonlinearity, interactions)." The SMC-FCS and chained equations approaches allow these unique features of the data to be directly incorporated in the imputation models. This approach allows for greater flexibility in the specification of the imputation distributions and simplifies the implementation.

3.4 Identifiability and Convergence

As with all missing data methods involving MNAR assumptions, one big concern is how to model the missingness mechanism (which will be unverifiable) (Molenberghs et al., 2008). Another concern is whether the resulting model parameters are identifiable (Little, 1995). Even when the parameters are technically identified, weak identifiability may also have implications on the numerical convergence of the proposed imputation algorithm. In this section, we briefly comment on some identifiability- and convergence-related issues that arise in the application of the proposed imputation algorithm.

3.4.1 Modeling the Missingness Mechanism

Under LMAR, we must specify a model for R^D (or some subset R^{-S} following *Lemma 3*). While we can conceive of many different models for R^D, the model parameter $\nu = (\phi, \rho)$ may not always be identifiable. In some specific settings (e.g. Wu and Carroll, 1988; Miao et al., 2016), identifiability has been demonstrated analytically, but exploring identifiability can be difficult in general. We explore identifiability in several particular modeling settings in **Appendices B and C**. In this chapter, we will not attempt to prove identifiability properties for general LMAR mechanisms. Instead, we will provide some guidance for applying the proposed methods in the presence of possible identifiability issues.

In order to reduce the potential for identifiability issues, many authors (e.g. Little, 2009) recommend that we avoid overburdening the missingness model with extra variables. However, if we leave out variables that should be in the model, we may introduce bias in estimating the parameter of interest as seen in our simulations. In our simulations, imputation with LMAR *outcome* missingness tended to be more susceptible to identifiability problems than *covariate* missingness. Some authors recommend performing a sensitivity analysis in which we specify the form of the missingness model and carry out analysis using fixed values for ϕ^D (e.g. Little, 2009). We can then perform the desired analysis many times using different values for ϕ^D. This approach allows us to directly study the impact of ϕ^D on inference and avoid estimating the parameters of the missingness model. Additionally, MNAR missingness mechanisms are known to be particularly sensitive to assumptions about the structure of the missingness mechanism, and we could perform a sensitivity analysis using different missingness model structures (Little, 1995).

We take this approach in our head and neck cancer example. These sensitivity approaches allow the proposed methods to be applied while avoiding some of the pitfalls of MNAR settings.

3.4.2 A Note on Convergence

When the conditional models used for imputation correspond to a well-defined joint distribution with identified parameters, our imputation algorithm is expected to converge to draws of the joint posterior distribution for the missing data (Liu et al., 2013; Hughes et al., 2014; Bartlett et al., 2014). When the imputation models do not correspond to a valid joint distribution (called incompatibility), our imputation method is not guaranteed to converge. However, several works have demonstrated that we can often still obtain "good" inference under incompatible imputation models (Van Buuren et al., 2006; Van Buuren, 2007).

We will not attempt to prove convergence or consistency properties for the proposed algorithm beyond what exists in the chained equations literature. Instead, we will use simulation and some minor analytical exploration to identify settings that may be particularly susceptible to concerns about convergence. In particular, identifiability concerns related to the missingness model have implications on the convergence of the algorithm. When parameters are not identifiable (in terms of the observed data likelihood having a unique maximizer), we may not expect the imputation algorithm to converge properly. Even when the parameters are all identifiable, we may run into numerical issues if the observed data likelihood is nearly flat. These issues appear to be of greater concern for outcome missingness. We note that in our experience, even when we have numerical convergence issues for ϕ (missingness model) and ω (model for $L|X$), the draws for θ (model for $Y|X, L$) may still converge to reasonable values. In such cases, the identifiability-related numerical problems may not strongly impact the draws for the primary parameter of interest, θ. It is important to monitor the convergence of all model parameters, and we may still be able to make inference about θ in the presence of some mild identifiability-related convergence issues for ϕ. We explore identifiability-related convergence issues further in **Section 3.5** and **Appendices B and C**.

3.5 Simulation Study

In this section, we present a simulation study with five parts. In the first three parts, we evaluate how the proposed algorithm performs in terms of bias, empirical variance, and coverage for outcome model parameters in linear mixed models (Simulation 1), CPH cure models (Simulation 2), and normal mixture models (Simulation 3). In Simulation 4, we explore convergence under a variety of modeling scenarios. In Simulation 5, we explore the impact of different types of final analysis on efficiency. Unless otherwise specified, imputations are drawn using kernels (3.6)-(3.8) rather than regression model approximations.

3.5.1 Simulation 1: Linear Mixed Model with Random Intercept

We consider data simulated under a linear mixed model with a random intercept. Each dataset contains two binary covariates, X_1 and X_2. X_1 takes the value 1 with a probability of 0.5, and X_2 is generated using $\text{logit}(P(X_2 = 1|X_1)) = 0.5X_1$. We draw random intercept $b_i \sim N(0,1)$ for each individual and then generate Y for each individual at each of three time-points using the model

$$Y_{ij} = \beta_{Intercept} + \beta_{X_1}X_{i1} + \beta_{X_2}X_{i2} + \beta_{Time}j + b_i + e_{ij}, \qquad j = 1,2,3$$

with independent $N(0,1)$ errors and with $\beta_{Intercept} = \beta_{X_1} = \beta_{X_2} = 0.5$ and $\beta_{Time} = 0.2$. In this simulation setting, $Y = (Y_1, Y_2, Y_3)$, $X = (X_1, X_2)$, and $L = b$. We impose $\sim 50\%$ missingness in X_2 using each of the following mechanisms:

(A) MAR with $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = -1.1 + Y_1$

(B) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = 0.5b$

(C) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = 0.1 + 1.2b$.

Mechanism (A) is MAR dependent on Y_1, the baseline value of Y. Mechanism (B) is LMAR with a moderate dependence between missingness and b, and mechanism (C) is LMAR with a strong dependence on b. Y and X_1 are fully observed.

We then impute values of X_2 and b using methods discussed in this chapter under various working models. When we impute under a LMAR working model, we model the covariate missingness indicator R_i^D using a logistic regression with different functions of b, X_1, and Y as predictors. For each simulated dataset, we create 10 imputed datasets. We then fit a linear mixed model to each of the imputed datasets and use Rubin's rules to obtain a single set of parameter estimates and their corresponding variances for each simulation. We then compute the bias, empirical variance, and coverage rates across the 500 simulations. We note that the APPROX simulations involve imputation of X_2 conditional on X_1, L and Y using a logistic regression form rather than using kernel (3.8), so the imputation distributions for X_2 and L in this case do not correspond to a coherent joint distribution.

Table 3.1 shows the simulation results. Complete case analysis produced biased parameter estimates in all three underlying missingness mechanisms considered. Under MAR missingness mechanism (A), the MAR-based imputation approach produces unbiased parameter estimates. LMAR imputation under mechanism (A) produces biased parameter estimates when an incorrect working missingness model is used. When the working model contains the underlying missingness model, however, the LMAR method results in essentially unbiased parameter estimates. Under mechanism (A), the MAR-based imputation approach and the LMAR imputation approach with the correct working model result in very similar coverage and relative variance. APPROX Imputation using a logistic regression model for imputing X_2 had similar performance to imputation using kernel (3.8).

Under mechanism (B), all imputation approaches produce essentially unbiased parameter estimates. The LMAR approaches, however, result in small increases in coverage and reductions in variance compared to the MAR imputation approach. Under mechanism (C), the MAR-based imputation approach produces noticeable bias in estimating the mixed model intercept and parameter associated with the imputed covariate. We see corresponding reductions in coverage for these parameters. In contrast, the LMAR-based imputation approaches produce unbiased parameter estimates. For mechanisms (B) and (C), the working model that uses $\mathbb{I}(b > 0)$ instead of b in the working model still shows good performance despite the fact that the working model does not contain the true model. We do not see evidence of problems arising from lack of identifiability or lack of convergence under any of the working models considered here. MAR-based imputation using a logistic regression model for imputing X_2 resulted in slightly greater bias than

MAR imputation using kernel (3.8).

Table 3.1: Linear Mixed Model Estimates using Proposed Imputation Methods

		Parameters			
	Contains	Intercept	X_1	X_2	Time
Method	Truth$^{\#}$	Bias (Var) CI†	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI
Full Data	-	0 (1.2) 95	0 (1.0) 94	-1 (1.0) 94	0 (0.10) 95
		Missingness Dependent on Y_{i1}, Independent of b_i (Mechanism A)			
Complete Case	-	-78 (1.9) 0	-8 (1.8) 91	-9 (1.7) 90	18 (0.19) 1
MAR Imputation	Y	0 (1.8) 93	0 (1.1) 95	-1 (2.5) 95	0 (0.10) 95
LMAR Imputation: b*	N	6 (1.5) 91	1 (1.0) 96	-9 (1.8) 93	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	N	6 (1.5) 92	1 (1.0) 95	-9 (1.9) 93	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.8) 93	0 (1.1) 95	-1 (2.6) 93	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	Y	0 (1.8) 94	0 (1.0) 95	-1 (2.5) 93	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.9) 93	0 (1.0) 96	-1 (2.6) 93	0 (0.10) 95
MAR APPROX Imputation	Y	0 (1.9) 94	0 (1.1) 95	0 (2.8) 93	0 (0.10) 95
LMAR APPROX Imputation: b	N	6 (1.6) 92	1 (1.1) 95	-9 (2.1) 94	0 (0.10) 95
		Missingness Moderately Dependent on b_i (Mechanism B)			
Complete Case	-	-23 (2.5) 96	0 (2.1) 93	0 (1.9) 95	0 (0.19) 94
MAR Imputation	N	-1 (1.7) 94	0 (1.1) 95	1 (2.2) 93	0 (0.10) 95
LMAR Imputation: b	Y	0 (1.5) 95	0 (1.1) 95	0 (2.0) 94	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	Y	1 (1.5) 96	0 (1.1) 94	0 (1.9) 94	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.6) 95	0 (1.1) 95	0 (2.0) 94	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	N	0 (1.5) 96	0 (1.1) 94	0 (2.0) 94	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.5) 96	0 (1.1) 94	0 (2.0) 94	0 (0.10) 95
		Missingness Strongly Dependent on b_i (Mechanism C)			
Complete Case	-	-47 (2.6) 13	0 (1.8) 95	-1 (1.9) 94	0 (0.22) 94
MAR Imputation	N	-6 (1.9) 91	0 (1.1) 95	6 (2.6) 90	0 (0.10) 95
LMAR Imputation: b	Y	0 (1.4) 96	0 (1.1) 95	-1 (2.0) 95	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	Y	0 (1.5) 95	0 (1.1) 95	-1 (2.1) 95	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.5) 96	0 (1.1) 94	-1 (2.0) 96	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	N	0 (1.5) 95	0 (1.1) 95	0 (2.1) 95	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.5) 95	0 (1.1) 95	0 (2.0) 95	0 (0.10) 95
MAR APPROX Imputation	N	-7 (2.0) 91	-1 (1.1) 95	8 (2.7) 89	0 (0.10) 95
LMAR APPROX Imputation: b	Y	0 (1.5) 96	0 (1.1) 95	0 (2.1) 97	0 (0.10) 95

*Variables after colon represent linear predictors in working model for R_i^D

† All values in table multiplied by 100. CI indicates coverage of 95% confidence intervals. Var indicates empirical variance.

# Indicates whether working missingness model contains true model.

APPROX: Imputation of X_2 uses a logistic regression with predictors X_1, b, Y_1, Y_2, Y_3 (instead of kernel (3.8))

3.5.2 Simulation 2: Cox Proportional Hazards Mixture Cure Model

We simulate 500 datasets of 500 subjects under a CPH mixture cure model. Covariates X_1 and X_2 are simulated as in Simulation 1. We simulate an underlying cure status using the relation $\text{logit}(P(\text{Not Cured}|X_{i1}, X_{i2})) = 0.5 + 0.5X_{i1} + 0.5X_{i2}$. This results in an average cure rate of 26%. For the non-cured group (G=1), we simulate an event time using the hazard function $\lambda(t) = 0.0005t^{0.3}e^{0.5X_{i1}+0.5X_{i2}}$. For cured subjects (G=0), the event time is taken to be infinity. We generate censoring times using the relation $\lambda_C(t) = 0.00015t^{0.5}$ for the first 400 subjects and impose administrative censoring at 3000 for the remaining 100 subjects. The observed event/censoring time T_i is taken as the minimum of the censoring and event time, and δ_i represents the event indicator. In this simulation setting, $Y = (T, \delta)$, $X = (X_1, X_2)$, and $L = G$. For the estimation, we assume subjects with T_i greater than a late cut-point are cured. We choose a cut-point of 50 as the Kaplan-Meier plots demonstrate a clear plateau by that point. We impose $\sim$50% missingness in X_2 using each of the following mechanisms:

(A) MCAR with missingness probability of 0.5

(B) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, G, T, \delta)) = -0.2 + 0.3G$

(C) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, G, T, \delta)) = -0.9 + 1.2G$.

Mechanism (A) is MCAR, mechanism (B) is LMAR with a moderate dependence on cure status (G), and mechanism (C) is LMAR with a strong dependence on cure status.

We assume a Weibull baseline hazard in the non-cured group for imputation. For each imputed dataset, we fit a CPH cure model, which consists of a logistic regression for the probability of not being cured and a Cox regression for the hazard of events in the not cured group. We fit this model using the package *smcure* in R (Cai et al., 2012). Variances were estimated using 100 bootstrap samples.

Table 3.2 shows the simulation results for the Cox proportional hazards mixture cure model. As expected, complete case analysis is essentially unbiased under covariate missingness mechanism (A) (MCAR), but the imputation-based methods are more efficient than the complete case analysis. When missingness depends on the underlying cure status, however, complete case analysis is biased. We see comparatively little bias in the imputation-based estimates across missingness mechanisms and imputation models using kernel (3.8). APPROX Imputation using a logistic regression model for imputing X_2 resulted in increased bias in all scenarios. For missingness mechanisms (A) and (B) and using kernel (3.8) for imputation, we see very little difference between the MAR and LMAR imputation approaches in terms of bias, coverage, and relative variance. APPROX imputation under LMAR resulted in slightly larger variances than APPROX imputation under MAR.

In mechanism (C) (when missingness depends strongly on cure status), we can begin to see a difference between the MAR and LMAR imputation methods using kernel (3.8) in terms of bias, but this difference is still small. Larger bias differences between MAR-based and LMAR-based imputation can be seen when covariate imputation uses a logistic regression instead of kernel (3.8). The LMAR imputation approaches using kernel (3.8) (which differ only in terms of the working missingness model) produce essentially unbiased estimates for all model parameters. LMAR imputation using G, X_1, and $G \times X_1$ in the working model resulted in some numerical convergence issues for several of the simulations (15 simulations failed), which may indicate issues with model identifiability (possibly due to collinearity). We included only the converging simulations (485 of them) in **Table 3.2**.

Table 3.2: CPH Cure Model Estimates using Proposed Imputation Methods

		- - - - - - - - Logistic Regression	- - - - - - - -	- - - - - - - -	- - - - Cox Regression	- - - -
	Contains	Intercept	X_1	X_2	X_1	X_2
Method	Truth#	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI
Full Data	-	1 (6.5) 94	1 (7.9) 95	0 (8.4) 95	0 (2.0) 95	0 (2.3) 95
		MCAR Missingness Independent of G_i (Mechanism A)				
Complete Case	-	2 (12.7) 97	1 (14.9) 97	1 (18.5) 96	1 (4.3) 95	0 (5.1) 94
MAR Imputation	Y	3 (9.1) 94	1 (8.4) 96	0 (18.0) 95	0 (2.1) 96	0 (4.8) 93
LMAR Imputation: G*	Y	3 (9.4) 94	1 (8.3) 96	0 (19.5) 95	0 (2.1) 96	0 (4.9) 94
LMAR Imputation: G, X_1, Y, δ	Y	3 (9.3) 94	1 (8.3) 96	0 (18.8) 95	0 (2.1) 96	0 (4.8) 95
LMAR Imputation: $G, X_1, G \times X_1$	Y	3 (9.5) 94	1 (8.4) 96	0 (18.9) 95	0 (2.1) 96	0 (4.8) 95
MAR APPROX Imputation	Y	6 (9.4) 93	1 (8.2) 95	-6 (19.6) 91	0 (2.1) 96	-1 (4.4) 93
LMAR APPROX Imputation: G	Y	6 (9.6) 93	1 (8.4) 96	-6 (21.1) 91	0 (2.1) 96	-2 (4.5) 91
		Missingness Moderately Dependent on G_i (Mechanism B)				
Complete Case	-	-13 (12.1) 93	3 (16.0) 97	0 (15.4) 97	0 (4.8) 95	1 (5.4) 93
MAR Imputation	N	3 (8.8) 95	0 (8.5) 96	0 (16.9) 96	0 (2.2) 96	0 (4.8) 93
LMAR Imputation: G	Y	3 (8.9) 96	0 (8.5) 96	0 (16.7) 96	0 (2.2) 95	0 (4.8) 94
LMAR Imputation: G, X_1, Y, δ	Y	3 (8.8) 94	1 (8.5) 96	0 (16.0) 96	0 (2.2) 95	1 (4.7) 94
LMAR Imputation: $G, X_1, G \times X_1$	Y	3 (9.0) 94	0 (8.6) 96	0 (16.3) 95	0 (2.2) 95	0 (4.8) 93
		Missingness Strongly Dependent on G_i (Mechanism C)				
Complete Case	-	-50 (9.9) 62	2 (12.9) 97	0 (12.2) 96	1 (5.4) 95	0 (5.8) 95
MAR Imputation	N	4 (8.3) 96	1 (8.6) 96	-1 (15.2) 95	0 (2.2) 96	0 (5.1) 94
LMAR Imputation: G	Y	2 (7.9) 95	0 (8.5) 97	0 (13.3) 96	0 (2.3) 96	0 (5.5) 93
LMAR Imputation: G, X_1, Y, δ	Y	2 (7.8) 95	1 (8.4) 97	1 (13.0) 96	0 (2.2) 95	0 (5.4) 93
LMAR Imputation: $G, X_1, G \times X_1$	Y	1 (7.5) 93	1 (8.1) 94	0 (13.0) 93	0 (2.2) 92	0 (5.4) 91
MAR APPROX Imputation	N	8 (8.4) 93	1 (8.5) 97	-12 (16.1) 91	0 (2.3) 96	-1 (5.3) 90
LMAR APPROX Imputation: G	Y	6 (8.0) 94	1 (8.4) 97	-9 (14.2) 93	1 (2.3) 96	-1 (5.2) 91

*Variables after colon represent linear predictors in working model for R_i^D

† All values multiplied by 100. CI indicates coverage of 95% confidence intervals. Var indicates empirical variance.

# Indicates whether working missingness model contains true model.

APPROX: Imputation of X_2 uses a logistic regression with predictors $X_1, G, G \times H_0(Y), G \times H_0(Y) \times X_1$.

3.5.3 Simulation 3: Mixture of Normals

We simulate 500 datasets of 500 subjects under a normal mixture model with two binary covariates and two latent classes. Covariates X_1 and X_2 are simulated as in Simulation 1. We generate the mixing variable C_i with $P(C_i = 1) = 0.62$ for each individual. We draw $N(0, 1)$ errors e_i and then generate Y using the model $Y_i = 0.5 + 0.5X_{i1} + 0.5X_{i2} + e_i$ if $C_i = 1$ and $Y_i = 2 + 3X_{i1} + 2X_{i2} + e_i$ if $C_i = 0$. In this simulation setting, $X = (X_1, X_2)$ and $L = C$. We then impose missingness in X_2 using each of the following mechanisms:

(A) MAR with $P(X_2 \text{ missing}|X_1, C, Y) = -0.5 + 0.2Y$

(B) LMAR with $P(X_2 \text{ missing}|X_1, C, Y) = -0.3 + 0.5C$

(C) LMAR with $P(X_2 \text{ missing}|X_1, C, Y) = -1.1 + 1.7C$.

Mechanism (A) is MAR dependent on Y. Mechanism (B) is LMAR with a moderate dependence on the latent class variable (C), and mechanism (C) is LMAR with a strong dependence on the latent class.

For each imputed dataset, we fit a latent class model (with two classes) using the package 'flexmix' in R to estimate θ through an EM algorithm (Leisch, 2004). The package 'flexmix' estimates the variance for $\hat{\theta}$ for each dataset by fitting a GLM weighted by estimated class membership probabilities for each individual. When parameters are drawn using latent class modeling, we may not be able to determine which value of C belongs to which subclass identified by the latent class modeling. In other words, we may not be able to differentiate which subset of θ belongs to which value of C. We can circumvent this issue by placing an additional assumption to differentiate between classes. We impose an identifying restriction that defines class $C_i = 1$ to be the cluster determined by the latent class modeling with a smaller intercept value. We note that the two clusters are well-separated in this example. We predict that we may encounter greater identifiability issues (in differentiating the clusters) when the clusters have parameters that are very close together.

Table 3.3 shows the simulation results for a mixture of normal distributions. Complete case analysis results in biased parameter values for mechanism (A) and mild or no bias for mechanisms (B) and (C). For mechanism (A), the MAR-based imputation approach produces essentially unbiased parameter estimates. The LMAR imputation approaches with working missingness models containing the true missingness model also produce very small bias. Mild increases in bias can be seen for the LMAR imputation approach using an incorrect working model. Compared to the MAR approach, the LMAR approach using the correct working model resulted in similar or slightly larger variances for all parameter estimates.

For mechanism (B), little bias can be seen across all of the imputation approaches. Similar coverage rates can be seen across imputation approaches. In this example, we see slightly smaller variances for the LMAR approaches with the more complicated working models. Under mechanism (C), we see increases in bias and small decreases in coverage for estimating mixture model parameters using the MAR-based imputation method (either using kernel (3.8) or logistic regression for imputing X_2). The LMAR-based imputation method using only C in the working missingness model produces essentially unbiased parameter estimates for all parameters. Compared to the approaches using the more complicated working model, the simpler LMAR approach using kernel (3.8) results in smaller variances for estimating model parameters.

Table 3.3: Mixture of Normals Model Estimates using Proposed Imputation Methods

		- - - - - - - - - - - - - C = 1 - - - - - - - - - - - - -			- - - - - - - - - - - - - C = 0 - - - - - - - - - - - - -		
	Contains	Intercept	X_1	X_2	Intercept	X_1	X_2
Method	Truth#	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI
Full Data	-	0 (1.6) 95	0 (1.9) 92	0 (1.5) 97	0 (3.6) 94	0 (3.3) 94	0 (3.4) 94
		Missingness Dependent on Y_i, Independent of C_i (Mechanism A)					
Complete Case	-	2 (2.8) 94	-7 (3.5) 92	-4 (2.7) 95	3 (12.6) 93	-12 (10.3) 90	-4 (10.3) 93
MAR Imputation	Y	1 (1.9) 95	0 (4.3) 94	1 (4.6) 95	-1 (6.1) 94	-2 (7.8) 94	0 (5.5) 95
LMAR Imputation: C*	N	3 (1.9) 94	1 (4.6) 94	-2 (4.5) 95	-3 (6.2) 94	-4 (7.8) 94	2 (5.6) 94
LMAR Imputation: C, Y, X_1	Y	1 (2.0) 94	0 (4.3) 93	1 (4.6) 95	-1 (6.4) 95	-2 (7.7) 94	0 (5.3) 95
LMAR Imputation: $C, Y, C \times Y$	Y	1 (2.0) 94	1 (4.6) 93	1 (4.6) 96	-2 (6.2) 94	-2 (8.1) 94	0 (6.0) 95
LMAR Imputation: $C, Y, C \times Y, X_1$	Y	1 (2.0) 94	0 (4.5) 93	1 (4.9) 95	-2 (6.3) 94	-2 (8.0) 93	0 (6.0) 94
MAR APPROX Imputation	Y	1 (2.0) 94	0 (2.9) 93	0 (3.2) 92	0 (6.3) 92	0 (6.4) 93	-3 (5.5) 94
LMAR APPROX Imputation: C	N	3 (1.9) 93	0 (3.4) 93	-3 (3.8) 93	-1 (6.3) 93	-2 (6.9) 92	0 (5.3) 93
		Missingness Moderately Dependent on C_i (Mechanism B)					
Complete Case	-	1 (5.4) 93	-1 (5.4) 93	0 (4.1) 95	-1 (8.6) 91	0 (7.9) 92	0 (6.9) 91
MAR Imputation	N	1 (2.3) 94	0 (2.7) 94	-1 (4.0) 93	-2 (5.8) 94	-2 (4.8) 95	2 (4.7) 93
LMAR Imputation: C	Y	0 (2.3) 94	1 (4.7) 94	1 (5.1) 93	-1 (5.6) 95	-2 (6.4) 95	0 (5.3) 94
LMAR Imputation: C, Y, X_1	Y	0 (2.4) 95	1 (5.1) 94	1 (5.5) 93	-1 (5.8) 94	-2 (7.1) 95	0 (5.7) 93
LMAR Imputation: $C, Y, C \times Y$	Y	0 (2.4) 94	1 (3.8) 94	1 (4.9) 92	-1 (5.7) 93	-2 (5.7) 94	0 (5.0) 94
LMAR Imputation: $C, Y, C \times Y, X_1$	Y	1 (2.4) 95	1 (3.6) 95	1 (4.7) 92	-1 (5.8) 94	-2 (5.6) 94	0 (5.0) 94
		Missingness Strongly Dependent on C_i (Mechanism C)					
Complete Case	-	3 (8.2) 93	-2 (7.6) 92	-1 (6.4) 93	-2 (5.8) 92	1 (6.1) 92	1 (5.3) 92
MAR Imputation	N	5 (2.5) 93	2 (7.2) 93	-7 (7.4) 91	-5 (4.9) 95	-4 (7.6) 94	5 (7.0) 92
LMAR Imputation: C	Y	1 (2.8) 94	1 (4.6) 94	0 (6.4) 92	-1 (4.5) 96	-2 (5.9) 95	0 (5.2) 94
LMAR Imputation: C, Y, X_1	Y	1 (2.9) 94	2 (5.9) 94	0 (7.2) 92	-2 (4.7) 96	-3 (6.9) 95	0 (5.9) 94
LMAR Imputation: $C, Y, C \times Y$	Y	1 (2.8) 94	1 (5.4) 94	0 (7.3) 92	-2 (4.5) 96	-2 (6.7) 95	0 (5.8) 93
LMAR Imputation: $C, Y, C \times Y, X_1$	Y	1 (2.8) 95	2 (6.6) 94	0 (7.8) 92	-2 (4.8) 94	-3 (8.0) 94	0 (6.1) 92
MAR APPROX Imputation	N	5 (2.5) 92	2 (5.7) 93	-7 (6.6) 88	-6 (4.9) 94	-3 (6.8) 94	6 (6.2) 91
LMAR APPROX Imputation: C	Y	1 (2.8) 92	2 (6.6) 94	0 (7.3) 87	-1 (4.4) 95	-3 (7.3) 95	0 (6.4) 94

*Variables after colon represent linear predictors in working model for R_i^D

† All values in table multiplied by 100. CI indicates coverage of 95% confidence intervals.Var indicates empirical variance.

# Indicates whether working missingness model contains true model.

APPROX: Imputation of X_2 uses a logistic regression with predictors $C, X_1, X_1 \times C, Y, Y \times C$ (instead of kernel (3.8))

3.5.4 Simulation 4: Exploring Identifiability and Convergence

One criticism of the selection model factorization in (3.3) is that it is often difficult to determine whether the parameters of the working missingness model are identifiable (Little, 2009). By "identifiable," we mean that the observed data likelihood has a unique maximizer. Even if the model parameters are technically identifiable, one additional concern is that the likelihood surface near the maximizer may be nearly flat. These identifiability concerns can lead to issues with model fitting and convergence of the imputation algorithm. In order to better understand possible identifiability-related convergence issues, we perform a set of simulations evaluating convergence of the imputation algorithm under a variety of modeling scenarios.

We simulate 500 complete datasets under a linear mixed model, cure model, and mixture of normals as in Simulations 1-3. We impose $\sim 50\%$ covariate or outcome missingness (but not both) under a variety of missingness models.

For covariate missingness, we generate MAR and LMAR missingness using missingness mechanisms (A) and (C) from Simulations 1-3. For both the linear mixed model and mixture of normals model, we generate outcome missingness under MCAR and LMAR using mechanism (C) from Simulations 1 and 3 applied to the outcome instead of the covariate. We also impose LMAR outcome missingness for the mixture of normals model using the relation $\text{logit}(P(Y \text{ missing}|X, C)) = -1.1 + 0.5X_1 - 0.5X_2 + 1.7C$. This results in $\sim 50\%$ outcome or covariate missingness in each scenario.

For each outcome model parameter, we estimate the fraction of missing information as described in (Little and Rubin, 2002). We also calculate the Gelman-Rubin convergence statistic (the potential scale reduction factor) for the outcome and missingness model parameter draws across imputation streams. The Gelman-Rubin statistic is a measure of the relative between and within-chain variance, and values less than 1.1 generally indicate satisfactory convergence (Gelman and Rubin, 1992). We also calculate a multivariate version of the Gelman-Rubin statistic to evaluate convergence overall across different model parameters (Brooks and Gelman, 1998).

Table 3.4 shows the simulation results. Under covariate missingness, the fractions of missing information tend to be generally small, particularly for parameters related to X_1, the fully-observed covariate. We see larger estimates for the fraction of missing information when we impose similar rates of missingness in the outcome. Additionally, we see good Gelman-Rubin convergence properties under covariate missingness and MAR outcome missingness. Under LMAR outcome missingness, the outcome model parameters appear to converge, but the parameters in the missingness model (in particular, the parameter attached to the latent variable) show some evidence of convergence problems. The drawn values of the outcome model parameters appear reasonable (with small or no bias) even when the missingness model parameters do not converge, but this may not be true in general. When we fix the value of the parameter related to the latent variable in the missingness, we see a large improvement in the convergence properties of the imputation algorithm.

Table 3.4: Fraction of Missing Information and Convergence Properties†

	Fraction of Missing Information						Gelman-Rubin Statistic								Overall
	Outcome Model Parameters*						Outcome Model Parameters*						$\phi_0^{\dagger}$	$\phi_1^{\ddagger}$	Gelman-Rubin
				Covariate Missingness											
Linear Mixed Model, MAR	0.27	0.07	0.54	0			1.01	1.00	1.02	1.01			-	-	1.03
Linear Mixed Model, LMAR: $b^{\#}$	0.24	0.06	0.52	0			1.01	1.00	1.01	1.01			1.00	1.03	1.04
Cure Model, MCAR	0.20	0.03	0.46	0.06	0.45		1.01	1.00	1.02	1.00	1.01		-	-	1.04
Cure Model, LMAR: G	0.16	0.02	0.36	0.07	0.52		1.01	1.00	1.01	1.00	1.02		1.00	1.00	1.04
Mixture of Normals, MAR	0.17	0.06	0.39	0.39	0.32	0.32	1.00	1.02	1.01	1.00	1.01	1.01	-	-	1.02
Mixture of Normals, LMAR: C	0.33	0.10	0.62	0.22	0.14	0.14	1.00	1.02	1.01	1.00	1.01	1.00	1.00	1.01	1.02
				Outcome Missingness											
Linear Mixed Model, MCAR	0.18	0.22	0.21	0.49			1.00	1.01	1.01	1.03			-	-	1.05
Linear Mixed Model, LMAR: b	0.19	0.22	0.20	0.57			1.01	1.01	1.01	1.07			1.00	1.10	1.14
Linear Mixed Model, LMAR: b	0.18	0.21	0.22	0.52			1.01	1.01	1.01	1.04			1.01	FIXED	1.06
Mixture of Normals, MCAR	0.52	0.54	0.53	0.54	0.54	0.54	1.02	1.03	1.02	1.02	1.02	1.01	-	-	1.06
Mixture of Normals, LMAR: C	0.57	0.57	0.57	0.46	0.46	0.47	1.02	1.02	1.01	1.01	1.03	1.01	1.55	1.65	1.66
Mixture of Normals, LMAR: C, X	0.58	0.62	0.57	0.47	0.48	0.47	1.02	1.02	1.02	1.01	1.03	1.01	1.38	1.64	1.65
Mixture of Normals, LMAR: C	0.68	0.68	0.66	0.35	0.35	0.35	1.04	1.04	1.02	1.01	1.03	1.01	1.05	FIXED	1.13
Mixture of Normals, LMAR: C, X	0.67	0.71	0.65	0.35	0.37	0.34	1.04	1.03	1.02	1.01	1.03	1.01	1.01	FIXED	1.08

† Imputations drawn using kernels (3.6) - (3.7)

*For each model, these are the parameters from the outcome model (same as **Tables 3.1 - 3.3**):

— Linear mixed model: intercept, X_1, X_2, and time

— Cure model: intercept, X_1, and X_2 in the logistic regression and X_1 and X_2 in the Cox regression

— Mixture of Normals: intercept, X_1, and X_2 for the $C = 1$ and $C = 0$ classes respectively

‡ ϕ_0 is the intercept in the missingness model. ϕ_1 is the parameter for the latent variable.

# Notation: True and working missingness models depend on variables after colon

3.5.5 Simulation 5: Comparison of Final Analysis with and without Imputed L

After imputation, we have several choices as to what combination of the imputed L and D we want to include in the final analysis. We first suppose that we will perform our final analysis ignoring the contribution of R. When both imputed D and L are included in the final analysis, R is ignorable. In *Lemma 4*, we show that R is also ignorable if only imputed D is included in the final analysis when missingness is MAR. When missingness is LMAR, we show in *Lemma 5* that final analysis using only the imputed D and ignoring R will be valid but not fully efficient. In this section, we want to briefly explore the practical impact of including or excluding the imputed values of L (assuming we are ignoring R) in the final analysis through simulation.

We generate simulated data under a linear mixed model, mixture of normals model, and Cox proportional hazards model as described for Simulations 1-3. We impose either MAR or LMAR (Strong Dependence) missingness in X_2 as in Simulations 1-3 and impute using a working missingness model with the correct structure (either MAR or LMAR dependent only on the latent variable) and kernels (3.6) - (3.8). After imputation, we perform the final analysis using the imputed values for X_2 and either ignoring or using the imputed values for the latent variable (and in both cases ignoring R). Additionally, in the course of our simulations, we observed that some simulations under the mixture of normals model had estimated variances that were very large when we used the imputed latent variable in the final model fit. This may be an indicator of inadequate convergence of the model fit. Therefore, we present the mixture of normals results 1) for all 500 simulations and 2) restricting to simulations in which the estimated variances were all less than 0.2 (20 in the scale presented in the table). This issue did not arise for the linear mixed model simulations. In **Tables 3.1-3.3**, we perform all final analyses ignoring the imputed latent variable and without restricting to simulations that have variance < 0.2, and the corresponding rows in this table are identical to the results in **Tables 3.1-3.3**.

Table 3.5 shows the simulation results. We first consider the results for the mixture of normals model. We first notice that analyses using the imputed latent variable in the final analysis result in substantial bias when we include all simulations in our estimation of bias. This is the result of just a few simulations with parameter estimates far from the true value. This suggests some instability or lack of convergence in the model fitting. However, when we restrict our focus to simulations that appear to have convergence (reasonable standard errors), we see that final analyses including and excluding the imputed latent variable perform similarly well. For some simulation settings, the variance estimates using C are slightly larger, and the reverse is true for other simulations, so there is not a clear trend in efficiency including or excluding the latent variable in the final analysis in these simulations.

Although not included in our results, it is worth mentioning that analysis including and ignoring the imputed L may be associated with different fractions of missing information, which could have implications on the number of imputations needed for good inference. Let $\bar{U}$ represent the average of the variance estimators for parameter θ across the m imputed datasets and B represent the sample variance of the estimates of θ across the m imputed datasets. Then, we can express the relative increase in variation due to the missing data (r) and the fraction of missing information (λ) as (Schafer, 1999):

$$r = \frac{(1+\frac{1}{m})B}{\bar{U}} \qquad \lambda = \frac{r}{1+r}$$

The relative efficiency of an estimate θ based on m imputations compared to the estimate based of in infinite number of imputations is:

$$RE = \frac{1}{1+\frac{\lambda}{m}}$$

We may expect an analysis that conditions on the imputed L in the final analysis to have larger relative between imputation variance vs. within imputation variance (r) compared to an analysis that does not condition on L in the final analysis for some parameters. This is because, when we include L in the final analysis, each fit treats the imputed L as known, resulting in substantially reduced “within imputation” standard error estimates for some parameters. This leads to larger values for the fraction of missing information, λ, for the same value of m when we include L in the final analysis compared to an analysis that ignores imputed L. In simulations (not shown), a final analysis using L did result in larger fractions of missing information compared to an analysis ignoring imputed L in the random intercept linear mixed model setting. We note that in practice this may translate into only a very small difference in relative efficiency between the two methods of analysis. However, several authors have noted practical issues regarding estimation of p-values and confidence intervals when a small number of imputations are used and the fraction of missing information is moderate to large (e.g. White and Royston, 2011; Bodner, 2008). Therefore, we may prefer to perform our final analysis using only the imputed D in the final analysis in an attempt to reduce the potential negative impact of larger fractions of missing information.

Table 3.5: Bias and Variance of Parameter Estimates under Different Final Analyses

Linear Mixed Model

Model$^{\#}$	Analysis	Intercept Bias (Var)†	X_1 Bias (Var)	X_2 Bias (Var)	Time Bias (Var)	SIMS
MAR	Ignoring b	0 (1.81)	0 (1.10)	-1 (2.58)	0 (0.1055)	500
MAR	Using b	0 (1.84)	0 (1.11)	-1 (2.59)	0 (0.1055)	500
LMAR	Ignoring b	0 (1.49)	0 (1.15)	-1 (2.03)	0 (0.1055)	500
LMAR	Using b	0 (1.50)	0 (1.08)	-1 (2.05)	0 (0.1055)	500

Mixture of Normals

		- - - - - - - - - - C = 1- - - - - - - - - -			- - - - - - - - - C = 0 - - - - - - - - -			SIMS
Model$^{\#}$	Analysis	Intercept Bias (Var)	X_1 Bias (Var)	X_2 Bias (Var)	Intercept Bias (Var)	X_1 Bias (Var)	X_2 Bias (Var)	
Variance Unrestricted								
MAR	Ignoring C	1 (1.98)	0 (4.35)	1 (4.64)	-1 (6.18)	-2 (7.80)	0 (5.53)	500
MAR	Using C	2 (2.22)	6 (7.55)	5 (6.38)	-3 (7.09)	-8 (10.38)	-3 (6.49)	500
LMAR	Ignoring C	1 (2.89)	1 (4.63)	0 (6.42)	-1 (4.56)	-2 (5.91)	0 (5.29)	500
LMAR	Using C	1 (3.28)	9 (9.71)	5 (7.76)	-3 (5.18)	-9 (10.03)	-4 (7.18)	500
Variance Restricted*								
MAR	Ignoring C	1 (2.04)	-1 (2.09)	0 (2.88)	-1 (6.06)	0 (5.34)	0 (5.02)	483
MAR	Using C	1 (2.07)	-1 (1.95)	0 (2.94)	0 (5.73)	-1 (5.14)	0 (4.25)	404
LMAR	Ignoring C	0 (2.79)	0 (2.07)	0 (5.39)	0 (4.28)	0 (3.77)	1 (3.94)	477
LMAR	Using C	0 (2.95)	0 (2.11)	0 (5.44)	0 (4.35)	-1 (3.75)	0 (3.95)	418

† All values in table multiplied by 100
# Indicates true and working missingness model
* Ignoring simulations in which the estimated variance was greater than 0.2 (20 in the scale of this table) for at least one parameter.

3.6 Application to Head and Neck Cancer Data

In practice, the missingness mechanism is rarely known, and we may want to explore the sensitivity of the model inference to assumptions about the missingness. In this section, we evaluate the impact of missingness assumptions on inference for a particular dataset.

We consider data from a cohort study of N=1226 patients treated for head and neck squamous cell carcinoma (HNSCC). This study was conducted by the University of Michigan Head and Neck Specialized Program of Research Excellence (SPORE) and followed patients who were treated at the University of Michigan Cancer Center for HNSCC between Nov. 2003 and July 2013. Details about this study can be found in Duffy et al. (2008) and Peterson et al. (2016). After treatment, patients were followed for recurrence. Covariate information was also collected at baseline. We are interested in studying the association between covariates and the time to HNSCC recurrence after treatment. We model the time to HNSCC recurrence using a Cox proportional hazards cure model.

HPV status was unavailable for 55.8% of the subjects, and small amounts of missingness was present in other study variables. **Chapter II** explores imputation for dealing with the missing covariate data for this study under MAR assumptions. However, an induced LMAR association between missingness in HPV status and cure status (denoted G) could occur if HPV missingness is related to an unmeasured variable that is also related to the cure probability. For example, a more experienced doctor may be more likely to recommend HPV testing and to have cured patients. Also, HPV missingness rate could be related to calendar time, which may be associated with the cure rate.

We are interested in comparing model inference assuming MAR to inference obtained when missingness in HPV is assumed to be LMAR. We assume missingness in all other variables is MAR. We consider three working assumptions for HPV status missingness: (A) MAR, (B) missingness dependent only on cure status, and (C) missingness dependent on cure status, age at diagnosis, cancer site, and enrollment year. Assumptions (B) and (C) are modeled using logistic regression.

We apply our proposed methods to impute the missing data. In this setting, G is the partially latent cure status, Y is the censored event time data (time and indicator), and X is the set of covariates. Here, the model $Y|G = 1, X$ is a Cox regression and the model for $G|X$ is a logistic regression. We impute cure status G using (3.6). As suggested in **Chapter II**, we impute missing values of each p^{th} covariate $X^{(p)}$ using a standard regression model with $X^{(-p)}$, G, $G \times \hat{H}_0(T)$, and $G \times \hat{H}_0(T) \times X^{(-p)}$ as predictors. Here, $\hat{H}_0(T)$ is an estimate of the cumulative baseline hazard of having an event in the non-cured group. Like in **Chapter II**, we will draw values for the regression model's parameter without conditioning on the imputed $X^{(p)}$ (as is done in usual chained equations). Variables included in $X^{(-p)}$ for the imputation include log-transformed number of sexual partners, PNI, comorbidities, smoking habits, alcohol use, age, cancer site, cancer stage, gender, and enrollment period (2003-2008, 2009-2011, 2012-2013).

Table 3.6 presents the cure model fit under different assumptions about the missingness mechanism. We see that the fits are nearly identical. The largest difference between the fits is in the estimate for the HPV effect on the time to recurrence in the non-cured group. We estimate a slightly stronger effect of HPV status under LMAR assumptions than under MAR assumptions, and the strongest effect is estimated when missingness is assumed to be LMAR dependent on G and other covariates. However, the HPV effect is not significant in any of the fits. We cannot make conclusions about the "correct" missingness mechanism, but regardless of the true missingness model, the CPH cure model inference appears to be very robust to different specifications of the working missingness model.

Table 3.6: Cure Model Fits to Head and Neck Data under LMAR Assumptions

Missingness model:	MAR		LMAR: *G* Only*		LMAR: *G* and Covariates*	
Patient Characteristic	Logistic OR, 95% CI	Failure Time HR, 95% CI	Logistic OR, 95% CI	Failure Time HR, 95% CI	Logistic OR, 95% CI	Failure Time HR, 95% CI
Age at Diagnosis						
10 Year ↑	1.14 (1.00, 1.31)†	1.08 (0.98, 1.19)	1.14 (0.99, 1.32)	1.08 (0.98, 1.18)	1.14 (1.00, 1.30)†	1.08 (0.98, 1.19)
Cancer Stage						
I/Cis (ref)						
II	1.25 (0.57, 2.74)	1.67 (0.70, 3.95)	1.25 (0.54, 2.89)	1.62 (0.69, 3.82)	1.25 (0.57, 2.74)	1.61 (0.66, 3.88)
III	2.36 (1.18, 4.72)†	2.42 (1.22, 4.79)†	2.32 (1.16, 4.61)†	2.40 (1.24, 4.66)†	2.33 (1.18, 4.63)†	2.42 (1.21, 4.84)†
IV	3.32 (1.74, 6.33)†	2.76 (1.48, 5.16)†	3.30 (1.74, 6.26)†	2.76 (1.47, 5.18)†	3.30 (1.80, 6.03)†	2.77 (1.45, 5.29)†
Cigarette Use						
Never (ref)						
Current	1.46 (0.97, 2.18)	0.98 (0.70, 1.38)	1.47 (0.96, 2.24)	0.97 (0.70, 1.35)	1.46 (0.98, 2.16)	0.97 (0.70, 1.33)
Former	1.27 (0.85, 1.90)	0.94 (0.66, 1.33)	1.28 (0.85, 1.93)	0.94 (0.67, 1.32)	1.28 (0.84, 1.95)	0.94 (0.67, 1.32)
HPV Status						
Negative (ref)						
Positive	0.34 (0.19, 0.58)†	0.91 (0.55, 1.48)	0.35 (0.19, 0.64)†	0.85 (0.51, 1.40)	0.34 (0.20, 0.56)†	0.81 (0.52, 1.28)
Comorbidities						
None (ref)						
Mild	1.14 (0.77, 1.69)	0.89 (0.65, 1.23)	1.15 (0.79, 1.68)	0.89 (0.65, 1.22)	1.15 (0.79, 1.68)	0.89 (0.65, 1.22)
Moderate	1.66 (1.08, 2.56)†	1.10 (0.75, 1.61)	1.66 (1.07, 2.58)†	1.09 (0.73, 1.61)	1.66 (1.07, 2.55)†	1.09 (0.75, 1.58)
Severe	1.94 (1.10, 3.43)†	1.07 (0.63, 1.80)	1.94 (1.08, 3.48)†	1.06 (0.64, 1.74)	1.97 (1.08, 3.57)†	1.06 (0.63, 1.80)
Cancer Site						
Larynx (ref)						
Hypopharynx	1.93 (0.88, 4.22)	1.43 (0.77, 2.67)	1.93 (0.86, 4.30)	1.42 (0.78, 2.60)	1.99 (0.91, 4.33)	1.42 (0.78, 2.58)
Oral Cavity	1.24 (0.81, 1.90)	1.33 (0.90, 1.97)	1.24 (0.81, 1.89)	1.32 (0.92, 1.90)	1.24 (0.81, 1.90)	1.32 (0.92, 1.89)
Oropharynx	1.68 (0.94, 3.02)	1.02 (0.62, 1.68)	1.64 (0.90, 2.97)	1.06 (0.66, 1.70)	1.68 (0.95, 2.96)	1.09 (0.69, 1.72)

*Predictors in working model for Prob(HPV missing). The LMAR model with *G* and covariates includes main effects for cancer site, age, and enrollment year group.

† Significant at p = 0.05

3.7 Discussion

We present a sequential imputation algorithm that can handle both MAR and LMAR covariate and outcome missingness for models with latent or partially latent variables. The proposed algorithm imputes the latent variable as part of the missing data.

We first propose an imputation approach assuming a fully-specified joint model for all the variables. In this setting, we demonstrate that the missingness mechanism can be ignored in the imputation steps for missing covariate and outcome values under MAR and LMAR when we condition on the latent variable. Additionally, we show that the missingness mechanism is not ignorable when imputing the latent variable under LMAR. We derive the forms of the posterior predictive distributions used for imputation under a fully-specified joint model. We then describe how we can use results based on a joint model to inform our imputation when we do not assume a fully-specified joint model, resulting in increased flexibility in the potential specification of imputation models used in practice.

The proposed imputation approach differs in several notable ways from existing approaches under MAR. In the joint modeling approach to imputation, the distributions used for imputation correspond to a valid joint distribution for the missing data. In the proposed algorithm, however, we allow missing covariate/outcome values to be imputed using distributions that do not correspond to a valid joint model (as is done in chained equations imputation). This allows for increased flexibility in the choice of the covariate and outcome imputation models over the joint modeling approach. However, unlike usual chained equations, we directly use the outcome model to inform our imputation of the latent variable and potentially the imputation of missing outcome/covariate values. Our proposed approach is similar to the covariate imputation approach in Bartlett et al. (2014) under MAR, except that our approach further addresses how to handling missingness in the outcome and latent variables. Therefore, the flexibility of the proposed imputation algorithm and the method's ability to incorporate outcome model assumptions into the imputation procedure are innovative even under MAR assumptions. Comparatively little work has been done to explore imputation under LMAR assumptions, and the proposed methods provide a flexible and novel approach to imputation under LMAR.

Simulations demonstrate that the proposed methods can result in "good" performance (in terms of bias, coverage, etc) under a variety of modeling scenarios as long as the working model contains the true model. We demonstrate that imputation assuming MAR can result in biased outcome model parameter estimates when missingness is truly LMAR. The proposed approach using LMAR assumptions can correct this bias.

Additional simulations explore the numerical convergence properties of the proposed imputation algorithm. We do not see evidence of convergence issues under MAR outcome missingness or MAR/LMAR covariate missingness except in the case where the working missingness model contains many highly correlated predictors. In some scenarios, we see convergence issues when we have LMAR outcome missingness, and parameters of the missingness model were particularly susceptible. Convergence problems can be substantially reduced by fixing parameters related to the latent variable in the missingness model.

We apply the imputation approach to a study of head and neck cancer recurrence. We impute missing values under MAR and LMAR assumptions, and the resulting model fits are very similar. In this application, the model inference is robust to the assumptions about missingness. We expect misspecification of the missingness model to have a greater impact when we have a larger amount of missingness in the latent variable or a stronger dependence between missingness and the latent variable.

One criticism of methods that do not assume a fully-specified joint distribution (e.g. chained equations) is that the algorithm is not guaranteed to converge to draws from a valid joint posterior predictive distribution for the missing values (Van Buuren et al., 2006). Our proposed imputation approach is similarly not guaranteed to converge to a valid joint distribution in general, and convergence can be impacted by identifiability issues. In this chapter, we do not prove convergence properties for the proposed algorithm beyond existing properties in the chained equations literature. Instead, we use simulation to identify settings that may be particularly susceptible to concerns about convergence. We demonstrate that the convergence of the proposed algorithm can be impacted by parameter identifiability. Care should be taken to monitor algorithm convergence, particularly in the setting of LMAR outcome missingness or with working missingness models containing many predictors. We similarly do not prove identifiability properties for general LMAR mechanisms. In some settings (e.g. Wu and Carroll, 1988; Miao et al., 2016), identifiability has been demonstrated analytically, but exploring identifiability can be difficult in general. We view proofs of identifiability for general LMAR mechanisms to be outside the scope of this work. Instead, we provide some guidance for applying the proposed methods in the presence of possible identifiability issues.

The proposed methods can be applied under MAR and LMAR outcome/covariate missingness. Unlike usual MAR-based imputation, the proposed imputation approach requires us to model the data missingness mechanism when missingness is assumed to be LMAR. However, this direct dependence on the missingness model provides a convenient framework for studying the sensitivity of outcome model inference to different assumptions about the missingness mechanism (Little, 1995; Molenberghs et al., 2008). Simulations suggest that the proposed LMAR-based imputation approach can be applied even in MAR settings as long as the working missingness model contains the true model and the LMAR-based model is well-identified. Additionally, when missingness is MAR, our proposed approach allows for greater flexibility in the specification of the covariate and outcome imputation models compared to joint modeling. The proposed method also allows us to incorporate the outcome model directly into the imputation of the latent variable (and possibly missing covariate/outcome values), potentially resulting in improved imputations and reduced bias in the downstream analysis compared to usual chained equations. Our proposed method, therefore, provides a flexible generalization of the usual MAR-based imputation that allows us to study a wider class of missingness models, of which MAR is a special case.

Chapter IV

Maximum Likelihood Estimation for Multistate Cure Models

4.1 Introduction

In medical applications, multistate models describe the rates at which individuals move between various health states. Multistate models have many valuable uses in medical research. Firstly, multistate models allow us to incorporate information from multiple event time outcomes (e.g. recurrence and death) in a unified way. Secondly, multistate models allow us to study which patient characteristics are relevant to which aspects of disease progression. Finally, multistate models are useful for making predictions, which can be valuable for medical decision-making.

The illness-death model is a popular multistate model explored in the literature and consists of three states: healthy (or no event), illness, and death (Andersen and Keiding, 2002). All subjects start out in the "healthy" state at baseline and can then move into the illness or death states as they develop illness or die from other causes. Subjects that develop illness can also transition into the death state. One common application of the illness-death model is in the study of cancer recurrence and death. In this setting, the "healthy" state represents subjects who have been treated for their initial cancer. For the remainder of this chapter, we will focus on the scenario with outcomes cancer recurrence and death, but these may be different types of events in general.

While the illness-death model is useful in many applications, one limitation is that the model implicitly assumes that all subjects can experience the illness event. In the context of cancer recurrence and death, this is equivalent to assuming that all subjects can experience a cancer recurrence. For many types of cancer, however, this may not be a reasonable assumption. In the case of head and neck cancer, for example, it has been well-established that some subjects can be completely cured of their initial disease, and these subjects will never experience a recurrence of their primary cancer (Taylor, 1995). We call the set of cured subjects the “cured fraction.”

In this chapter, we consider a generalization of the illness-death model called the multistate cure model that accounts for the cured fraction. The Bayesian multistate cure model developed in Conlon et al. (2013) breaks the “healthy” state of the illness-death model into two baseline states: cured and non-cured. The non-cured subjects can then experience either cancer recurrence or death under an illness-death model. The cured subjects can only experience the death event.

One challenge to fitting the multistate cure model is that cure status is partially latent. Subjects with an observed recurrence are known to be non-cured, but subjects censored for recurrence have unknown cure status. Cure status is unknown for all subjects at baseline. A natural question that arises in the context of cure models is our ability to identify the cured population when the event time distribution in the non-cured subjects may have a long tail (Farewell, 1982; Conlon et al., 2013). We will assume that we have sufficient follow-up after the last observed event time. One indicator of sufficient follow-up in a cure setting is that a Kaplan-Meier estimator applied to the time to recurrence outcome should have a clear plateau, indicating that there is a time-point after which recurrence events are no longer being observed. In the illness-death model setting, we require further follow-up for death before and after recurrence, so a lack of sufficient follow-up may be less of a concern for datasets that are well-suited for illness-death models.

Another problem for illness-death models in general and multistate cure models in particular is that the follow-up may not be the same for both outcomes of interest (Conlon et al., 2013). We call this situation “unequal follow-up.” For example, death status may be more easily obtained through death records, while assessment of cancer recurrence status requires a clinic visit, so it may often be the case that death status is known and recurrence status is unknown at a particular time t. Conlon et al. (2013) propose a Gibbs-sampling algorithm for fitting the multistate cure model in which values of cure status are drawn using a data augmentation approach and unequal follow-up is handled through a modification to the likelihood involving an integral. Their proposed algorithm performs well, but it requires substantial custom programming, it requires specification of prior distributions and tuning parameters, and it can take a long time to reach convergence. Additionally, the algorithm discussed in Conlon et al. (2013) assumes that covariates are fully-observed, which may not be the case in practice.

The Expectation-Maximization (EM) algorithm is an alternative, maximum-likelihood-based method in the literature for fitting models with latent variables or other types of missing data (Dempster et al., 1977). One advantage of the EM algorithm over Bayesian methods is that, in some cases, an EM algorithm can be more readily implemented using standard software. Additionally, the EM algorithm does not require specification of prior distributions or tuning parameters. In some complicated missing data scenarios, however, the conventional EM algorithm can be difficult to implement. The Monte Carlo EM (MCEM) algorithm proposed in Wei and Tanner (1990) provides a convenient, imputation-based approach for handling more complex missing data within a modified EM algorithm. EM and MCEM algorithms have not been previously explored in the context of multistate cure models in the literature, and development of such fitting algorithms could make the multistate cure model much more accessible to investigators.

In this chapter, we first propose a simple EM algorithm for fitting the standard multistate cure model. We then propose a MCEM algorithm for fitting the model in the presence of covariate missingness and/or unequal follow-up of the outcomes. The proposed algorithms can incorporate either parametric or nonparametric baseline hazards for the transitions between states and can incorporate different assumptions about the rate of death from other causes. The proposed EM algorithm makes use of a weighted likelihood representation, allowing it to be easily implemented using standard software. We provide software for implementing the EM and MCEM algorithms. We describe a novel approach for estimating standard errors for the MCEM algorithm. Simulations demonstrate the performance of the EM and MCEM algorithms under different modeling assumptions. We apply the proposed MCEM to a study of cancer recurrence and death of head and neck patients. We then derive expressions for estimating state occupancy probabilities, which can used to make predictions for individual patients.

In **Section 4.2**, we present details about the multistate cure model structure. In **Sections 4.3 and 4.4**, we propose an EM and MCEM algorithm for fitting the model. In **Section 4.5**, we discuss how to estimate standard errors. In **Section 4.6**, we derive state occupancy probabilities. We present a simulation study in **Section 4.7**, and we apply the proposed methods to head and neck cancer data in **Section 4.8**. In **Section 4.9**, we include a discussion.[2]

[2] A version of this chapter will be published in *Biostatistics* in 2018 under the title "EM Algorithms for Fitting Multistate Cure Models" by Lauren J Beesley and Jeremy M G Taylor.

4.2 Multistate Cure Model Specification

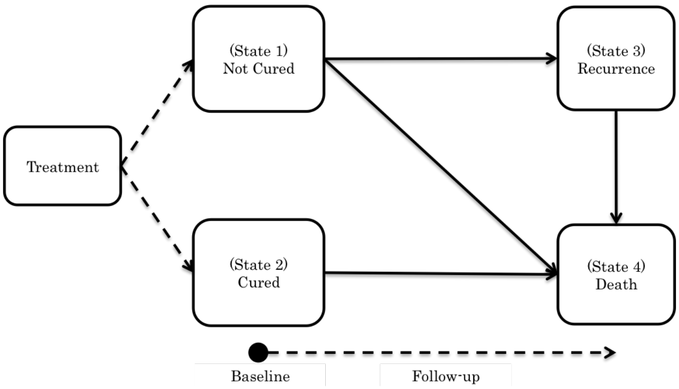
Suppose we have two semi-competing events: recurrence and death. By semi-competing, we mean that we can observe death after cancer recurrence but cannot observe recurrence after death. We further suppose that there is some subset of the subjects that are cured of their initial cancer and will never experience a cancer recurrence (even with very long follow-up).

Let T_{ir} and T_{id} be the underlying recurrence and death times for subject i. For cured subjects, $T_{ir} = \infty$. Let C_{ir} be the censoring time for recurrence (loss to follow-up) and C_{id} be the censoring time for death. Initially, we assume that the follow-up for recurrence and death is equal and define $C_i = C_{ir} = C_{id}$. For all subjects, we observe censored recurrence time information, $Y_{ir} = \min(T_{ir}, C_i, T_{id})$ and $\delta_{ir} = \mathbb{I}(Y_{ir} = T_{ir})$, and censored death time information, $Y_{id} = \min(T_{id}, C_i)$ and $\delta_{id} = \mathbb{I}(Y_{id} = T_{id})$. Let X_i denote the covariates for subject i, which we will initially suppose are fully observed.

We assume that all subjects have been previously treated for their initial cancer and did not have observable cancer at baseline. For non-cured subjects, however, some unobservable cancer cells remain and will grow until they are eventually observable, called cancer recurrence. Subjects with an observed recurrence are known to have been non-cured at baseline, but all other subjects have unknown baseline cure status. Let G_i be a variable indicating baseline non-cure status: $G_i = 0$ if cured, $G_i = 1$ if not cured. While we never know for sure that subjects are cured, we may strongly believe subjects still at risk after some time t_0 are cured and assign $G_i = 0$ for these subjects. Assigning some subjects to be cured improves the stability of the proposed methods, and we use this approach in our simulations and data example. Similar assumptions are often implicitly made for standard cure rate models through restrictions on the event rate in the non-cured subjects (Peng and Dear, 2000; Cai et al., 2012).

Figure 4.1 shows the conceptual structure of the model proposed in Conlon et al. (2013). Solid arrows represent potential state transitions given baseline cure status. If we removed state 4, the multistate cure model would reduce to the popular Cox proportional hazards (CPH) mixture cure model (Kuk and Chen, 1992; Sy and Taylor, 2000). We will assume that the underlying transition times between states are conditionally independent given covariates. We will further assume that C_{ir} and C_{id} are independent of all

Figure 4.1: Diagram of the Multistate Cure Model



underlying transition times given covariates.

We model the probability of not being cured by initial treatment using a logistic model: $\text{logit}(P(G_i = 1|X_i)) = \alpha_0 + \alpha_1^T X_{iG}$, where X_{iG} is a subset of X_i. We model the transition rate from state j to state k for all transitions except $3 \rightarrow 4$ using proportional hazards model $\lambda_{jk}(t) = \lambda_{jk}^0(t)\exp(\beta_{jk}^T X_{ijk})$, where X_{ijk} is the subset of X_i used in the model for transition $j \rightarrow k$. One important decision in multistate modeling is whether we reset time back to zero upon entering a new state (Putter et al., 2007; Meira-Machado et al., 2009). In the model for the $3 \rightarrow 4$ transition, we use the "clock reset" method in which time is reset to zero upon entering state 3, and use a proportional hazards regression to model the residual time in state 3 before entering state 4 as follows: $\lambda_{34}(t - T_{ir}) = \mathbb{I}(t > T_{ir})\lambda_{34}^0(t - T_{ir})\exp(\beta_{34}^T X_{i34})$. We can incorporate the time spent in state 1, T_{ir}, as a covariate in X_{i34} if desired.

Let $\Lambda_{jk}(t)$ and $\Lambda_{jk}^0(t)$ represent the cumulative hazard and cumulative baseline hazard for transition $j \rightarrow k$, and let $S_j(t)$ represent the probability of remaining in state j at time t. We have that $S_1(t) = \exp\{-\Lambda_{13}(t) - \Lambda_{14}(t)\}$, $S_2(t) = \exp\{-\Lambda_{24}(t)\}$, and $S_3(t - T_{ir}) = \exp\{-\Lambda_{34}(t - T_{ir})\}$ for $t > T_{ir}$ given T_{ir}. We may use a parametric or non-parametric form for the baseline hazards.

We may place additional assumptions on the hazards for the $2 \rightarrow 4$ and $1 \rightarrow 4$ transitions. Since these two transitions represent typically death from other causes, it may be reasonable to assume that the hazards are identical ($\Lambda_{14}(t) = \Lambda_{24}(t)\ \forall\ t \geq 0$). In this case, the multistate cure model reduces to a CPH cure model for recurrence time with two additional regressions for time to death with and without recurrence (Conlon et al., 2013). However, suppose we do not want to assume the hazards are equal. We may instead assume the hazards are proportional ($\Lambda_{14}(t) = \Lambda_{24}(t)\exp\{\beta_0\}$) or that the baseline hazards are equal ($\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$, β_{24} and β_{14} unrestricted), or we may make no equality assumptions ($\Lambda_{14}(t)$ and $\Lambda_{24}(t)$ unrestricted).

Let θ represent the set of model parameters. With parametric baseline hazards, θ includes $\alpha = (\alpha_0, \alpha_1)$, $\beta = (\beta_{13}, \beta_{24}, \beta_{14}, \beta_{34})$, and baseline hazard parameters. With nonparametric baseline hazards, θ includes α and β, and we treat the baseline hazards as unknown but fixed within the algorithm. We will assume that α and β are distinct. Let $\mathbb{D} = (Y, \delta, G, X)$ denote the complete data. The complete data log-likelihood for the multistate cure model takes the following form:

$$\begin{aligned} l(\theta|\mathbb{D}) = \sum_{i=1}^{n} &(1 - G_i)\log\big(P(G_i = 0)\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id})\big) \\ &+ G_i \log\Big(P(G_i = 1)\left[\lambda_{14}(Y_{id})^{\delta_{id}} S_1(Y_{id})\right]^{1-\delta_{ir}} \\ &\qquad \times \left[\lambda_{13}(Y_{ir}) S_1(Y_{ir}) \lambda_{34}(Y_{id} - Y_{ir})^{\delta_{id}} S_3(Y_{id} - Y_{ir})\right]^{\delta_{ir}}\Big) \end{aligned} \tag{4.1}$$

4.3 EM Algorithm

The EM algorithm is an approach for maximum likelihood estimation in the presence of missing data, which in our setting is the partially latent cure status (Dempster et al., 1977). Let $\mathbb{D}^{(obs)}$ represent the observed data and $\mathbb{D}^{(mis)}$ represent the missing data. The goal is to maximize the *observed* data log-likelihood, $l(\theta|\mathbb{D}^{(obs)})$, with respect to θ. The algorithm breaks the problem of maximizing $l(\theta|\mathbb{D}^{(obs)})$ into iterations of two simpler steps: the E-Step and the M-Step. In the E-Step, we calculate the expected value of the *complete* data log-likelihood conditioning on the observed data and the most recent parameter estimate, $\theta^{(t)}$, to obtain

$$Q(\theta|\theta^{(t)}) = \int l(\theta|\mathbb{D}^{(obs)}, \mathbb{D}^{(mis)}) f(\mathbb{D}^{(mis)}|\mathbb{D}^{(obs)}, \theta^{(t)}) d\mathbb{D}^{(mis)}$$

In the M-Step, we maximize $Q(\theta|\theta^{(t)})$ with respect to θ. We iterate these steps many times until convergence of the estimated θ to the MLE.

The EM algorithm is a common estimation method in the literature for models with latent classes. Frydman and Kadam (2004) proposed an EM algorithm for estimation for a continuous time multistate model in which the underlying population is split into movers and stayers such that only the movers are eligible to experience a state transition. This setting is very similar to our setting except that our model allows both the movers and stayers to experience death.

4.3.1 E-Step

In the E-Step for fitting the multistate cure model, we take the expectation of (4.1). Since (4.1) is linear in G_i, we can obtain $Q(\theta|\theta^{(t)})$ by replacing G_i in (4.1) with $E(G_i|X_i, Y_{id}, Y_{ir}, \delta_{id}, \delta_{ir}; \theta^{(t)})$ when G_i is unknown. Let $R_i = \mathbb{I}(G_i \text{ known})$. For all subjects, we replace G_i with

$$\begin{aligned} p_i &= \delta_{ir} + (1-\delta_{ir})(1-R_i)P(G_i = 1|X_i, Y_{id}, Y_{ir}, \delta_{id}, \delta_{ir} = 0; \theta^{(t)}) \qquad (4.2) \\ &= \delta_{ir} + (1-\delta_{ir})(1-R_i) \left. \frac{P(G_i=1)\lambda_{14}(Y_{id})^{\delta_{id}} S_1(Y_{id})}{P(G_i=1)\lambda_{14}(Y_{id})^{\delta_{id}} S_1(Y_{id}) + P(G_i=0)\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id})} \right|_{\theta=\theta^{(t)}} \end{aligned}$$

In order to use the formula in (4.2), we need estimates of the baseline hazard functions for the 1 → 3, 1 → 4, and 2 → 4 transitions. Under parametric assumptions, the complete form of the baseline hazards are determined in the M-Step. When the baseline hazards are non-parametric, we estimate the baseline hazards prior to calculating (4.2). In **Appendix E**, we use the profile likelihood method to derive estimators for the baseline hazards. The form of the estimator depends on the estimate of p_i from the previous iteration and whether we assume that the baseline hazards for the 2 → 4 and 1 → 4 transitions are equal.

4.3.2 M-Step

In the M-Step, we maximize $Q(\theta|\theta^{(t)})$ with respect to θ. After replacing G_i with p_i in (4.1) and reorganizing the terms, we have that

$$\begin{aligned}
Q(\theta|\theta^{(t)}) = &\sum_{i=1}^{n}(1-p_i)\log\left[P(G_i=0)\right] + p_i\log\left[P(G_i=1)\right] \\
&+ (1-p_i)\log\left[\lambda_{24}(Y_{id})^{\delta_{id}}\exp\{-\Lambda_{24}(Y_{id})\}\right] \\
&+ p_i\log\left[\lambda_{14}(Y_{id})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}(Y_{id})\}^{1-\delta_{ir}}\exp\{-\Lambda_{14}(Y_{ir})\}^{\delta_{ir}}\right] \\
&+ p_i\log\left[\lambda_{13}(Y_{ir})^{\delta_{ir}}\exp\{-\Lambda_{13}(Y_{ir})\}^{\delta_{ir}}\exp\{-\Lambda_{13}(Y_{id})\}^{1-\delta_{ir}}\right] \\
&+ p_i\delta_{ir}\log\left[\lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}}\exp\{-\Lambda_{34}(Y_{id}-Y_{ir})\}\right]
\end{aligned}$$

Since we assume that censoring times for recurrence and death are the same, $\delta_{ir} = 0$ implies $Y_{ir} = Y_{id}$. Additionally, we note that $\delta_{ir} = 1$ implies $p_i = 1$. We can therefore rewrite Q as:

$$\begin{aligned}
Q(\theta|\theta^{(t)}) = &\sum_{i=1}^{n}(1-p_i)\log\left[P(G_i=0)\right] + p_i\log\left[P(G_i=1)\right] \\
&+ (1-p_i)\log\left[\lambda_{24}(Y_{id})^{\delta_{id}}\exp\{-\Lambda_{24}(Y_{id})\}\right] \\
&+ p_i\log\left[\lambda_{14}(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}(Y_{ir})\}\right] \\
&+ p_i\log\left[\lambda_{13}(Y_{ir})^{\delta_{ir}}\exp\{-\Lambda_{13}(Y_{ir})\}\right] \\
&+ \delta_{ir}\log\left[\lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}}\exp\{-\Lambda_{34}(Y_{id}-Y_{ir})\}\right]
\end{aligned} \tag{4.3}$$

The terms involving α and β separate, so we can maximize (4.3) with respect to α and β separately. The terms involving α resemble the log-likelihood for a logistic model with p_i as the outcome. We can estimate α by fitting a logistic regression to p_i using predictors X_{iG}. We can estimate β (and perhaps baseline parameters) by maximizing the last four terms of (4.3) with respect to the parameters.

We can perform the maximization for β by fitting a single survival model to an augmented version of the data. This single-survival model maximization strategy is convenient in settings where we want to impose additional parameter or baseline restrictions across transitions. We first note that each of the last four summands of (4.3) takes the form of a weighted proportional hazards regression model for a different state transition. We will use this property to combine the four terms in the above sum into a single weighted proportional hazards regression model.

We consider an augmented version of the data that contains four rows for each subject (one for each transition in the multistate cure model). Each row contains a variable indicating the transition being considered (S), the time the subject was at risk for that transition (T), an indicator for whether the subject experienced that transition (D), a weight variable (W), and covariates (Z). **Table 4.1** shows the form of the rows in the augmented dataset for each subject i.

Table 4.1: Augmented Data Structure for Subject i

Transition	S	T	D	W	Z
1 → 3	13	Y_{ir}	δ_{ir}	p_i	X_i
2 → 4	24	Y_{id}	δ_{id}	$1 - p_i$	X_i
1 → 4	14	Y_{ir}	$\delta_{id}(1 - \delta_{ir})$	p_i	X_i
3 → 4	34	$Y_{id} - Y_{ir}$	δ_{id}	δ_{ir}	X_i

We note that subjects with an observed recurrence have $p_i = 1$, so we could equivalently replace δ_{id} in row 2 of **Table 4.1** with $\delta_{id}(1 - \delta_{ir})$. Using the augmented data structure, we can rewrite the last four terms in (4.3) as

$$\sum_{m=1}^{4n} W_m log\left(\left[\lambda^0_{S_m}(T_m)\exp\{g(Z_m, S_m; \beta)\}\right]^{D_m} \exp\{-\Lambda^0_{S_m}(T_m)\exp[g(Z_m, S_m; \beta)]\}\right) \quad (4.4)$$

where $g(Z_m, S_m; \beta)$ is a function of Z_m and S_m that may include linear functions of Z_m and S_m along with interactions between Z_m and S_m. The sum in (4.4) takes the form of a single weighted log-likelihood for a proportional hazards regression model.

When the baseline hazards are modeled parametrically, we can maximize (4.4) directly with respect to β by fitting a single survival model to the outcome data (T, D) with weights W and some function of S and Z as predictors. We can include interactions between Z and S in the function g in order to allow the β's to differ across transitions. In order to impose different baseline hazards for different transitions, we can stratify the baseline hazard by S (or a grouped version of S). We can accommodate different covariate sets across transitions by not including particular covariate-strata combinations (effectively setting some parameters in a covariate-strata interaction to zero).

Suppose we include the same set of covariates in each one of the transitions, so the elements of Z are the same within subjects. Below, we describe how we can specify g and stratify the baseline hazard to incorporate different assumptions about the $1 \rightarrow 4$ and $2 \rightarrow 4$ transitions. In each setting, we can then fit this model to (T,D) using weights W.

Example 1*: No Restrictions*

Suppose first that we do not impose any restrictions on the $1 \rightarrow 4$ and $2 \rightarrow 4$ transitions. Then, we can formulate the survival regression model as follows:

$$g(Z, S; \beta) = \beta_{13} Z * \mathbb{I}(S = 13) + \beta_{24} Z * \mathbb{I}(S = 24) + \beta_{14} Z * \mathbb{I}(S = 14) + \beta_{34} Z * \mathbb{I}(S = 34)$$

where the baseline hazard is stratified by S.

Example 2*: Equal Hazards*

Suppose instead that we restrict the $1 \rightarrow 4$ and $2 \rightarrow 4$ hazard functions to be equal. Then, we can formulate the survival regression model as follows:

$$g(Z, S; \beta) = \beta_{13} Z * \mathbb{I}(S = 13) + \beta_{1424} Z * \mathbb{I}(S = 24 \text{ or } S = 14) + \beta_{34} Z * \mathbb{I}(S = 34)$$

where the baseline hazard is stratified into three categories in which $S = 14$ and $S = 24$ are merged into one group.

Example 3*: Equal Baseline Hazards*

Suppose instead that we restrict the $1 \rightarrow 4$ and $2 \rightarrow 4$ baseline hazard functions to be equal while allowing the corresponding β's to be different. Then, we can formulate the survival regression model as follows:

$$g(Z, S; \beta) = \beta_{13} Z * \mathbb{I}(S = 13) + \beta_{24} Z * \mathbb{I}(S = 24) + \beta_{14} Z * \mathbb{I}(S = 14) + \beta_{34} Z * \mathbb{I}(S = 34)$$

where the baseline hazard is stratified into three categories in which $S = 14$ and $S = 24$ are merged into one group.

Example 4: *Proportional Hazards*

Suppose instead that we assume that the hazards for the 1 → 4 and 2 → 4 transitions are proportional rather than equal. Then, we can formulate the survival regression model as follows:

$$g(Z, S; \beta) = \beta_{13} Z{*}\mathbb{I}(S = 13) + \beta_{1424} Z{*}\mathbb{I}(S = 24 \text{ or } S = 14) + \beta_{34} Z{*}\mathbb{I}(S = 34) + \beta_0 \mathbb{I}(S = 14)$$

where the baseline hazard is stratified into three categories in which $S = 14$ and $S = 24$ are merged into one group.

When the baseline hazards are modeled nonparametrically, we approximate (4.4) by a single weighted Cox partial log-likelihood and maximize with respect to β by fitting a Cox regression model as above. Through this single model fit, we can obtain estimates of β and, if the baseline hazards are parametric, the parameters related to the baseline hazard.

The proposed method for estimating β is similar to the methods used by *mstate* in R and other multistate modeling software except that it incorporates transition-specific weights and involves a different augmented data structure (de Wreede et al., 2011).

4.4 Extension to Handle Additional Missing Data

The EM algorithm in **Section 4.3** assumes that cure status is the only source of missing data. However, additional missing data often arises in practice. One source of missing data is missingness in the covariates. Another common source of missingness occurs when the follow-up for recurrence is shorter than the follow-up for death. We call this phenomenon "unequal follow-up." In order to follow up for recurrence, patients must come into the clinic, while death status can be more easily obtained from death registries. In this case, recurrence status may only be known up to time t, while death status may be known up to time $s > t$. This results in missing information about recurrence status on the interval $(t, s]$, which we will treat as missing data. This setting is similar to interval censoring and panel data for illness-death models (Jackson, 2011).

Conlon et al. (2013) handles the problem of unequal follow-up by constructing a modified likelihood function involving an integral. Another potential solution is to censor death back to the follow-up time for recurrence for subjects with unequal follow-up. This approach is unappealing since it throws out valuable information about death. A third solution is to modify the conventional EM algorithm so that the E-Step takes the expectation over all types of missing data. However, when we have complicated patterns of missing data, these expectations can be difficult to compute. We consider an alternative approach called the Monte Carlo EM Algorithm, which takes an imputation-based approach to dealing with missing data.

4.4.1 Monte Carlo EM Algorithm

The Monte Carlo EM algorithm (MCEM) proposed in Wei and Tanner (1990) provides a convenient approach to handling complicated missing data within a modified EM algorithm. The strategy is to replace the usual E Step from the EM algorithm with a step in which we obtain M imputations $\mathbb{D}^{(t,1)}, \mathbb{D}^{(t,2)}, \ldots, \mathbb{D}^{(t,M)}$ of $\mathbb{D}$ by drawing the missing data from $f(\mathbb{D}^{(mis)}|\mathbb{D}^{(obs)}, \theta^{(t)})$.

The M Step of the MCEM algorithm then involves maximizing the complete data log-likelihood mixed over the imputed values:

$$Q_{mix}(\theta|\theta^{(t)}) = \frac{1}{M}\sum_{m=1}^{M} l(\theta|\mathbb{D}^{(t,m)}) \propto \sum_{m=1}^{M} l(\theta|\mathbb{D}^{(t,m)})$$

Suppose we create a stacked version of the dataset, called $\mathbb{D}^{(t)}$, obtained by stacking the imputed versions of the data such that each subject appears in the dataset M times. We then have that $Q_{mix}(\theta|\theta^{(t)}) \propto l(\theta|\mathbb{D}^{(t)})$. We can estimate θ by maximizing $l(\theta|\mathbb{D}^{(t)})$ with respect to θ. We then iterate between the imputation and M steps until "convergence", where successive estimates of θ fall around the $\theta = \hat{\theta}$ line with noise (Wei and Tanner, 1990). We can estimate θ by taking the mean parameter estimate across the last few iterations of the MCEM algorithm.

The imputation step involves drawing the missing data M times from $f(\mathbb{D}^{(mis)}|\mathbb{D}^{(obs)}, \theta^{(t)})$. Unlike conventional multiple imputation, missing data is drawn from the predictive distribution of the missing data evaluated at a single estimated parameter value, $\theta^{(t)}$, rather than independent draws of the parameter θ (Wei and Tanner, 1990; Neath, 2012). Therefore, the imputations produced are "improper" as described in Little and Rubin (2002). In addition to imputing missing covariate or outcome values, the Monte Carlo EM algorithm will also involve imputing values for the partially latent cure status, and we will impute each type of missing data separately.

4.4.2 Imputation for Unequal Follow-up

Unequal follow-up is very common when the outcomes of interest are recurrence and death and occurs in many other semi-competing risks settings. In **Appendix G**, we present a derivation of the proposed imputation approach and provide recommendations for implementation. Here, we include a brief description of the general approach.

Let C_r be the censoring time for recurrence and C_d be the censoring time for death, but now assume that $C_r \leq C_d$ and for some subjects, $C_{ir} < C_{id}$. For all subjects, we observe C_{ir}-censored recurrence information, $Y_{ir}^0 = \min(T_{ir}, C_{ir}, T_{id})$ and $\delta_{ir}^0 = \mathbb{I}(Y_{ir} = T_{ir})$, and C_{id}-censored death information, $Y_{id} = \min(T_{id}, C_{id})$ and $\delta_{id} = \mathbb{I}(Y_{id} = T_{id})$. Our goal is to impute values of $Y_{ir} = \min(T_{ir}, C_{id}, T_{id})$ and $\delta_{ir} = \mathbb{I}(Y_{ir} = T_{ir})$ that would have been observed if we had followed subjects for recurrence as long as we followed them for death.

Suppose we treat previously imputed G and X as known. Values of Y_{ir} and δ_{ir} are only unknown for subjects with imputed $G_i = 1$ and observed $Y_{ir}^0 < Y_{id}$ and $\delta_{ir}^0 = 0$. Define $\mathbb{Z} = (Y_{ir}^0, \delta_{ir}^0, Y_{id}, \delta_{id}, G_i, X_i)$. We impute missing δ_{ir} from a Bernoulli distribution with probability

$$P(\delta_{ir} = 1|\mathbb{Z}; \theta^{(t)}) = \frac{\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(t)S_1(t)S_3(Y_{id} - t)\lambda_{34}(Y_{id} - t)^{\delta_{id}}dt}{\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(t)S_1(t)S_3(Y_{id} - t)\lambda_{34}(Y_{id} - t)^{\delta_{id}}dt + \lambda_{14}^{\delta_{id}}(Y_{id})S_1(Y_{id})}$$

If imputed $\delta_{ir} = 0$, we set $Y_{ir} = Y_{id}$. Otherwise, we draw $Y_{ir} = T_{ir}$ from

$$f(T_{ir} = t|\delta_{ir} = 1, \mathbb{Z}; \theta^{(t)}) \propto \lambda_{13}(t)S_1(t)S_3(Y_{id} - t)\lambda_{34}(Y_{id} - t)^{\delta_{id}}\mathbb{I}(Y_{ir}^0 < t < Y_{id})$$

4.4.3 Imputation of Cure Status

The cure status imputation approach will depend on whether we have unequal follow-up in the outcome. First, we will assume there is equal follow-up for all subjects. In this case, we can draw missing G_i using $P(G_i = 1|X_i, Y_{id}, Y_{ir}, \delta_{id}, \delta_{ir} = 0)$ as shown in equation (4.2).

Suppose instead that we have unequal follow-up. We perform imputation of G_i conditioning on the observed Y_{ir}^0 and δ_{ir}^0 but not the imputed values of Y_{ir} and δ_{ir}, which allows imputations to more easily move between $G_i = 0$ and $G_i = 1$ in successive iterations. We can impute missing G_i from a Bernoulli distribution using probability $P(G_i = 1|X_i, Y_{id}, Y_{ir}^0, \delta_{id}, \delta_{ir}^0 = 0)$:

$$\frac{P(G_i = 1)\left[\lambda_{14}(Y_{id})^{\delta_{id}}S_1(Y_{id}) + C_i\right]}{P(G_i = 1)\left[\lambda_{14}(Y_{id})^{\delta_{id}}S_1(Y_{id}) + C_i\right] + P(G_i = 0)\lambda_{24}(Y_{id})^{\delta_{id}}S_2(Y_{id})}$$

$$\text{where } C_i = \mathbb{I}(Y_{ir}^0 < Y_{id} \text{ and } \delta_{ir}^0 = 0)\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(t)S_1(t)S_3(Y_{id} - t)\lambda_{34}(Y_{id} - t)^{\delta_{id}}dt$$

4.4.4 Imputation of Missing Covariates

Many methods can be used to perform the covariate imputation. For a detailed discussion of imputation methods, we refer the reader to Little and Rubin (2002). Chained equations is one popular approach to imputation in which we specify regression models for each covariate with missingness and impute each covariate one-by-one (Van Buuren et al., 2006).

A modification of chained equations proposed in Bartlett et al. (2014) uses the structure of the outcome model (rather than a simple regression model) to obtain the imputation distributions. We use this approach in our simulations and data example. Let $X^{(p)}$ denote the p^{th} covariate in X and $X^{(-p)}$ denote all but the p^{th} covariate. Under MAR assumptions (as defined in Little and Rubin, 2002), we impute each $X^{(p)}$ with missingness from its full conditional distribution, which is proportional to $l(\theta^{(t)}|\mathbb{D})f(X^{(p)}|X^{(-p)};\psi^{(t)})$, where $f(X^{(p)}|X^{(-p)};\psi^{(t)})$ is the conditional distribution of $X^{(p)}$ given $X^{(-p)}$. This expression is viewed as a function of $X^{(p)}$, treating all other imputed variables as fixed and using estimated values of θ and ψ.

4.5 Estimating Standard Errors

The EM and MCEM algorithms provide an estimate of θ, but they do not readily provide corresponding standard errors. Many methods have been explored in the literature for estimating standard errors for parameters estimated using an EM algorithm. Some methods involve $l(\theta|\mathbb{D})$ and its derivatives (e.g. Louis, 1982). Bootstrap methods are also popular. This approach is commonly used to estimate standard errors for the CPH cure model, and we use this approach for estimating standard errors for the EM algorithm in our simulations (Sy and Taylor, 2000).

A similar bootstrap approach can be used to estimate the standard errors from the MCEM algorithm, but we do not recommend this approach due to the relative slowness of the MCEM fitting algorithm. The usual approach for estimating standard errors after an MCEM algorithm estimation of θ is a generalization of Louis's method proposed in Wei and Tanner (1990) with:

$$I(\theta) = -\frac{1}{M}\sum_{m=1}^{M}\frac{D^2 l(\theta|\mathbb{D}^{(t,m)})}{D^2\theta} - \frac{1}{M}\sum_{m=1}^{M}\left(\frac{Dl(\theta|\mathbb{D}^{(t,m)})}{D\theta}\right)^2 + \left(\frac{1}{M}\sum_{m=1}^{M}\frac{Dl(\theta|\mathbb{D}^{(t,m)})}{D\theta}\right)^2$$

where $D^{(t,1)}, \ldots, D^{(t,M)}$ are the M imputed versions of $\mathbb{D}$ from the last iteration of the MCEM algorithm. The estimated covariance matrix for $\hat{\theta}$ is $I(\hat{\theta})^{-1}$. This approach is usually implemented using large M for the last few iterations of the model fitting algorithm. This approach requires us to directly compute first and second derivatives of $l(\theta|\mathbb{D})$ with respect to θ, which may not be convenient. Additionally, in the proposed MCEM algorithm, the M imputations at a given iteration depend on the M imputations from the previous iteration, so we cannot easily change the value of M across iterations.

As an alternative to the "standard" approach for estimating standard errors, we propose a post-processing method (below) to obtain proper multiple imputations of $\mathbb{D}$. After fitting the multistate cure model to each imputed dataset separately, we can then use Rubin's multiple imputation combining rules to obtain standard errors that correctly account for the uncertainty related to the missing data (Little and Rubin, 2002). This approach is convenient because 1) it does not require us to use large M for any iterations and 2) it does not require us to directly compute derivatives of the observed data log-likelihood. In simulations (not shown), we found similar performance under our proposed estimation method and the approach proposed in Wei 1990. However, the post-processing step is important, and skipping the post-processing resulted in undercoverage of multi-state cure model parameters.

We propose the following method to obtain proper multiple imputations of $\mathbb{D}$ using the (improper) multiple imputations obtained within the MCEM algorithm. Our goal is to obtain M independent draws from $f(D^{(mis)}|D^{(obs)})$, which are our proper multiple imputations. At the end of the MCEM algorithm, we have M independent draws from $f(D^{(mis)}|D^{(obs)};\theta^{(t)})$, where $\theta^{(t)}$ is the estimate of θ at the last iteration. Let $\mathbb{D}^{(t,1)},\mathbb{D}^{(t,2)},\ldots,\mathbb{D}^{(t,M)}$ denote the imputations at the final iteration t of the MCEM algorithm. We can obtain M approximate draws from $f(D^{(mis)}|D^{(obs)})$ by performing the following for each $\mathbb{D}^{(t,m)}$. [Step 1]: Estimate θ on a bootstrap sample of the most recent $\mathbb{D}^{(m)}$. Since $\mathbb{D}^{(m)}$ contains no missingness, this estimation is easy to perform. This results in an approximate draw of θ from $f(\theta|\mathbb{D}^{(m)})$ under a flat prior (Little and Rubin, 2002). [Step 2]: Using the draw of θ, obtain an updated imputation $\mathbb{D}^{(m)}$ of $\mathbb{D}$ as described in **Section 4.4**. [Step 3]: Repeat Steps 1-2 several times. We can then use these proper multiple imputations of $\mathbb{D}$ for variance estimation. We provide some theoretical justification for this approach in **Appendix H**.

4.6 Prediction

We can use the multistate cure model fit to estimate the probability that a subject will be in a particular state at time t given that subject's baseline predictors. These probabilities can be useful for predicting prognosis or exploring the potential impact of different treatments. Below, we provide expressions for estimating state occupancy probabilities over time given only baseline covariate information. Then, we provide similar expressions that also incorporate some limited post-baseline follow-up.

State Occupancy Probabilities Given Baseline Covariates

We are interested in estimating quantities related to the unmeasured (or incompletely measured) variables, T_r and T_d. While cure status theoretically exists at baseline, it is not observed at the baseline time. Thus, we assume this is unknown for prediction. Instead, we will only assume that the baseline predictors, X, are known. We recall that T_r is defined as infinity when $G = 0$, so we have $P(T_r < t|X, G = 0) = 0$. Using the structure of the multistate cure model, we derive the following probabilities, which sum to 1 for a given t:

$$\begin{aligned}
P(T_r < T_d < t|X) &= P(G = 1|X) \int_0^t [1 - S_3(t - T_r)]\, \lambda_{13}(T_r) S_1(T_r)\, dT_r \\
P(T_r < t < T_d|X) &= P(G = 1|X) \int_0^t S_3(t - T_r) \lambda_{13}(T_r) S_1(T_r)\, dT_r \\
P(T_r > t, T_d > t|X) &= P(G = 1|X) S_1(t) + P(G = 0|X) S_2(t) \\
P(T_d < T_r, T_d < s|X) &= P(G = 1|X) \int_0^t \lambda_{14}(T_d) S_1(T_d)\, dT_d + P(G = 0|X)(1 - S_2(t))
\end{aligned}$$

These expressions can be calculated using numerical integration and a multistate cure model fit. We can estimate these probabilities for different values of t to create probability curves over time.

State Occupancy Probabilities Given Baseline Covariates and Alive, Non-Recurrent at Time t^*

Suppose now that we know the subject is alive and non-recurrent at time t^*. We then want to estimate the state occupancy probabilities at time $s \geq t^*$. For $s \geq t^*$, we have that:

$$
\begin{aligned}
P(T_r < T_d < s|X, T_d > t^*, T_r > t^*) =& \frac{\pi_1(s) - \pi_1(t^*)}{\pi_3(t^*)} C(t) \\
P(T_r < s < T_d|X, T_d > t^*, T_r > t^*) =& \frac{\pi_2(s) - \pi_2(t^*)}{\pi_3(t^*)} C(t) \\
P(T_r > s, T_d > s|X, T_d > t^*, T_r > t^*) =& \frac{\pi_3(s)}{\pi_3(t^*)} C(t) + \frac{S_2(s)}{S_2(t^*)}(1 - C(t)) \\
P(T_d < T_r, T_d < s|X, T_d > t^*, T_r > t^*) =& \frac{\pi_4(s) - \pi_4(t^*)}{\pi_3(t^*)} C(t) + \frac{S_2(t^*) - S_2(s)}{S_2(t^*)}(1 - C(t))
\end{aligned}
$$

where

$$
C(t) = P(G = 1|X, T_d > t^*, T_r > t^*) = \frac{\pi_3(t^*)P(G = 1|X)}{\pi_3(t^*)P(G = 1|X) + S_2(t^*)P(G = 0|X)}
$$

State Occupancy Probabilities Given Baseline Covariates and Alive at Time t^* with Prior Recurrence

Suppose that we know the subject is alive at time t^* and had a prior recurrence at time $t_0 \leq t^*$. We then want to estimate the state occupancy probabilities at time $s \geq t^*$. For $s \geq t^*$, we have that:

$$
\begin{aligned}
P(T_r < T_d < s|X, T_d > t^*, T_r = t_0) =& 1 - \frac{S_3(s - t_0)}{S_3(t^* - t_0)} \\
P(T_r < s < T_d|X, T_d > t^*, T_r = t_0) =& \frac{S_3(s - t_0)}{S_3(t^* - t_0)} \\
P(T_r > s, T_d > s|X, T_d > t^*, T_r = t_0) =& 0 \\
P(T_d < T_r, T_d < s|X, T_d > t^*, T_r = t_0) =& 0
\end{aligned}
$$

4.7 Simulation Study

4.7.1 Simulation 1: Bias, Efficiency, and Coverage of Multistate Cure Model Parameters

We simulate 500 datasets with 2000 subjects each under a multistate cure model with two bivariate normal covariates (standard normal with correlation of 0.5) and Weibull baseline hazards. We then generate cure status using $\text{expit}(P(G_i = 1|X)) = 0.5 + 0.5X_1 + 0.5X_2$. For cured subjects, we simulate a death time using a proportional hazards model with $\Lambda^0_{24}(t) = 0.002t^{1.4}$. For non-cured subjects, we generate time to recurrence, time to death from other causes, and time to death after recurrence with $\Lambda^0_{13}(t) = 0.005t^2$, $\Lambda^0_{14}(t) = 0.002t^{1.4}$, and $\Lambda^0_{34}(t) = 0.08t^{1.9}$ respectively. These baseline hazards were chosen to mimic relative event rates that we might expect to see in real data. We may expect that the rate of death from other causes will be low relative to the rate of recurrence. Additionally, we may expect that the death rate after recurrence is very high relative to both the rate of death from other causes and the rate of recurrence. For all transitions, $\beta = (0.5, 0.5)^T$. We simulate event times such that the hazards for the $1 \rightarrow 4$ and $2 \rightarrow 4$ transitions are equal.

We consider three simulation scenarios: 1) no covariate missingness or unequal follow-up, 2) covariate missingness, and 3) unequal follow-up. For scenarios 1 and 2, an outcome censoring time was generated from $U(10, 80)$. This provides sufficient follow-up so that a clear plateau can be observed in the Kaplan-Meier plot of time to recurrence, allowing the cure rate to be well-estimated. For scenario 3, censoring time for death was generated from $U(10, 80)$. For all but the first 750 subjects, we impose an earlier $U(10, 40)$ censoring time for recurrence. We use these simulated values to determine the observed data for each subject. This leads to roughly 25% of the subjects needing imputation for unequal censoring. For scenario 2, we impose $\sim 30\%$ MCAR missingness in X_2. In all scenarios, we assume subjects still at risk for recurrence after time 50 are cured. This value was chosen as to point in which the Kaplan-Meier plots for time to recurrence show a clear plateau, indicating no more recurrence events are observed after that point.

For each simulated dataset, we fit a multistate cure model using the proposed EM algorithm (scenario 1) or the MCEM algorithm (scenarios 2 and 3). For the MCEM algorithm, we use $M = 10$. Within each scenario, we consider different assumptions regarding baseline hazards (Weibull or Cox) and restrictions for the $1 \rightarrow 4$ and $2 \rightarrow 4$ transition hazards. Simulations using the MCEM algorithm and Cox baseline hazards used at least 50 iterations under Weibull baseline hazards and then switched to 50 iterations under Cox baseline hazards. All other simulations used 100 iterations of the EM/MCEM algorithm. Variances of the parameter estimates from the EM algorithm are estimated using 50 bootstrap samples. Variances from the Monte Carlo EM algorithm are obtained using the Rubin's rules-based approach described in **Section 4.5** with 5 iterations of post-processing. Unequal censoring imputation under Weibull and Cox baseline hazards use the rejection sampling method and Metropolis-Hastings method respectively. We then compute the bias, empirical variance, and coverage rates of the multistate cure model parameter estimates across the 500 datasets. We also record the median run time and the number of simulations with numerical issues (non-converging M-Step or difficulty with variance estimation) for each scenario.

Tables 4.2-4.3 show the results. When we assume $\Lambda_{14}(t) = \Lambda_{24}(t)$ under Weibull or Cox baseline hazard assumptions, the proposed algorithms result in essentially unbiased parameter estimates with nominal coverage rates in all scenarios. When we assume $\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$, we again see good bias and coverage properties under Weibull or Cox baseline hazards for scenarios 1 and under Weibull baseline hazards for scenario 2 and 3. For scenarios 2 and 3 under Cox baseline hazards, we see increased bias and/or undercoverage for the parameters related to the $1 \rightarrow 4$ and $2 \rightarrow 4$ transitions and the logistic regression. When we assume $\Lambda_{14}(t) \propto \Lambda_{24}(t)$, we generally obtain good bias and coverage properties for all failure time model parameters. For the intercept in the logistic model, we tend to see some undercoverage, particularly when we assume Cox baseline hazards and when we have unequal follow-up. We explore possible causes of this issue in **Appendix I**. When we do not restrict $\Lambda_{14}(t)$ and $\Lambda_{24}(t)$, we obtain good bias and covariate properties in scenario 1 under Weibull baseline hazards, but we see increased bias and undercoverage in all other settings. Failing simulations provide additional evidence of numerical instability. Overall, the proposed algorithms can provide good numerical properties in all three scenarios when the assumptions for the $1 \rightarrow 4$ and $2 \rightarrow 4$ hazards are sufficiently restrictive. When the restrictions are relaxed (particularly when the baseline hazards are not equal), we can run into numerical problems, and these problems tend to be greater under Cox baseline hazards and when we have unequal follow-up.

Table 4.2: Multistate Cure Model β Estimates using Proposed Methods

Results across 500 simulations are presented using the following notation: Bias (Empirical Variance) Coverage of 95% Confidence Interval, each multiplied by 100. States 1 and 2 are the non-cured and cured baseline states. States 3 and 4 represent recurrence and death.

Baseline	$2 \rightarrow 4$, $1 \rightarrow 4^{\ddagger}$	Failure Time Models*							
Hazard	Assumption	$1 \rightarrow 3$		$2 \rightarrow 4$		$1 \rightarrow 4$		$3 \rightarrow 4$	
		X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
Scenario 1: No Covariate Missingness or Unequal Follow-up									
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.17) 94	0 (0.15) 97	0 (0.47) 94	0 (0.45) 96	0 (0.47) 94	0 (0.45) 96	0 (0.14) 95	0 (0.15) 95
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.18) 95	0 (0.16) 96	0 (1.03) 94	0 (0.97) 96	-1 (3.22) 93	0 (3.19) 94	0 (0.14) 94	0 (0.15) 94
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.17) 95	0 (0.15) 96	0 (0.48) 95	0 (0.47) 95	0 (0.48) 95	0 (0.47) 95	0 (0.14) 95	0 (0.15) 94
Weibull	None	0 (0.18) 95	0 (0.16) 96	0 (1.12) 95	1 (1.04) 96	0 (4.54) 95	0 (5.30) 95	0 (0.15) 95	0 (0.15) 95
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.18) 94	0 (0.15) 97	0 (0.48) 94	0 (0.45) 96	0 (0.48) 94	0 (0.45) 96	0 (0.15) 95	0 (0.16) 95
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.18) 94	0 (0.16) 96	0 (1.05) 94	0 (0.97) 96	-1 (3.40) 94	0 (3.24) 93	0 (0.17) 95	0 (0.16) 94
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.18) 94	0 (0.16) 96	0 (0.48) 94	0 (0.47) 94	0 (0.48) 94	0 (0.47) 94	0 (0.15) 94	0 (0.16) 95
Cox	None	-1 (0.83) 70	-1 (0.82) 68	-6 (5.07) 60	-4 (5.19) 58	67 (15.1) 31	65 (16.7) 35	0 (0.15) 95	0 (0.16) 95
Scenario 2: Covariate Missingness									
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.17) 94	0 (0.22) 95	0 (0.45) 97	0 (0.63) 97	0 (0.45) 97	0 (0.63) 97	0 (0.15) 95	0 (0.17) 96
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.19) 95	0 (0.23) 95	0 (0.97) 96	0 (1.22) 96	-1 (3.36) 95	0 (4.20) 93	0 (0.15) 95	0 (0.17) 96
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.17) 95	0 (0.22) 96	0 (0.47) 96	0 (0.64) 96	0 (0.47) 96	0 (0.64) 96	0 (0.47) 97	0 (0.64) 96
Weibull	None	0 (0.19) 94	0 (0.24) 94	1 (1.11) 94	0 (1.42) 95	-1 (3.81) 72	0 (5.28) 73	0 (0.14) 96	0 (0.18) 95
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.18) 95	0 (0.22) 95	0 (0.46) 97	0 (0.64) 96	0 (0.46) 97	0 (0.64) 96	0 (0.16) 96	0 (0.19) 97
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.20) 95	0 (0.24) 96	1 (1.07) 96	-3 (1.30) 95	0 (3.52) 94	0 (4.20) 94	0 (0.16) 95	0 (0.21) 96
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.18) 95	0 (0.23) 95	1 (0.49) 96	-2 (0.61) 95	1 (0.49) 96	-2 (0.61) 95	0 (0.16) 97	0 (0.19) 96
Cox	None	0 (0.53) 76	1 (0.52) 81	6 (6.13) 53	5 (5.97) 60	15 (28.8) 37	16 (26.7) 44	0 (0.15) 96	0 (0.18) 97
Scenario 3: Unequal Follow-up									
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.19) 95	0 (0.21) 94	0 (0.50) 94	0 (0.50) 95	0 (0.50) 94	0 (0.50) 95	0 (0.16) 95	0 (0.16) 95
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.21) 94	0 (0.22) 95	0 (1.05) 96	0 (1.03) 95	-1 (3.53) 94	-1 (3.39) 94	0 (0.16) 96	0 (0.16) 95
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.19) 95	0 (0.21) 94	0 (0.50) 95	0 (0.51) 96	0 (0.50) 95	0 (0.51) 96	0 (0.16) 97	0 (0.16) 95
Weibull	None	0 (0.21) 94	0 (0.22) 95	1 (1.23) 95	0 (1.29) 93	-2 (4.30) 71	0 (3.63) 76	0 (0.17) 95	0 (0.17) 94
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.20) 95	0 (0.21) 94	-1 (0.46) 94	-2 (0.49) 95	-1 (0.46) 94	-2 (0.49) 95	0 (0.16) 96	0 (0.16) 95
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	1 (0.23) 94	1 (0.24) 93	3 (1.39) 94	1 (1.40) 93	-3 (3.98) 92	-2 (4.15) 91	0 (0.16) 95	0 (0.16) 95
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.21) 94	1 (0.23) 93	1 (0.55) 95	1 (0.55) 95	1 (0.55) 95	1 (0.55) 95	0 (0.16) 98	0 (0.16) 95
Cox	None	0 (0.65) 70	1 (0.61) 71	4 (6.49) 43	3 (5.75) 51	15 (30.5) 35	16 (25.0) 37	0 (0.16) 95	0 (0.17) 95

Table 4.3: Multistate Cure Model α Estimates using Proposed Methods

Results across 500 simulations are presented using the following notation: Bias (Empirical Variance) Coverage of 95% Confidence Interval, each multiplied by 100. The number of simulations (out of 500) with numerical issues and the median run time per simulation are also shown.

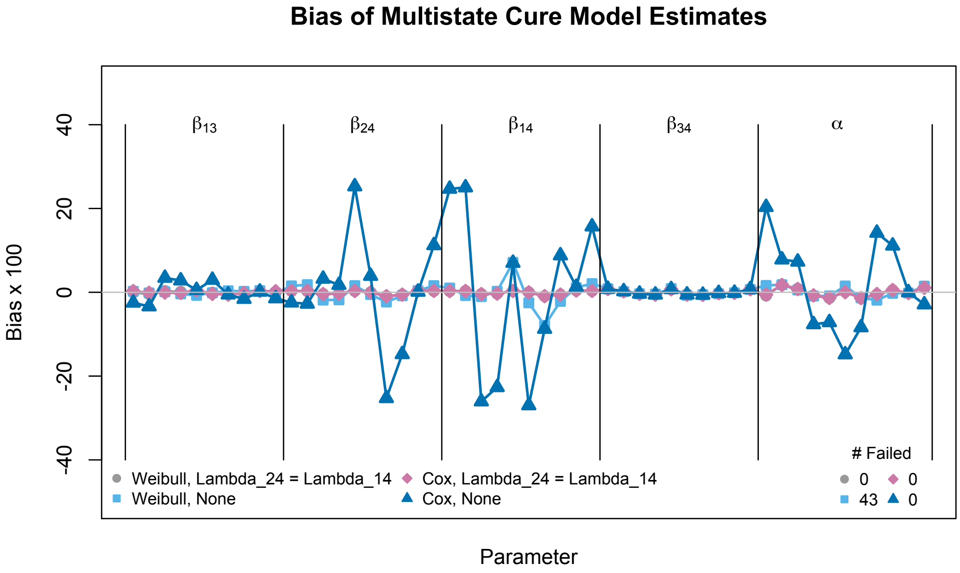
Baseline	$2 \rightarrow 4$, $1 \rightarrow 4$		Logistic Model*		# Failed	Run Time
Hazard	Assumption	Intercept	X_1	X_2	(out of 500)	(mins/sim)
		Scenario 1: No Covariate Missingness or Unequal Follow-up				
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.33) 93	0 (0.42) 95	0 (0.44) 94	0	2.02
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.34) 94	0 (0.51) 94	0 (0.51) 94	0	2.12
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.59) 93	0 (0.41) 95	0 (0.45) 95	0	2.07
Weibull	None	1 (0.71) 93	0 (0.49) 96	0 (0.51) 97	46	2.16
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.33) 93	0 (0.42) 97	0 (0.44) 95	0	7.68
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.34) 93	0 (0.51) 95	0 (0.52) 95	0	8.04
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	1 (0.89) 84	0 (0.43) 94	0 (0.48) 95	0	7.98
Cox	None	11 (0.37) 50	2 (2.64) 60	2 (2.83) 64	1	8.46
			Scenario 2: Covariate Missingness			
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.32) 96	0 (0.44) 97	0 (0.60) 95	0	5.65
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.33) 96	0 (0.54) 97	0 (0.70) 95	1	5.71
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.63) 90	0 (0.47) 97	0 (0.60) 95	0	5.73
Weibull	None	2 (0.72) 89	0 (0.59) 95	0 (0.73) 93	141	5.57
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.32) 96	0 (0.46) 96	0 (0.60) 95	0	27.5
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.33) 94	1 (0.58) 96	-2 (0.71) 93	0	27.3
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	1 (1.59) 81	1 (0.50) 96	-3 (0.60) 94	0	27.4
Cox	None	6 (0.63) 80	-2 (2.15) 61	-2 (2.28) 67	82	26.7
			Scenario 3: Unequal Follow-up			
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.35) 96	0 (0.51) 95	0 (0.48) 96	0	8.49
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.35) 96	0 (0.64) 95	0 (0.63) 94	0	8.61
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.63) 92	0 (0.54) 94	0 (0.51) 96	0	8.56
Weibull	None	3 (0.69) 92	0 (0.66) 95	0 (0.64) 93	102	8.58
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	-2 (0.35) 94	1 (0.53) 93	1 (0.48) 95	0	18.3
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	4 (0.43) 87	-2 (0.77) 93	-2 (0.81) 90	0	18.5
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	5 (1.95) 74	-2 (0.60) 94	-2 (0.57) 94	0	18.5
Cox	None	4 (0.70) 83	-1 (2.65) 53	-2 (2.44) 60	19	17.8

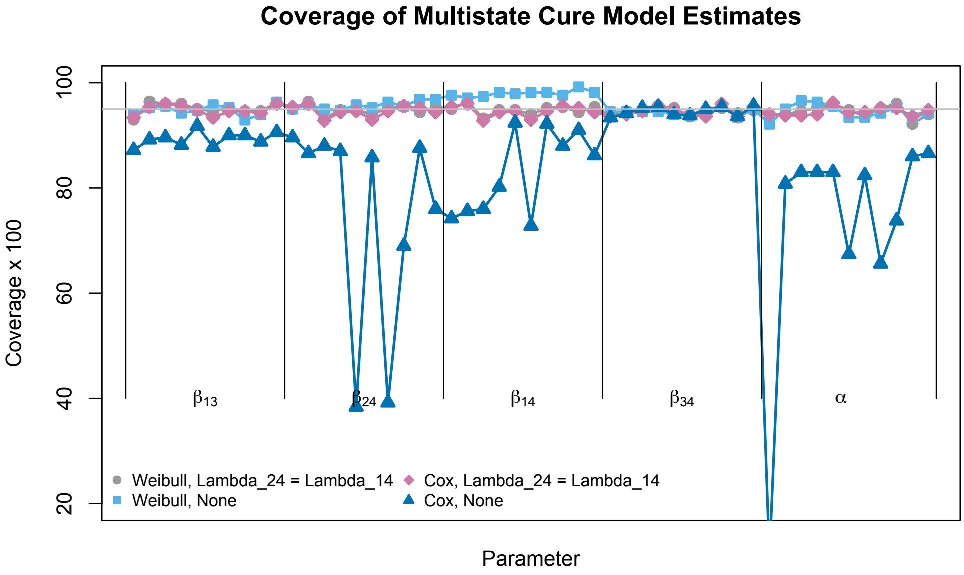
4.7.2 Simulation 2: Multistate Cure Model Estimates with More Covariates

In the second set of simulations, we simulate 500 datasets with 2000 subjects each. Each dataset contains 10 multivariate normal covariates with correlations of 0.5 for each pair. We then fit four different multistate cure models to the simulated data with different baseline hazards (Weibull and Cox) and different assumptions about the $2 \rightarrow 4$ and $1 \rightarrow 4$ transitions (equal hazards and no restrictions) using 100 iterations of the EM algorithm. Standard errors are estimated using 50 bootstrap samples.

Figure 4.2 presents the bias and coverage for the estimated θ from each model fit. When we assume $\Lambda_{24}(t) = \Lambda_{14}(t)$, we obtain essentially unbiased parameter estimates with nominal coverage under both Weibull and Cox baseline hazard assumptions. When we assume Weibull baseline hazards and no $2 \rightarrow 4$ and $1 \rightarrow 4$ hazard restrictions, we see some increased bias and overcoverage for estimating $1 \rightarrow 4$ parameters, but otherwise we have good bias and coverage properties. However, we do see evidence of model instability as 119 out of the 500 simulations had numerical issues. When we assume Cox baseline hazards and no $2 \rightarrow 4$ and $1 \rightarrow 4$ hazard restrictions, we see substantial bias and/or undercoverage, particularly in estimating β_{24}, β_{14} and α. As in Simulation 1, we see evidence of estimation instability in general when we place no restrictions on the hazards for the $2 \rightarrow 4$ and $1 \rightarrow 4$ transitions.

Figure 4.2: Bias and Coverage of Multistate Model Estimates with Ten Covariates





4.8 Application to Head and Neck Cancer Data

We consider again consider the head and neck cancer data discussed in previous chapters, but for this analysis, we have obtained an updated version of the dataset containing data N=1519 patients treated for head and neck squamous cell carcinoma (HNSCC).

After treatment, patients were followed for recurrence and death. Covariate information was also collected at baseline. We are interested in studying the association between these covariates and the time to HNSCC recurrence and death after treatment. Additionally, it is been well-established that some head and neck cancer patients can be cured of their cancer through their primary treatment, and we are interested in identifying factors related to the underlying cure probability (Taylor, 1995; Grau et al., 1997; Cognetti et al., 2008). The analysis of an earlier version of these data presented in **Chapter II** explores time to recurrence and cure probability in a Cox proportional hazards cure model, but this analysis does not incorporate death information.

Missing data, however, poses an additional complication. For many patients (62.3%), follow-up for recurrence was substantially shorter than follow-up for death, resulting in unequal follow-up of recurrence and death for many subjects. Also, HPV status was unavailable for 50.1% of the subjects, and a small amount of missingness was present in other study variables. **Table 4.4** provides descriptives of the analytical sample. We restricted our analyses to subjects with the following cancer sites: oropharynx, oral cavity, larynx, and hypopharynx. We further restricted these analyses to subjects that appeared to clear their cancer through the initial treatment. As a result, we excluded 193 subjects with persistent disease from our analysis, resulting in our dataset of size N=1519. Deaths were observed for 556 (36.6%) of subjects, and recurrences were observed for 354 (23.3%) of subjects. Median survival was 129 months [95% CI (108, 142)], and median follow-up time for death was 65.6 months [95% CI (62.9, 72.5)]. The median follow-up time for recurrence was 47.2 months [95% CI (38.4, 48.0)]. **Table 4.4** provides descriptives of the analytical sample. **Table 4.5** provides a breakdown of the observed outcome information. We assume that subjects still at risk for recurrence and death at 80 months are cured (starting in State 2). We have a large portion of the subjects with unequal follow-up for recurrence and death. Additionally, baseline cure status is unknown for many subjects.

Table 4.4: Characteristics of Study Patients at HNSCC Diagnosis

Characteristic	N (%) or Mean (SD)	Missing N (%)	Characteristic	N (%) or Mean (SD)	Missing N (%)
Age at Diagnosis	59.5 (11.4)		Comorbidities		4 (0.2)
Cancer Stage		0 (0)	None	387 (25.4)	
I/Cis	245 (16.1)		Mild	667 (43.9)	
II	183 (12.0)		Moderate	318 (20.9)	
III	222 (14.6)		Severe	143 (9.4)	
IV	869 (57.2)		Cancer Site		0 (0)
Cigarette Use		28 (1.8)	Larynx	334 (21.9)	
Never	352 (23.1)		Hypopharynx	61 (4.0)	
Current	673 (44.3)		Oral Cavity	509 (33.5)	
Former	466 (30.6)		Oropharynx	615 (40.4)	
HPV Status		761 (50.1)	Gender		1 (0.06)
Negative	404 (26.5)		Female	386 (25.4)	
Positive	354 (23.3)		Male	1132 (74.5)	

Table 4.5: Observed Outcome Information for HNSCC Dataset

	N (% of 1519)
Initial State	
State 1	354 (23.3)
State 2	160 (10.5)
Unknown	1005 (66.1)
Observed Transition	
Recurrence	
1 → 3	354 (23.3)
Death	
3 → 4	287 (18.8)
2 → 4	22 (1.4)
1 → 4	0 (0)
1 or 2 (Unknown) → 4	247 (16.2)
Unequal Follow-up	
No	563 (37.0)
Yes	956 (62.9)

Figure 4.3 provides a visual display of the unequal follow-up in this dataset. The black bars represent follow-up for death, and the red and blue dots indicate recurrence censoring and events respectively. Censoring of recurrence time often occurs at yearly check-ups, resulting in the banded pattern. However, observed recurrences do no follow this banded pattern. We notice that for many subjects, there is a substantial difference in follow-up for recurrence and death, so the method for handling unequal follow-up is particularly important. Additionally, we notice that there are some (6) subjects with very late observed recurrences among the subjects with recurrence events. The general rule is that primary recurrences do not usually happen for head and neck cancer after

60 months past treatment (Taylor, 1995; Grau et al., 1997). We were initially concerned that these subjects were not experiencing recurrence of the primary tumor, but review of the medical records does not provide enough evidence to rule out a classification of primary recurrence. **Figure 4.4** provides the Kaplan-Meier estimator applied to the time to recurrence data. The red vertical lines indicate recurrence events. We can see that, although there are a few late events, the majority of the events occur prior to 60 months, and very few occur after 80 months.

Figure 4.3: Plot of Observed and Censored Recurrence Times

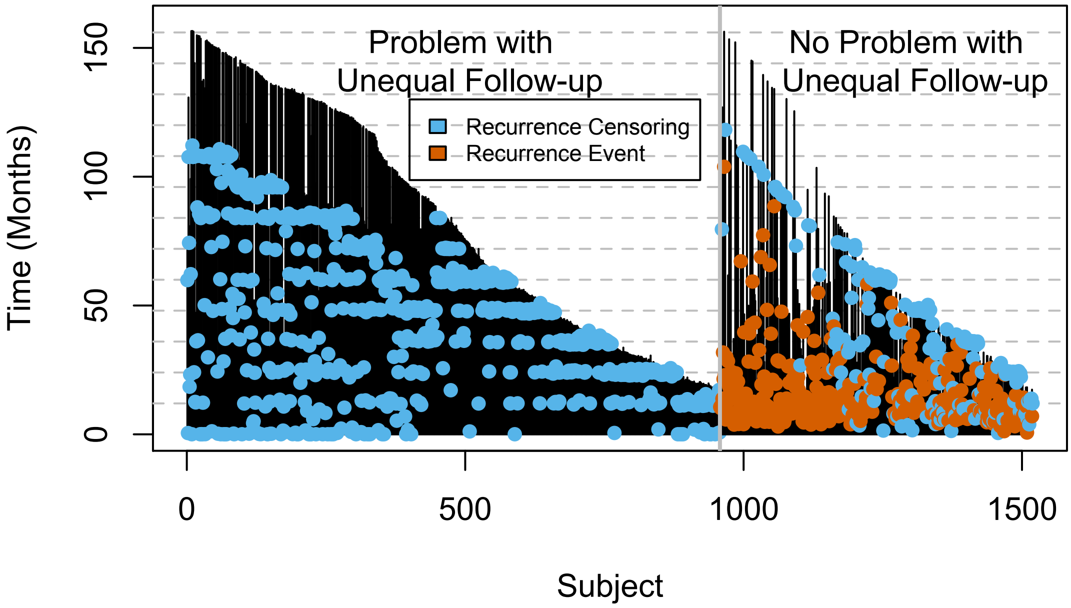
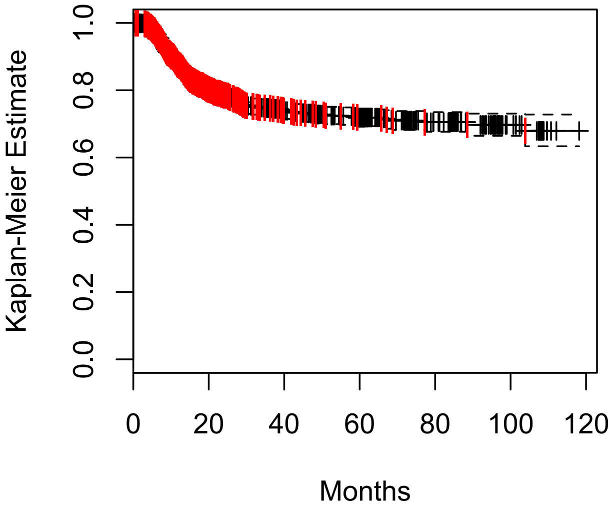


Figure 4.4: KM Plot of Time to Recurrence



We fit a multistate cure model assuming Weibull baseline hazards and equal $1 \to 4$ and $2 \to 4$ hazards using 100 iterations of the MCEM algorithm. A particularly tricky element of the multistate cure model is that subjects experiencing the $1 \to 4$ transition always have missing cure status. This means that there are no known events for that transition, making estimation for that transition particularly difficult. We assume that the hazards for the $1 \to 4$ and $2 \to 4$ transitions are equal, which greatly improves our ability to estimate model parameters. Missing data were imputed using the method proposed in Bartlett et al. (2014) as described in **Section 4.4.4**. Unequal follow-up was handled using the approach described in **Section 4.4.2**. Standard errors were estimated by treating the most recent imputations as proper imputations (after post-processing) and estimating variance within each imputed dataset using 50 bootstrap samples. Rubin's rules were then used to obtain the final estimates for the standard errors.

Figure 4.5 presents the results of the multistate cure model fit to the head and neck cancer dataset. Higher cancer stage and HPV negativity were associated with higher rates of recurrence for non-cured subjects. Greater age, higher cancer stage, worse comorbidities, and increased smoking history were associated with higher rates of death from other causes for both cured and non-cured subjects. Higher cancer stage and increased smoking history were associated with higher rates of death after recurrence, and larynx site was associated with lower rates of death after recurrence compared to oral cavity cancer. Higher cancer stage and HPV negativity were associated with lower probabilities of being cured by treatment.

One way to evaluate the convergence of an EM algorithm is by plotting the observed data log-likelihood across iterations of the algorithm. We expect the observed data log-likelihood to increase and then flatten out once convergence has been reached. In the case of the Monte Carlo EM algorithm, we can obtain an estimate of the observed data log-likelihood by taking the mean of the complete data log-likelihood across the imputations of the data at a given iteration. With few imputations (we use 10), this is a noisy estimate of the observed data log-likelihood, so the estimated observed data log-likelihood may jump around the true observed data log-likelihood curve. **Figure 4.6** shows the estimated observed data log-likelihood. The estimate appears to stabilize around iteration 20 and then follow a flat line with noise. This suggests that the algorithm has adequately converged.

Figure 4.5: Multistate Cure Model Fit to Head and Neck Data

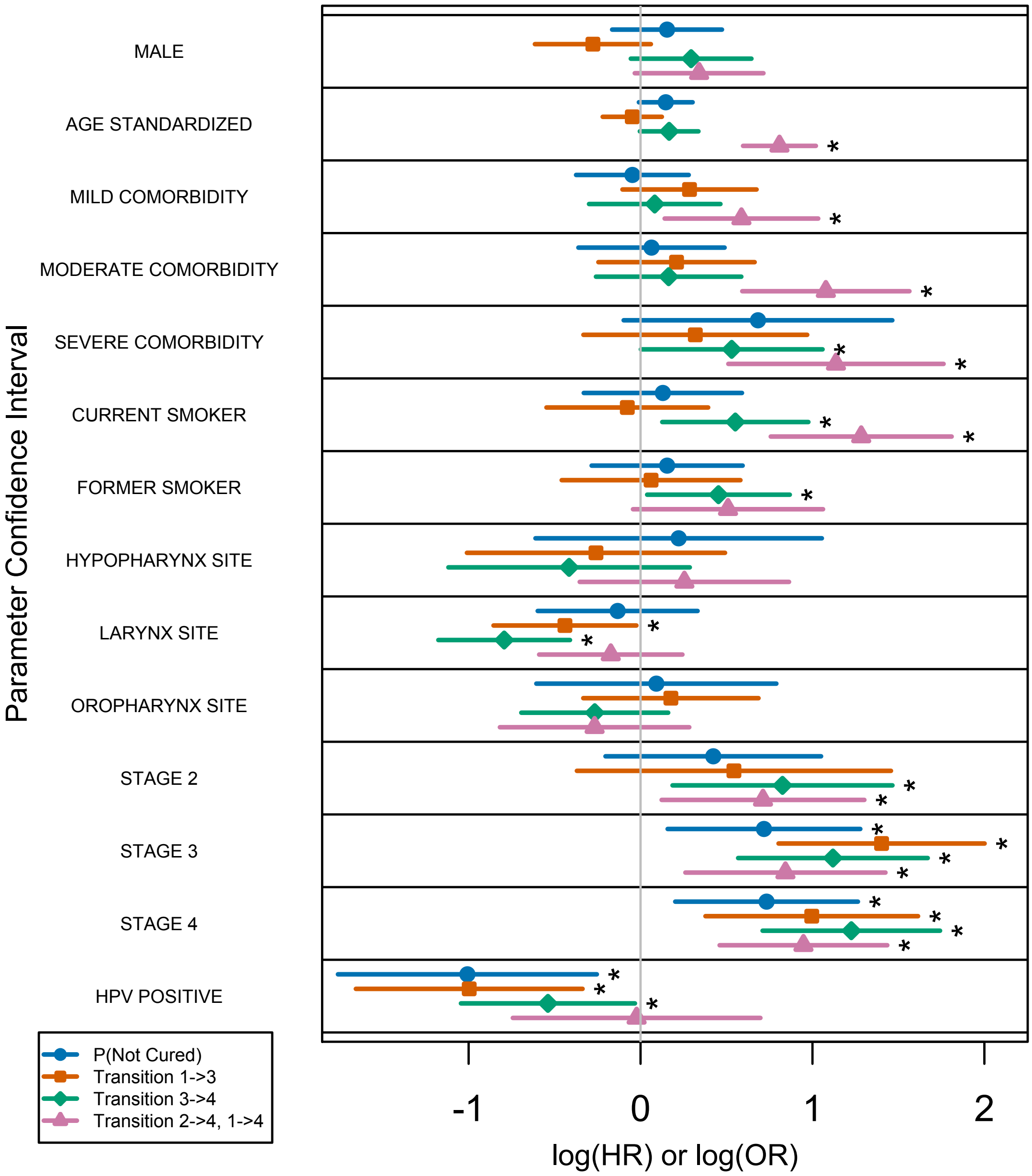
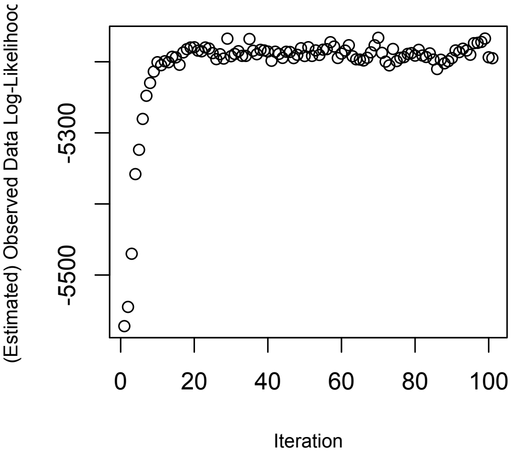
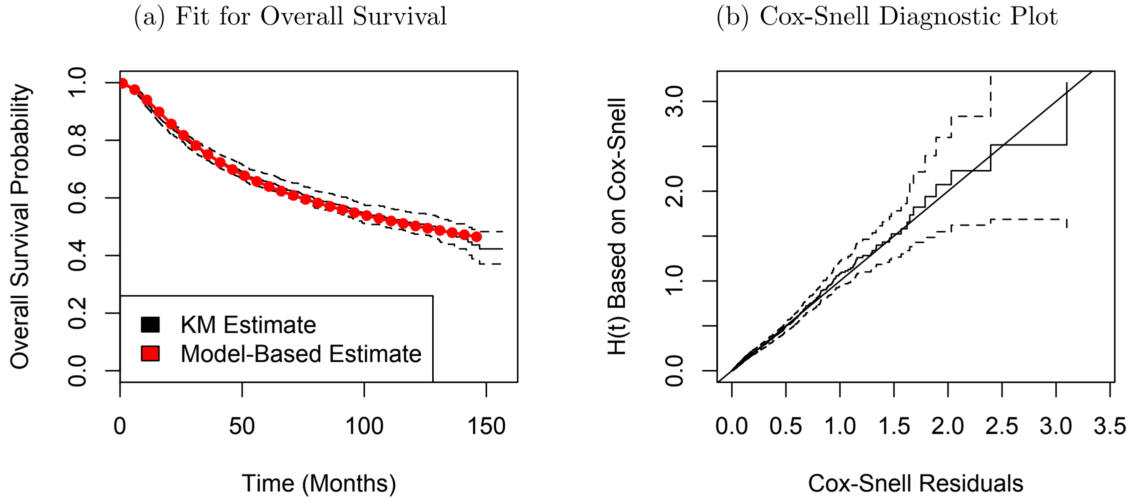


Figure 4.6: Estimation of Observed Data Log-Likelihood



While we can develop methods to fit the multistate cure model, it is important to evaluate whether the model itself is well-specified for a particular dataset. **Figure 4.7** presents some goodness of fit diagnostics for the multistate cure model fit to the head and neck cancer data. **Figure 4.7a** compares the Kaplan-Meier estimate of Overall Survival with the model-predicted survival curve. The model-predicted curve is the average of predicted survival curves for each individual. Each individual's predicted survival curve is an average of predicted survival probabilities across the 10 imputed datasets from the Monte Carlo EM algorithm. The survival probability at time t given the subject's covariate values can be calculated using the expressions in **Section 4.6**. The multistate cure model fit does an excellent job at predicting the marginal survival probability. **Figure 4.7b** displays a Cox-Snell diagnostic plot. The Cox-Snell residuals are calculated as $r_i = -log(P(T_d > Y_d|X))$ for each subject, where $P(T_d > Y_d|X)$ is the average of the estimated survival probability for subject i at Y_{id} across the 10 imputed datasets. If the model fits well, we expect these residuals to be exponentially-distributed. We then use the Kaplan-Meier method to estimate a cumulative hazard function with event time r_i and event indicator δ_{id} for each subject. If the residuals are exponentially-distributed, we expect the cumulative baseline hazard to lie on the $y = x$ line. This plot again indicates that the multistate cure model fits these data well.

Figure 4.7: Goodness of Fit Diagnostic Plots


(a) Fit for Overall Survival

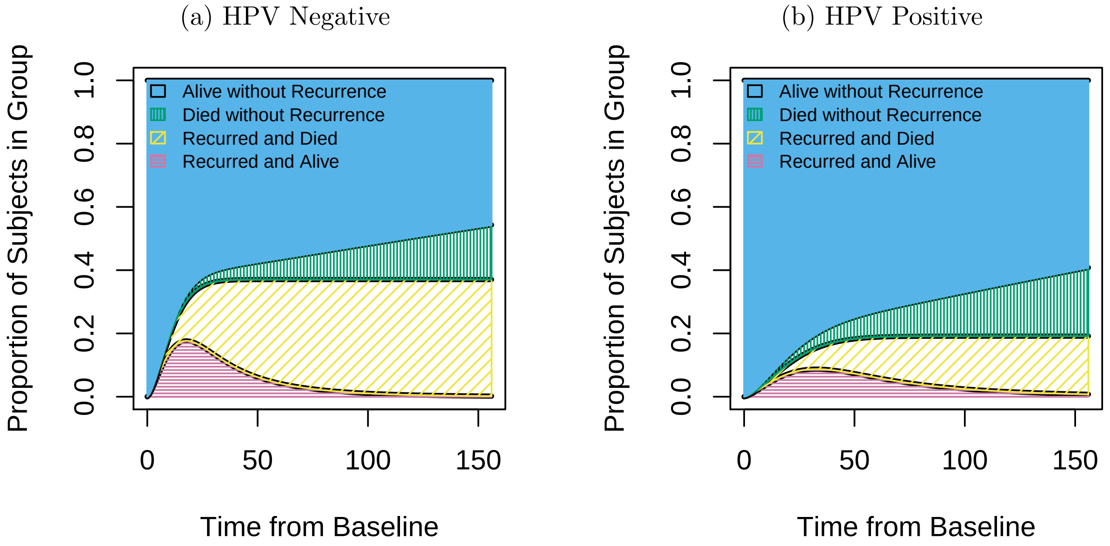
(b) Cox-Snell Diagnostic Plot

Once we have established that the model fit has converged and appears reasonable for the data, we can use the model to predict outcomes for new subjects given only the baseline covariates. We note that cure status is not known at baseline. We can use the expressions in **Section 4.6** to estimate the probabilities of different outcome events given the subject's baseline characteristics. **Figure 4.8** shows the predicted state occupancy probabilities over time for a subject with particular baseline characteristics.

We note that other multistate cure model specifications could have been made. For

Figure 4.8: Predicted State Occupancy Probabilities

Male 60-year-old with Stage 4 Oropharyngeal Cancer, Mild Comorbidities, and No History of Smoking by HPV Status


(a) HPV Negative

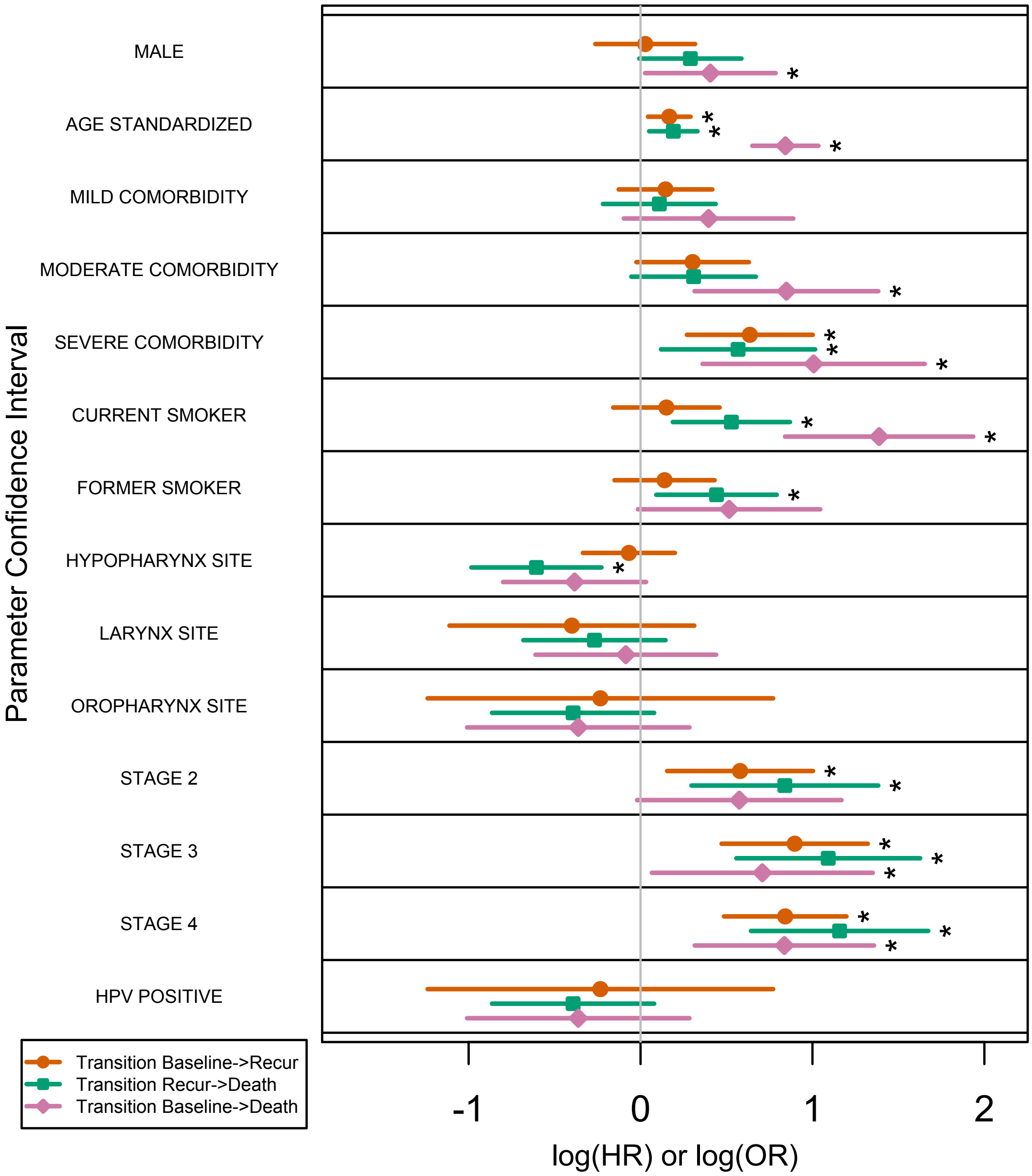
(b) HPV Positive

this dataset, we ran into numerical issues when we tried to relax the assumption of equal 1 → 4, 2 → 4 transition hazards to equal baseline hazards. This is probably due to the large number of predictors included in the model for each transition along with the large amount of missing data. Additionally, imputation of the missing covariates proceeded only using the covariates in the final model fit, but information regarding p16 mutations for oropharynx subjects was also available, and it was highly correlated with observed HPV status in these patients. When we incorporate the p16 information into the imputation of HPV status within the MCEM algorithm, we obtain a very similar model fit as in **Figure 4.5**. An additional modification to this model would include interactions between HPV status and cancer site, as the effect of HPV status may be believed to differ across site. A model including a full site-HPV status interaction ran into numerical troubles due to the small number of subjects with the hypopharynx subsite in our dataset. We were able to fit a model including a separate term for the HPV effects within the oropharynx subsite for each transition, and the resulting interaction terms were all non-significant. An additional assumption we made when fitting the multistate cure model to the head and neck data was that subjects at risk for recurrence at time = 80 months are cured. This threshold was chosen as a point in which the Kaplan-Meier plot of time to recurrence has reached a plateau, but other thresholds could be chosen. We repeated our analysis use a threshold of $t_0 = 100$, and we obtained very similar results.

While the multistate cure model may appear to be well-suited to these data, we may wonder whether a simpler illness-death model may provide an adequate fit to the data while avoiding some of the numerical complications. We fit the illness-death model to the head and neck cancer data (grouping the cured and non-cured groups in the multistate cure model). We handle missing covariates through imputation using SMC-FCS imputation based on the illness-death model structure and we handle unequal censoring using the imputation approach proposed (Bartlett et al., 2014).

Figure 4.9 provides the illness-death model parameter estimates. As expected, the parameter estimates for transitions to death from other causes and death after recurrence are very similar between the two model fits. Where we expect differences is in the estimation for the transition from the baseline state (a grouping of the cured and non-cured state) and the recurrence state. We may expect covariate effects for this transition to be a combination of the covariate effects on the probability of cure and the recurrence rate in the non-cured subjects from the multistate cure model. We first note that some parameter estimates, particularly the covariate effect of HPV status on the transition to recurrence, have a lot of uncertainty. Additionally, we consider the gender effect. The multistate cure model fit to these data suggests a possible effect of gender on the recurrence rate in the non-cured group and a possible effect in the other direction for the probability of cure. In the illness-death model fit, it appears that these two effects cancel out to produce no effect of gender on the recurrence rate. This provides an example of how the multistate cure model can provide more granular inference that may not be attainable from the standard illness-death model. For both the multistate cure model and illness-death models, we use the imputed datasets and parameter estimates to obtain estimates of the Akaike Information Criterion (AIC). The AIC for the multistate cure model (63 parameters) was 10536.98, and the AIC for the illness-death model (48 parameters) was 11876.85, indicating that the multistate cure model provides a superior fit to these data.

Figure 4.9: Illness-Death Model Fit to Head and Neck Cancer Data



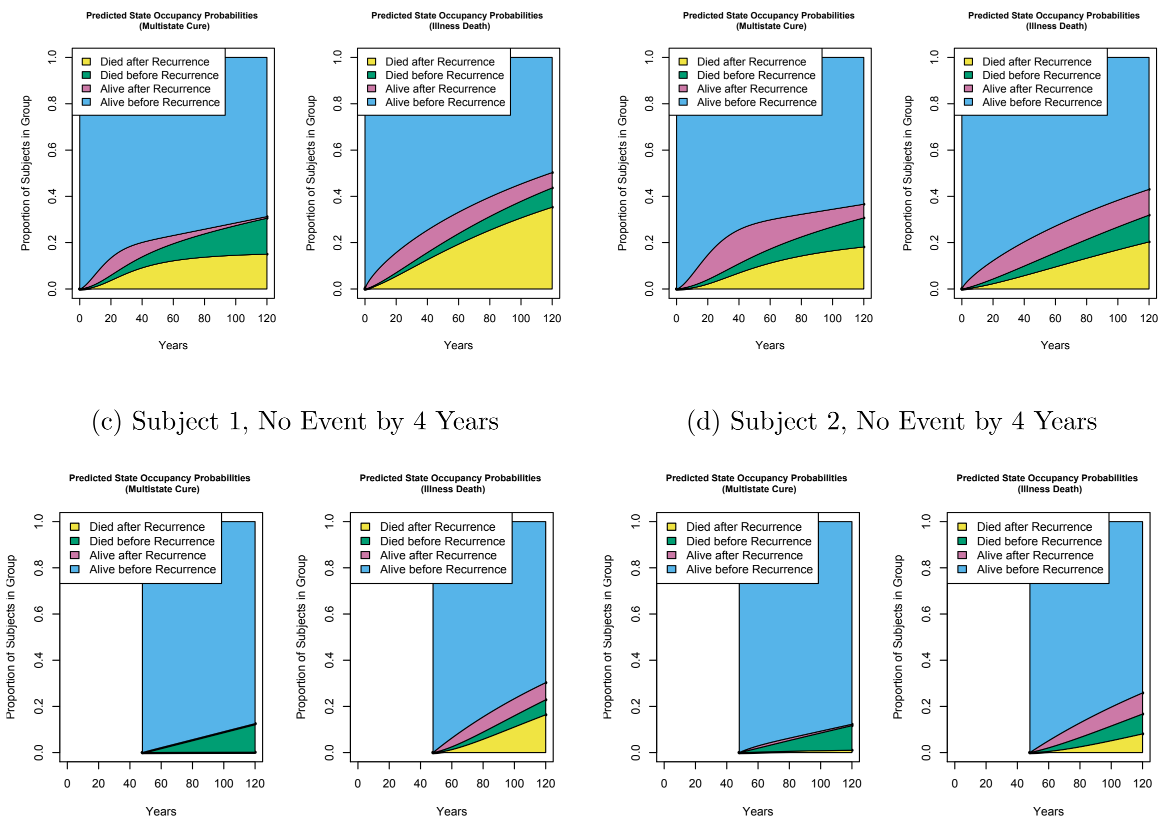
Additionally, the multistate cure model fit can produce very different subject-specific predictions compared to the illness-death model. Both models can provide the probabilities of being in different states (alive without recurrence, alive with recurrence, dead before recurrence, dead after recurrence) at a given time t. We compare the predicted probabilities for two example patients in **Figure 4.10**. For the first patient, the predicted probabilities given baseline covariates are very different, and for the second patient they are quite similar. For each patient, we also show the predicted probabilities if we also assume that subject was known to be event-free at 4 years. For both subjects, the predicted probabilities look quite different when we incorporate post-baseline follow-up. In particular, the predicted probability of having a recurrence given no recurrence or death by 4 years is much smaller for the multistate cure model. This is because this model weights the probability of recurrence by the probability that the subject is non-cured given the observed data. When we have observed that the subject has had no events by time t, we have greater evidence that the subject is cured, and the recurrence rate is therefore shrunk toward zero as time t increases. This is not the case for the illness-death model. These two example patients provide an illustration of the advantages of using the multistate cure model over the illness-death model for this setting in terms of prediction, particularly when incorporating post-baseline follow-up.

Figure 4.10: Comparing State Occupancy Probabilities

Subject 1 Covariates: Female, HPV +, 46 years old, Stage III, Mild Comorbidities, Current Smoker, Oral Cavity subsite
Subject 2 Covariates: Female, HPV -, 67 years old, Stage I, Moderate Comorbidities, Never Smoker, Oral Cavity subsite

(a) Subject 1, Baseline

(b) Subject 2, Baseline



(c) Subject 1, No Event by 4 Years

(d) Subject 2, No Event by 4 Years

4.9 Discussion

In the study of cancer, multistate cure models can be used to identify factors related to the rate of cancer recurrence, the rate of death before and after recurrence, and the probability of being cured by initial treatment. Additionally, multistate cure models can be very useful for prediction. However, the previous method for fitting multistate cure models requires substantial custom programming, making multistate cure models less accessible to analysts. We are interested in developing methods for fitting multistate cure models that can be implemented more easily and can incorporate different modeling assumptions into the fitting procedure.

In this chapter, we proposed an Expectation-Maximization (EM) algorithm for fitting the multistate cure model using maximum likelihood. The proposed algorithm can be fit using standard software, can incorporate either parametric or nonparametric baseline hazards for the state transition rates, and can integrate parameter restrictions for the transitions to death from other causes for cured and non-cured subjects. We then propose a Monte-Carlo EM (MCEM) Algorithm for fitting the multistate cure model in the presence of covariate missingness and/or unequal follow-up of the two outcomes, and we provide some software.

In simulations, the proposed EM and MCEM algorithms demonstrate good bias and coverage properties when the modeling assumptions are sufficiently restrictive. Additionally, we can still see good model fitting performance when we include more covariates in the model. When the $1 \to 4$ and $2 \to 4$ restrictions are relaxed (particularly when the baseline hazards are not restricted to be equal), we can run into numerical problems in some settings, suggesting that care should be taken to make reasonably restrictive modeling assumptions for these transitions. Additional exploration (not shown) suggests that these numerical issues stem from identifiability issues related to the $1 \to 4$ and $2 \to 4$ model parameters that occur when we have weaker restrictions for the two transitions. When applying the multistate cure model to a particular dataset, we recommend fitting a model under different assumptions about the $1 \to 4$ and $2 \to 4$ transitions and evaluating convergence properties and goodness of fit to determine if the $1 \to 4$ and $2 \to 4$ hazard restrictions can be reasonably relaxed. See Conlon et al. (2013) for a discussion of goodness of fit diagnostics for multistate cure models. We explore a shrinkage-based method to help guard against numerical issues related to relaxing the baseline hazard assumptions in **Appendix J**.

We applied the proposed MCEM algorithm to study cancer recurrence and death rates in subjects with head and neck cancer. Higher age, worse comorbidities, and increased smoking were associated with higher rates of death from other causes for both cured and non-cured subjects. Increased smoking was also associated with higher rates of death after recurrence, and larynx subsite was associated with lower rates of death after recurrence. HPV positive subjects had significantly lower rates of recurrence and higher rates of cure. Higher cancer stage was significantly associated with all transition rates and the probability of being cured.

Given the various types of imputation required, the relative advantages of the MCEM algorithm over the Bayesian MCMC algorithm should be considered. One disadvantage of the Bayesian approach (even if we apply our proposed imputation methods) is that we require accept-reject methods to draw the parameters, which involves careful consideration of parameter tuning, acceptance rates, and mixing. In contrast, the M Step of the MCEM algorithm is very simple to perform. As a result, the MCEM algorithm can perform estimation more quickly than the Bayesian MCMC. However, unlike with Bayesian approaches, the standard errors of θ using MCEM are not readily available. Additionally, the Bayesian approach allows the user to more directly incorporate prior assumptions into the estimation. The relative merits of the two approaches may depend on the data and the experience of the analyst.

As illustrated by the head and neck cancer example, multistate cure models offer investigators with an extremely useful tool for identifying factors involved in different parts of the disease process, and they can be used for prediction for future patients and in medical decision-making. Additionally, multistate cure models can be applied in a variety of settings and are certainly not limited to the study of cancer. In this chapter, we developed methods to make multistate cure models easier to fit in practice. Previous work focused on the setting with fully parametric baseline hazards, and our proposed methods allow us to choose parametric or non-parametric baselines and incorporate different assumptions about the transitions to death from other causes. The novel imputation-based approach for dealing with unequal follow-up proposed as part of the MCEM algorithm is not specific to the multistate cure model setting, and it can be applied in general semi-competing risks settings. Additionally, we propose a novel approach for obtaining standard errors for an MCEM algorithm, which can be applied in other MCEM settings. The proposed methods provide a convenient estimation method for fitting the multistate cure model and increased flexibility in model specification over existing methods.

Software

Software in the form of an R package called *MultiCure* is available on GitHub at `https://github.com/lbeesleyBIOSTAT`. This package provides functions for fitting the multistate cure model via the proposed EM and Monte Carlo EM algorithms and for estimating corresponding standard errors. This package also includes applications for estimating the derived state occupancy probabilities. *MultiCure* includes several vignettes describing how to use the software. Example code is also provided on GitHub. The package is also available on request from the corresponding author (lbeesley@umich.edu).

Chapter V

Comparison of Selection and Shrinkage Strategies for a Multistate Model of Head and Neck Cancer

5.1 Introduction

In medical applications, multistate models describe the rates at which individuals move between various health states. A common model considers the times it takes subjects to move between healthy, cancer recurrence, and death states. Multistate models have many valuable uses in medical research. Firstly, multistate models allow us to incorporate information from multiple event time outcomes (e.g. time to recurrence, time to death) in a unified way. These models are well-suited to handle issues of competing risks and recurrent events. Secondly, multistate models allow us to study which patient characteristics are relevant to which aspects of disease progression. In a model incorporating both recurrence and death times, we can identify factors related to time to recurrence and time to death with and without recurrence. These models provide appealing interpretations in terms of the disease process. Finally, multistate models are useful for making predictions for new patients based on their individual characteristics, which can be incredibly valuable for medical decision-making.

One challenge for using multistate modeling in practice is related to the large number of parameters. Multistate models often contain many component models (which we will call submodels). With even a modest number of covariates in each submodel, the multistate cure model as a whole can quickly end up with a very large number of parameters. We can easily run into issues of overfitting and other numerical issues when fitting such large models in practice. As such, we would like to explore variable selection/shrinkage methods to improve estimation for multistate models with a large number of parameters.

Many methods have been developed to deal with issues of variable selection in both the frequentist and Bayesian context. Bayesian methods usually approach selection/shrinkage through the specification of the prior distributions. Some common examples include spike and slab priors and horseshoe priors (George and McCulloch, 1993, 1997; Carvalho et al., 2009). Popular frequentist methods include ridge and LASSO penalization (Hoerl and Kennard, 1970).

While many variable selection/shrinkage methods have been developed, it is not known how these methods will perform in the multistate modeling context. In particular, multistate models often involve incorporating the same covariate in multiple places in the model, resulting in highly correlated (in fact, perfectly correlated) predictors in the multistate model as a whole. To our knowledge, this setting has not be explored in the literature. In this chapter, **we are interested in comparing how various existing variable selection and shrinkage strategies perform in a particular multistate modeling setting**. The explored methods, however, can be applied in other multistate modeling settings.

We focus our attention on the head and neck cancer data explored previously. For these data, we define the time to recurrence as the time from initial treatment to the time at which the tumor becomes observable. However, for some subjects with particularly aggressive tumors, initial treatment never reduces the size of the tumor to a point at which it is not observable. We call subjects that never appear to clear their initial cancer through treatment "persistent."

At the time of treatment, we do not know which patients are persistent and which are not. However, imaging and other medical diagnostics shortly after treatment (for example, in the following weeks) can be used to determine if patients have persistent disease. In the head and neck cancer data, we classified subjects with observable tumor within one month of treatment as being persistent, and this amounted to about 10% of subjects. The time period used to define persistence may vary from application to application, but we may expect that recurrences observed very soon after treatment are evidence of persistence disease, since it takes time for an unobservable tumor to grow to become observable.

Our interest is in modeling time to cancer recurrence and death for the head and neck data when some subjects are persistent. We may expect persistent subjects to behave differently than non-persistent subjects in terms of overall survival. Additionally, persistent subjects cannot experience a primary recurrence after initial treatment, because their cancer never appeared to go away. For the head and neck dataset, we also have subjects who were cured of their initial disease by treatment. These cured subjects, too, would be expected to have a different survival rate than the other subjects, and they also cannot experience recurrence. When developing a model of recurrence and death in such a heterogeneous population, it seems intuitive to separate out these different subgroups of subjects within the model for recurrence and death.

In this chapter, we develop a generalization of the multistate cure model that can account for both cured and persistent subpopulations. The multistate cure model in Conlon et al. (2013) and **Chapter IV** of this dissertation consists of cured and non-cured baseline states and models the recurrence and death rates. In our proposed model, we break the population into three non-overlapping baseline states: never appearing to clear their cancer (persistent), cured of their cancer, and appearing to clear their cancer but will eventually recur (non-cured/non-persistent). For the cured and persistent subjects, the only possible event is death. Non-cured/non-persistent subjects can experience recurrence and/or death.

This model formulation is useful for a variety of reasons. Firstly, the model allows us to study survival dynamics separately for the persistent and non-persistent subjects. Additionally, we can identify covariates related to the rate of persistence. This allows for more granular study of covariate effects on different parts of the disease process. Secondly, by separating out the persistent subjects from the non-cured/non-persistent subjects, we can directly model recurrence in the group of interest (the subjects that can have a recurrence). Thirdly, it is useful to be able to incorporate persistence information into predictions for new patients, which may allow us to generate improved predictions.

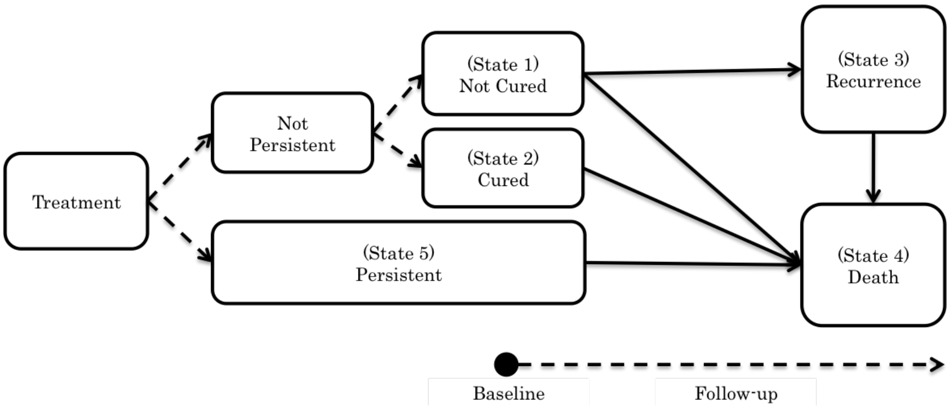
Parameter estimation for the proposed multistate model can be straightforward when the covariate set is not too large. Bayesian methods in Conlon et al. (2013) can be easily extended to accommodate the modified model formulation, and the EM and MCEM algorithms in **Chapter IV** can also be easily modified to include estimation of the added parameters.

In this chapter, we are particularly interested in the setting where the number of covariates in X_i is moderate to large. In this setting, estimation methods presented in Conlon et al. (2013) and **Chapter IV** can begin to break down. In this chapter, we compare various parameter estimation methods that incorporate variable selection and/or shrinkage to help improve the estimation for this particular multistate model setting. We first describe how we can perform Bayesian estimation incorporating modifications of standard Bayesian variable selection methods. We then propose an extension of the EM and MCEM algorithms in **Chapter IV** that can fit the proposed model and incorporate parameter shrinkage. We apply the proposed estimation methods to the head and neck cancer data, and we compare the resulting inference across the variables selection and shrinkage strategies.

5.2 Model Formulation and Parameter Estimation

In **Chapter IV**, we developed a multistate cure model. Here, we present a generalization of the multistate cure model incorporating persistence. The population is broken into three baseline states: persistent, cured, and non-persistent/non-cured. **Figure 5.1** shows the multistate model structure. Solid arrows represent possible state transitions. We note that only the non-cured/non-persistent subjects can experience a recurrence. For the non-persistent subjects, the model is identical to the multistate cure model in **Chapter IV** with a logistic regression for the probability of being non-cured (in this setting, given non-persistent) and Cox regression models for the state transition rates. We add a Cox regression model for the death rate in the persistent subjects and a logistic regression for the probability of persistence.

Figure 5.1: Diagram of Multistate Cure Model with Persistence



Let G_i be a categorical variable indicating the baseline state. As before, let $G_i = 0$ indicate cured. For persistent subjects, define $G_i = 2$, and define $G_i = 1$ for non-cured/non-persistent subjects. All subjects are known to be either $G_i = 2$ or $G_i \neq 2$. For subjects with observed recurrences, we know $G_i = 1$. Non-persistent subjects without an observed recurrence will have unknown G_i. Let X_i be a set of covariates, and for the sake of simplicity, we will assume that X_i is used as the set of predictors for all regressions. For now, we will assume that the covariates are fully observed and that we do not have unequal censoring. We will also assume that recurrence time is not included in the model from recurrence to death, although the proposed methods can be easily adapted to allow for this. We will further assume that we have parametric baseline hazards for each one of the transitions. We define $T_d, Y_{ir}, \delta_{ir}, Y_{id}$, and δ_{id} as in **Chapter IV**. For non-cured, non-persistent subjects, T_r is the underlying recurrence time. For persistent and cured subjects, T_r is defined as infinity. **Using notation developed in Chapter IV**, we have the following models:

$$\begin{aligned}
\text{logit}(P(G_i = 2|X_i)) =&\omega_0 + \omega_1^T X_i \\
\text{logit}(P(G_i = 1|X_i, G_i \neq 2)) =&\alpha_0 + \alpha_1^T X_i \\
\lambda_{13}(t) =&\lambda_{13}^0(t) \exp(\beta_{13}^T X_i) \\
\lambda_{14}(t) =&\lambda_{14}^0(t) \exp(\beta_{14}^T X_i) = \lambda_{24}(t) \\
\lambda_{34}(t - T_{ir}) =&\mathbb{I}(t > T_{ir})\lambda_{34}^0(t - T_{ir}) \exp(\beta_{34}^T X_i) \\
\lambda_{54}(t) =&\lambda_{54}^0(t) \exp(\beta_{54}^T X_i)
\end{aligned}$$

We call these different Cox and logistic regression models the "submodels," which together form the multistate model of interest. Rearranging the logistic regressions, we have that

$$\begin{aligned}
P(G_i = 2|X_i) =&\frac{e^{\omega_0+\omega_1^T X_i}}{1 + e^{\omega_0+\omega_1^T X_i}} \\
P(G_i = 1|X_i) =&\frac{1}{1 + e^{\omega_0+\omega_1^T X_i}} \times \frac{e^{\alpha_0+\alpha_1^T X_i}}{1 + e^{\alpha_0+\alpha_1^T X_i}} \\
P(G_i = 0|X_i) =&\frac{1}{1 + e^{\omega_0+\omega_1^T X_i}} \times \frac{1}{1 + e^{\alpha_0+\alpha_1^T X_i}}
\end{aligned}$$

We will use $S_j(t)$ to denote the probability of remaining in state j for time t. Let $\Lambda_{jk}(t)$ denote the cumulative hazard for the $j \to k$ transition. The multistate modeling assumptions result in the following complete data likelihood:

$$\begin{aligned}
L^{(com)} =&\prod_{i=1}^{n} \left[\frac{e^{\omega_0+\omega_1^T X_i}}{1 + e^{\omega_0+\omega_1^T X_i}} S_5(Y_{id})\lambda_{54}(Y_{id})^{\delta_{id}}\right]^{\mathbb{I}(G_i=2)} \left[\frac{1}{1 + e^{\omega_0+\omega_1^T X_i}}\right]^{\mathbb{I}(G_i\neq 2)} \\
&\times \left[\frac{1}{1 + e^{\alpha_0+\alpha_1^T X_i}} S_2(Y_{id})\lambda_{24}(Y_{id})^{\delta_{id}}\right]^{\mathbb{I}(G_i=0)} \\
&\times \left[\frac{e^{\alpha_0+\alpha_1^T X_i}}{1 + e^{\alpha_0+\alpha_1^T X_i}} S_1(Y_{id})\lambda_{14}(Y_{id})^{\delta_{id}}\right]^{\mathbb{I}(G_i=1,\delta_{ir}=0)} \\
&\times \left[\frac{e^{\alpha_0+\alpha_1^T X_i}}{1 + e^{\alpha_0+\alpha_1^T X_i}} S_1(Y_{ir})\lambda_{13}(Y_{ir}) S_3(Y_{id} - Y_{ir})\lambda_{34}(Y_{id} - Y_{ir})^{\delta_{id}}\right]^{\mathbb{I}(G_i=1,\delta_{ir}=1)}
\end{aligned}$$

The observed data likelihood is as follows:

$$
\begin{aligned}
L^{(obs)} = & \prod_{i=1}^{n} \left[\frac{e^{\omega_0+\omega_1^T X_i}}{1+e^{\omega_0+\omega_1^T X_i}} S_5(Y_{id}) \lambda_{54}(Y_{id})^{\delta_{id}} \right]^{\mathbb{I}(G_i=2)} \left[\frac{1}{1+e^{\omega_0+\omega_1^T X_i}} \right]^{\mathbb{I}(G_i\neq 2)} \\
& \times \left[\frac{1}{1+e^{\alpha_0+\alpha_1^T X_i}} \left\{ S_2(Y_{id})\lambda_{24}(Y_{id})^{\delta_{id}} + e^{\alpha_0+\alpha_1^T X_i} S_1(Y_{id})\lambda_{14}(Y_{id})^{\delta_{id}} \right\} \right]^{\mathbb{I}(\delta_{ir}=0, G_i\neq 2)} \\
& \times \left[\frac{e^{\alpha_0+\alpha_1^T X_i}}{1+e^{\alpha_0+\alpha_1^T X_i}} S_1(Y_{ir})\lambda_{13}(Y_{ir}) S_3(Y_{id}-Y_{ir})\lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}} \right]^{\mathbb{I}(\delta_{ir}=1, G_i\neq 2)}
\end{aligned}
$$

We notice that the complete data likelihood is very similar to the complete data likelihood for the multistate cure model in **Chapter IV**, and the terms coming from the added persistence category (terms involving ω, $S_5(t)$, and $\lambda_{54}(t)$) can be separated multiplicatively in both likelihoods (complete and observed) from the terms coming from the original multistate cure model. Under distinctness of the parameters, we can estimate ω and β_{54} separately from the parameters in the multistate cure model. Additionally, we can estimate ω and the 5 → 4 failure time parameters separately from each other. This separability property makes parameter estimation easily handled once we have methods for fitting the standard multistate cure model.

Assuming X_i is not too large and we have no covariate missingness, we can estimate the parameters $\beta_{13}, \beta_{24}, \beta_{34}$, and α by fitting a multistate cure model to the data excluding the persistent subjects using the methods described in **Chapter IV**. We can then estimate ω by fitting a standard logistic regression using the outcome $\mathbb{I}(G_i = 2)$, which is known for all subjects. Similarly, we can estimate the 5 → 4 failure time parameters by fitting a Cox regression model for (Y_d, δ_d) directly on the subjects with $G_i = 2$. If we do have covariate missingness, we can incorporate the estimation of ω and 5 → 4 parameters into an MCEM algorithm very similar to the algorithm proposed in **Chapter IV** for the standard multistate cure model. This will allow the covariate imputation to take advantage of observed covariates for both the persistent and non-persistent subjects. The imputation step of the MCEM algorithm will be exactly the same except, if we use SMC-FCS (Bartlett et al., 2014) to do the covariate imputation, we will want to use the likelihood for the proposed multistate model with persistence rather than the standard multistate cure model. The M-Step of the MCEM algorithm can be easily modified to include two additional regression model fits (logistic regression for $\mathbb{I}(G_i = 2)$ and Cox regression for survival in persistent subjects) given the most recent imputed covariates.

Like the maximum likelihood methods proposed in this paper, the Bayesian methods in Conlon et al. (2013) can also be easily extended to accommodate the modified model formulation.

In this chapter, we are particularly interested in the setting where the number of covariates in X_i is moderate to large. In this setting, estimation methods presented in Conlon et al. (2013) and **Chapter IV** can begin to break down. Through the estimation, we are interested in making inference about parameter Θ, which consists of α, ω, β, and baseline hazard parameters. Suppose that X_i contains p elements. Then, Θ would contain at least $6p+2$ parameters in addition to baseline hazard parameters. Even with moderate p, this can quickly result in a large number of parameters to estimate, which can lead to issues of overfitting and numerical problems. In this chapter, we will describe two estimation methods that incorporate variable selection/shrinkage to improve estimation.

5.3 Estimation using Bayesian Variable Selection

In this section, we explore Bayesian methods for parameter estimation that incorporate variable selection and/or shrinkage. Bayesian methods usually approach selection/shrinkage through the specification of the prior distributions of parameters corresponding to the covariate effects. We will first describe several common prior specifications in the literature, and we will describe how these priors can be easily applied to multistate modeling in general and our proposed multistate model in particular.

5.3.1 Common Priors

Suppose θ is a subset of parameters Θ on which we want to perform some sort of selection/shrinkage. In our problem, θ may consist of parameters related to the covariates in each submodel component of the multistate model. Let θ_k be a single element of θ.

We accomplish this variable selection/shrinkage by putting a prior on θ_k that will shrink it toward 0 (nearly or exactly zero) when we determine that the corresponding covariate is not "important." We consider three popular formulations of the prior distribution of θ_k.

Horseshoe Priors

The horseshoe prior is described in Carvalho et al. (2009) and has gained a lot of popularity in the variable selection literature. The prior takes the following form:

$$f(\theta_k|\gamma_k) = N(0, \lambda_k^2\sigma^2)$$
$$f(\lambda_k) = Cauchy^+(0,1)$$

where σ^2 is a tuning parameter with smaller values corresponding to greater shrinkage. Carvalho et al. (2009) suggests using $\sigma^2 = 1$, but a hyperprior can also be used. $Cauchy^+$ indicates the half-Cauchy distribution. This prior will strongly shrink weak signals to zero while allowing very large signals to remain large (little shrinkage). This prior has proven very useful in the setting of sparse signals. This is called a "horseshoe" prior due to the horseshoe shape of the density of κ (a function of λ_k), which is a measure of the amount of shrinkage. Under this prior, estimation is straightforward and can proceed using a

Gibbs Sampler in which each parameter is drawn one-by-one in an iterative algorithm. Each parameter draw can be performed using Metropolis-Hastings methods.

Mixture of Normals Priors

In this prior formulation, we assume that θ_k is generated from a mixture of normal distributions, one of which has the majority of its mass very close to 0 (called the spike) and one of which is a more diffuse normal distribution centered around zero (called the slab). We define a set of binary latent variables, γ, such that γ_k takes the value one if the covariate associated with θ_k should be included in the model. γ_k describes the component of the mixture distribution generating θ_k, where $\gamma_k = 0$ corresponds to the tight normal distribution around zero. The prior takes the following form:

$$\begin{aligned} f(\theta_k|\gamma_k) &= (1-\gamma_k)N(0,v_0) + \gamma_k N(0,v_1) \\ f(\gamma_k) &= p_k^{\gamma_k}(1-p_k)^{1-\gamma_k} \end{aligned}$$

v_0 is taken to be a value near 0 resulting in a tight normal distribution around zero (the spike) and v_1 is taken to be a value larger than v_0 corresponding to the more diffuse normal distribution. For now, we will treat v_0 and v_1 as constants, but we can put hyperpriors on their values. We can think of p_k as the prior probability that we should include the covariate associated with θ_k in the model. Other formulations of this mixture of normals prior exist (including different hyperpriors, different variance structures, etc). As with the horseshoe prior, estimation is fairly straightforward. We treat the elements of γ as parameters and draw their values within a “Stochastic Search” Gibbs sampling algorithm along with other parameter values (George and McCulloch, 1993, 1997). Each individual parameter draw can be performed using a Metropolis-Hastings draw. More details about this prior can be found in George and McCulloch (1993), George and McCulloch (1997), and Ishwaran and Rao (2005).

Point Mass at Zero Priors

In this formulation, we assume that θ_k is either from a normal distribution (called the slab) or is exactly zero (called the spike). We can then imagine θ_k is generated from a mixture of a normal distribution and a distribution with point mass at zero. We again define indicator γ_k which takes value 1 when θ_k is nonzero. We model

$$f(\theta_k|\gamma_k) = (1-\gamma_k)I(\theta_k = 0) + \gamma_k N(0, v_1)$$

In Newcombe et al. (2017), v_1 is treated as a hyperparameter with its own hyperprior. In this paper, we will assume v_1 is pre-specified.

Several different models for γ have been used in the literature, but they tend to directly model the sum of the elements of γ, which corresponds to the number of covariates to be included in the model. Hastie and Green (2012) suggests using a truncated Poisson or negative binomial prior. Newcombe et al. (2017) suggests a prior for the sum of γ involving the beta-binomial distribution. We might also consider a simple binomial distribution. We will use a Bernoulli prior for γ_k. Assuming prior independence across γ, this results in a binomial prior for the sum of γ as follows

$$\begin{aligned} f(\gamma_k) &= p^{\gamma_k}(1-p)^{1-\gamma_k} \\ f(\sum \gamma_k) &\propto \prod_k p^{\sum \gamma_k}(1-p)^{d-\sum \gamma_k} \end{aligned}$$

where d is the length of γ, and p is the probability of inclusion for each covariate, assumed to be equal for all k.

Estimation is more challenging under this prior than the other two. Unlike the other priors, θ_k is restricted to be exactly zero with probability one when $\gamma_k = 0$. This in effect is saying that the covariate corresponding to θ_k should not be included in the model, and it reduces the size of the parameter set, θ. When we go from $\theta_k = 0$ to a nonzero value of θ_k (or vice versa), we are essentially changing the dimension of the parameter set. In order to have good operating characteristics and to properly account for changes in model dimension, a reversible jump algorithm (Green, 1995) to obtain draws of γ and θ at each iteration of the multistate cure model fitting algorithm. Without going into too many details here, reversible jump is a MCMC algorithm which allows us to go explore models with different numbers of parameters/dimension. At a given iteration of the parameter sampling algorithm, we can decide to keep the parameter set from the previous iteration or to change the set of parameters included in the model. When no change in dimension is made, parameter updating then comes from usual Metropolis-Hastings draws. When we propose a change the dimension of the parameter, however, we need to modify the Metropolis-Hastings proposal distribution to account for the changing dimension of the parameter set. Rather than drawing individual elements of γ and θ, we draw the entire vectors γ and θ jointly. More details about reversible jump can be found in Newcombe et al. (2017), Troughton and Godsill (1997), and Green (1995). We discuss how to implement reversible jump in our setting in **Section 5.3.3**.

5.3.2 Applying the Priors to Our Model

When applying the standard Bayesian variable selection priors in our setting, several issues arise. One distinguishing feature of multistate models is that it consists of many submodels, each of which may include the same set of covariates. Therefore, the same covariate may appear in many different parts of the model. This presents several challenges. Firstly, for the point mass at zero prior, we could draw the entire vector γ jointly, or we could break γ up into the components corresponding to each one of the submodels. Restated, we could perform variable selection for the entire model as a whole or for each submodel separately. We propose performing variable selection for each submodel separately. This allows us to avoid drawing multiple elements of γ corresponding to the same covariate at once. Additionally, this will allow us to explore the model space more easily as at each iteration of the reversible jump fitting algorithm, we consider small changes to each submodel rather than a single small change to the entire model.

Bayesian variable selection methods can have problems with autocorrelation, poor mixing, and can spread posterior weight across many very similar models when we have many highly correlated predictors (Chipman et al., 2001). In the multistate modeling context in which the same covariates appear multiple places in the model, we have perfectly correlated predictors. For usual multistate models with fully observed outcome data (up to censoring), this may not present much of a concern since, for these models, the parameters in each submodel can often be separated in the observed data likelihood. This suggests that inclusion/exclusion of a variable in one submodel should not impact the inclusion/exclusion of a variable in another submodel. In the presence of missing data, however, the observed data likelihood may not be separable. In our setting, G_i is unknown for non-recurrent/non-persistent subjects. We have the following observed data likelihood:

$$
\begin{aligned}
L^{(obs)} = & \prod_{i=1}^{n} \left[\frac{e^{\omega_0+\omega_1^T X_i}}{1+e^{\omega_0+\omega_1^T X_i}} S_5(Y_{id}) \lambda_{54}(Y_{id})^{\delta_{id}} \right]^{\mathbb{I}(G_i=2)} \left[\frac{1}{1+e^{\omega_0+\omega_1^T X_i}} \right]^{\mathbb{I}(G_i\neq 2)} \\
& \times \left[\frac{1}{1+e^{\alpha_0+\alpha_1^T X_i}} \left\{ S_2(Y_{id}) \lambda_{24}(Y_{id})^{\delta_{id}} + e^{\alpha_0+\alpha_1^T X_i} S_1(Y_{id}) \lambda_{14}(Y_{id})^{\delta_{id}} \right\} \right]^{\mathbb{I}(\delta_{ir}=0, G_i\neq 2)} \\
& \times \left[\frac{e^{\alpha_0+\alpha_1^T X_i}}{1+e^{\alpha_0+\alpha_1^T X_i}} S_1(Y_{ir}) \lambda_{13}(Y_{ir}) S_3(Y_{id}-Y_{ir}) \lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}} \right]^{\mathbb{I}(\delta_{ir}=1, G_i\neq 2)}
\end{aligned}
$$

In the above likelihood, we cannot separate out α from the parameters in $S_2(t)$ and $S_1(t)$. This suggests that there is the potential for correlation of parameter estimates across submodels and, consequently, the inclusion and exclusion of covariates across submodels. In an extreme case, we could have that a covariate bounces back and forth between being included in each of two submodels. In our experience, this has not been too much of a problem, but it is worth consideration. Some work has been done in the literature to explore variable selection when we have highly correlated predictors through dilution priors, but we will not explore these here (George, 2010).

We may also want to apply restrictions to which covariates can and cannot be included in the model together. It is common to include categorical covariates that enter the model through a set of dummy variables. We would like to define our priors such that a group of related dummy variables (e.g. dummies representing cancer stage) are included or excluded from the model jointly. Note that this issue only arises for the two spike and slab prior formulations as the horseshoe prior does not perform variable selection. This problem is known as "grouped" variable selection in the literature, and it has been explored by many authors in the context of Bayesian variable selection (George and McCulloch, 1997; Farcomeni, 2010). The methods involve breaking up the covariate set into groups of covariates to be included/excluded jointly. We then replace γ with a vector representing inclusion of each *group* of covariates into the model. We then define the spike and slab priors in terms of this new set of latent indicators and model the individual group inclusion indicators rather than the individual elements of γ. This allows us to ensure that our variable selection algorithm is not exploring unintuitive models such as a model that includes only a dummy variable for cancer stage 2 but not stages 3 and 4 (with stage 1 as the reference).

5.3.3 Estimation

Under the horseshoe and mixture of normal priors, estimation is fairly straightforward. As we mentioned earlier, the proposed multistate cure model with persistence is just a multistate cure model with two additional regressions. As such, we can easily modify the Bayesian MCMC algorithm proposed for the multistate cure model in Conlon et al. (2013) for our generalized multistate model with persistence and substitute either a horseshoe or a mixture of normals prior for each element of θ. For the point mass at zero prior, however, we cannot modify the existing methods so easily.

Rather than the usual Gibbs sampling algorithm, we use a reversible jump algorithm to jointly draw the elements of γ (or a version reflecting the covariate groups). We can use the following algorithm.

Reversible Jump Algorithm

At each iteration of the model fitting process, we will:

(1) Use standard Metropolis-Hastings method to draw the baseline hazard parameters and the intercepts of the logistic regressions. Standard priors without variable selection are used for these parameters.

(2) We perform a parameter drawing step for each of the six submodels separately (regressions for transitions 1 → 4, 1 → 3, 3 → 4, 5 → 4 and the two logistic regressions). Let γ_s and θ_s represent the parts of γ and θ corresponding to the covariates in the s^{th} submodel. For each of the submodels, we perform the following:

(2a) Draw candidate γ_s^* and θ_s^* from the reversible jump proposal distribution $q(\gamma_s^*, \theta_s^*|\gamma_s, \theta_s)$ where γ_s and θ_s are the current values. We choose a common specification in the literature as follows: $q(\gamma_s^*, \theta_s^*|\gamma_s, \theta_s) = q_1(\theta_s^*|\gamma_s^*)q_2(\gamma_s^*|\gamma_s)$.

Given the current value of γ_s, we first draw γ_s^* using the following rules. At each iteration of the reversible jump algorithm, we can do one of the following moves: add a covariate into the model, remove a covariate, swap in a covariate for one already in the model, and keep the covariate set the same (called a null move). Note that we can incorporate grouping by performing the following for groups of covariates rather than individual covariates. At a particular iteration of the algorithm, we choose our move type as in **Table 5.1**.

Table 5.1: Proposal Distribution for γ_s^*

γ_s	Possible Moves	Probability
No covariates included	Add a covariate	1/6
	Null (keep the same)	5/6
All covariates included	Subtract a covariate	1/6
	Null (keep the same)	5/6
else	Add a covariate	1/6
	Subtract a covariate	1/6
	Swap covariates	1/6
	Null (keep the same)	1/2

For non-null moves, we then randomly select which covariate/s to move with equal probabilities among candidate covariates. Set $\theta_s^* = \theta_s$. If we subtract a covariate, we then set the corresponding elements of θ_s^* to zero. If we add a covariate, we draw from a normal distribution centered at the previous value to update the corresponding element of θ_s^*. For null moves, we re-draw all nonzero elements from normal distributions centered at the previous values.

(2b) Accept draw (γ_s^*, θ_s^*) with probability

$$\text{Prob(Accept Draw)} = \frac{P(Data|\gamma_s^*, \theta_s^*) f(\theta_s^*|\gamma_s^*) f(\gamma_s^*)}{P(Data|\gamma_s, \theta_s) f(\theta_s|\gamma_s) f(\gamma_s)} \times \frac{q(\gamma_s, \theta|\gamma_s^*, \theta_s^*)}{q(\gamma_s^*, \theta_s^*|\gamma_s, \theta_s)}$$

(2c) If the candidate move was Add, Subtract, or Swap, draw new parameter values for each of the covariates included in the model conditioning on the given model (no change in model dimension).

(3) Draw G for subjects with G unknown. We can use the same distribution as in Conlon et al. (2013).

5.3.4 Inference

After we run our MCMC algorithms to fit the model with variable selection through one of the spike and slab priors, we may take either of two general approaches when it comes to inference. The approach taken should depend on the motivations for performing variable selection.

Suppose we are very interested in identifying the "best" model according to some metric. This suggests that our goal is really to identify one or a few sets of covariates which together are most strongly associated with the outcome. This might be our goal if we are, for example, building a model for prediction purposes. In this case, we can determine which combination of variables (value of γ) has the highest posterior probability or choosing a model formulation including all covariates that have posterior inclusion probabilities (posterior probability that the corresponding element of γ equals 1) over a particular threshold (often 0.5). We then take the corresponding model formulation and use that as "the model" for inference.

Suppose instead that our goal is to make inference on individual parameters associated with covariates and identify individual covariates which seem to be "important." In this case, we are less interested in identifying the "best" model and more interested in getting a good sense of important covariates and their parameter values. In this case, we can use Bayesian model averaging, which makes inference about the model parameter by averaging across all the different values of vector γ (all the different covariate combinations) drawn within the MCMC algorithm (Hoeting et al., 1999). Under the horseshoe prior, we do not introduce the latent variables γ into the modeling framework, and inference proceeds directly using the posterior draws of θ as usual.

5.4 Maximum Likelihood Estimation with Shrinkage

As discussed earlier, we can easily adapt the EM and MCEM algorithms developed in **Chapter IV** to fit the multistate cure model with persistence. In this section, we describe how we can incorporate shrinkage (in particular, ridge penalties) into the estimation. We consider the following ridge-penalized complete data log-likelihood:

$$log(L^{(ridge)}(\theta)) = log(L^{(obs)}(\theta)) + \sum_{jk} M_{jk}||\beta_{jk}||_2^2 + M_\alpha||\alpha_1||_2^2 + M_\omega||\omega_1||_2^2$$

where jk indexes the possible state transitions and $M = (M_{jk}, M_\alpha, M_\omega)$ is the set of penalty tuning parameters. Here, $|| * ||_2^2$ represents the squared l^2 norm. We impose shrinkage on the covariate effects for each submodel, and we allow the shrinkage parameter to vary by submodel. We recall that the E-Step of the EM algorithm involves taking the expectation of the complete data log-likelihood with respect to the missing data. The imputation step of the MCEM algorithm involves drawing imputations of the missing data. Both of these steps condition on the most recent estimate of θ. Therefore, the expectation and imputation steps of the EM and MCEM algorithm will not be impacted by the ridge penalty given the current estimate of θ.

The penalty will impact the maximization step of the EM and MCEM algorithms. Here, we will focus on the EM algorithm. The MCEM algorithm is similar. Defining p_i as in (4.2) from **Chapter IV**, we obtain the following expected complete data log-likelihood:

$$\begin{aligned}
Q(\theta|\theta^{(t)}) = &\sum_{i=1}^{n}(1-p_i)\mathbb{I}(G_i \neq 2)\log\left[P(G_i = 0|G_i \neq 2)\right] \\
&+ p_i\mathbb{I}(G_i \neq 2)\log\left[P(G_i = 1|G_i \neq 2)\right] \\
&+ (1-p_i)\mathbb{I}(G_i \neq 2)\log\left[\lambda_{24}(Y_{id})^{\delta_{id}}\exp\{-\Lambda_{24}(Y_{id})\}\right] \\
&+ p_i\mathbb{I}(G_i \neq 2)\log\left[\lambda_{14}(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}(Y_{ir})\}\right] \\
&+ p_i\mathbb{I}(G_i \neq 2)\log\left[\lambda_{13}(Y_{ir})^{\delta_{ir}}\exp\{-\Lambda_{13}(Y_{ir})\}\right] \\
&+ \delta_{ir}\log\left[\lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}}\exp\{-\Lambda_{34}(Y_{id}-Y_{ir})\}\right] \\
&+ \mathbb{I}(G_i = 2)\log\left[\lambda_{54}(Y_{id})^{\delta_{id}}\exp\{-\Lambda_{54}(Y_{id})\}\right] \\
&+ \mathbb{I}(G_i \neq 2)\log\left[P(G_i \neq 2)\right] + \mathbb{I}(G_i = 2)\log\left[P(G_i = 2)\right]
\end{aligned}$$

$$+\sum_{jk} M_{jk}||\beta_{jk}||_2^2 + M_\alpha||\alpha_1||_2^2 + M_\omega||\omega_1||_2^2 \tag{5.1}$$

As before, the terms involving α, ω, and β separate, so we can maximize (5.1) with respect to α, ω, and β separately. The terms involving α resemble the log-likelihood for a logistic model with p_i as the outcome. We can estimate α by fitting a logistic regression to p_i (excluding the subjects with $G_i = 2$) and applying a ridge penalty to α_1. We can estimate ω by fitting a logistic regression to $\mathbb{I}(G_i = 2)$ and applying a ridge penalty to ω_1.

As in **Chapter IV**, we can perform the maximization for β by fitting a single survival model to an augmented version of the data. We consider an augmented version of the data that contains five rows for each subject (one for each transition in the multistate cure model). Each row contains a variable indicating the transition being considered (S), the time the subject was at risk for that transition (T), an indicator for whether the subject experienced that transition (D), a weight variable (W), and covariates (Z). **Table 5.2** shows the form of the rows in the augmented dataset for each subject i.

Table 5.2: Augmented Data Structure for Subject i

Transition	S	T	D	W	Z
$1 \rightarrow 3$	13	Y_{ir}	δ_{ir}	$p_i\mathbb{I}(G_i \neq 2)$	X_i
$2 \rightarrow 4$	24	Y_{id}	δ_{id}	$(1-p_i)\mathbb{I}(G_i \neq 2)$	X_i
$1 \rightarrow 4$	14	Y_{ir}	$\delta_{id}(1-\delta_{ir})$	$p_i\mathbb{I}(G_i \neq 2)$	X_i
$3 \rightarrow 4$	34	$Y_{id} - Y_{ir}$	δ_{id}	δ_{ir}	X_i
$5 \rightarrow 4$	54	Y_{id}	δ_{id}	$\mathbb{I}(G_i = 2)$	X_i

Using the augmented data structure, we can rewrite the last four terms in (5.1) as

$$\sum_{m=1}^{4n} W_m log\left(\left[\lambda^0_{S_m}(T_m)\exp\{g(Z_m, S_m; \beta)\}\right]^{D_m} \exp\{-\Lambda^0_{S_m}(T_m)\exp[g(Z_m, S_m; \beta)]\}\right) + \sum_{jk} M_{jk}||\beta_{jk}||_2^2$$

where $g(Z_m, S_m; \beta)$ is a function of Z_m and S_m that may include linear functions of Z_m and S_m along with interactions between Z_m and S_m. The sum in the above equation takes the form of a single weighted log-likelihood for a proportional hazards regression model with separate ridge penalties for each one of the β's.

A natural question is how we can choose the values for the tuning parameters M.

We choose the software default methods for choosing tuning parameters. We fit the ridge-penalized logistic regressions using the R function *glmnet*, and the penalty term is chosen through cross-validation. We estimate the failure time parameters in the M-Step using the *survreg* or *coxph* functions in the *survival* package in R. These functions use a method based on degrees of freedom to determine a "good" choice for the tuning parameters. We use this approach in our application of the proposed methods to the head and neck dataset. We can obtain standard errors using the same methods discussed in **Chapter IV**.

5.5 Predictions

We may also be interested in estimating the state occupancy probabilities. The state occupancy probabilities are the probabilities of being in each state of the multistate model at a given time t. We estimate these probabilities for various values of t to get a sense of the overall death and recurrence rates over time. First, we will provide expressions for estimating these probabilities conditioning only on baseline covariate information. Then, we will show how we can estimate these probabilities incorporating additional post-baseline follow-up. These probabilities rely on an estimate of the multistate model parameter, θ. We may choose to use the posterior mean of θ in the Bayesian estimation case or the MLE of θ in the maximum likelihood estimation case.

State Occupancy Probabilities Given Baseline Covariates

Let T_r and T_d denote the underlying event times for recurrence and death. For cured and persistent subjects, $T_r = \infty$. Additionally, note that persistence status is assumed to be unknown at baseline. We have the following:

$$\begin{aligned}
P(\text{Recurred and then died by time } t) =& P(T_r < T_d < t|X) \\
=& P(T_r < T_d < t|X, G = 1)P(G = 1|X) \\
P(\text{Alive at } t \text{ with prior recurrence}) =& P(T_r < t < T_d|X) \\
=& P(T_r < t < T_d|X, G = 1)P(G = 1|X) \\
P(\text{Alive at } t \text{ without prior recurrence}) =& P(T_r > t, T_d > t|X) \\
=& P(T_r > t, T_d > t|X, G = 1)P(G = 1|X) \\
&+ P(T_d > t|X, G = 0)P(G = 0|X) \\
&+ P(T_d > t|X, G = 2)P(G = 2|X) \\
P(\text{Died by time } t \text{ without prior recurrence}) =& P(T_d < T_r, T_d < t|X) \\
=& P(T_d < T_r, T_d < t|X, G = 1)P(G = 1|X) \\
&+ P(T_d < t|X, G = 0)P(G = 0|X) \\
&+ P(T_d < t|X, G = 2)P(G = 2|X)
\end{aligned}$$

We recall that, given $G \neq 2$, the proposed multistate model is identical to the multistate model developed in **Chapter IV**. Therefore, the form for probabilities conditioning on $G = 1$ or $G = 0$ will take the same form as for the standard multistate cure model. Define the following:

$$
\begin{aligned}
\pi_1(t) =& P(T_r < T_d < t|X, G = 1) = \int_0^t [1 - S_3(t-u)] \, \lambda_{13}(u) S_1(u) \, du \\
\pi_2(t) =& P(T_r < t < T_d|X, G = 1) = \int_0^t S_3(t-u) \lambda_{13}(u) S_1(u) \, du \\
\pi_3(t) =& P(T_r > t, T_d > t|X, G = 1) = S_1(t) \\
\pi_4(t) =& P(T_d < T_r, T_d < t|X, G = 1) = \int_0^t \lambda_{14}(u) S_1(u) \, du
\end{aligned}
$$

Then we have that

$$
\begin{aligned}
P(T_r < T_d < t|X) =& \pi_1(t) \frac{1}{1 + e^{\omega_0 + \omega_1^T X}} \times \frac{e^{\alpha_0 + \alpha_1^T X}}{1 + e^{\alpha_0 + \alpha_1^T X}} \\
P(T_r < t < T_d|X) =& \pi_2(t) \frac{1}{1 + e^{\omega_0 + \omega_1^T X}} \times \frac{e^{\alpha_0 + \alpha_1^T X}}{1 + e^{\alpha_0 + \alpha_1^T X}} \\
P(T_r > t, T_d > t|X) =& \pi_3(t) \frac{1}{1 + e^{\omega_0 + \omega_1^T X}} \times \frac{e^{\alpha_0 + \alpha_1^T X}}{1 + e^{\alpha_0 + \alpha_1^T X}} \\
&+ S_3(t) \frac{1}{1 + e^{\omega_0 + \omega_1^T X}} \times \frac{1}{1 + e^{\alpha_0 + \alpha_1^T X}} \\
&+ S_5(t) \frac{e^{\omega_0 + \omega_1^T X}}{1 + e^{\omega_0 + \omega_1^T X}} \\
P(T_d < T_r, T_d < t|X) =& \pi_4(t) \frac{1}{1 + e^{\omega_0 + \omega_1^T X}} \times \frac{e^{\alpha_0 + \alpha_1^T X}}{1 + e^{\alpha_0 + \alpha_1^T X}} \\
&+ (1 - S_3(t)) \frac{1}{1 + e^{\omega_0 + \omega_1^T X}} \times \frac{1}{1 + e^{\alpha_0 + \alpha_1^T X}} \\
&+ (1 - S_5(t)) \frac{e^{\omega_0 + \omega_1^T X}}{1 + e^{\omega_0 + \omega_1^T X}}
\end{aligned}
$$

State Occupancy Probabilities Given Post-Baseline Follow-up

Suppose we have followed a subject past baseline and want to use the observed outcome information up to the current time $t^* > 0$ to predict outcomes for $t > t^*$. We assume that we are making predictions at t^* such that baseline persistence status (yes/no) is known. Let t be greater than t^* and $s = t - t^*$.

Predictions Given Persistent

The only event for subjects that are persistent is death. We have

$$P(T_d > t|X, G = 2, T_d > t^*) = \frac{P(T_d > t|X, G = 2)}{P(T_d > t^*|X, G = 2)} = \frac{S_5(t)}{S_5(t^*)}$$

This is the probability that a subject is still alive at time t given that they are alive at $t^* < t$.

Predictions Given Not Persistent and Have Recurred

We now want to make predictions for a subject that is not persistent and has recurred by t^*. Suppose that the recurrence time is $0 < r^* \leq t^*$. After recurrence, the only possible event is death. For $t > t^*$, we estimate:

$$P(T_d > t|X, G = 1, T_d > t^*, T_r = r^*) = \frac{P(T_d > t|X, G = 1, T_r = r^*)}{P(T_d > t^*|X, G = 1, T_r = r^*)} = \frac{S_3(t - r^*)}{S_3(t^* - r^*)}$$

This is the probability that a subject is still alive at time t given that they are alive at $t^* < t$ and had a recurrence at time $r^* \leq t^*$.

Predictions Given Not Persistent and Have Not Recurred

Suppose that we know that a subject is alive, non-persistent, and has not recurred by time t^*, and we want to predict outcomes for $t > t^*$. If the subject is known to be non-persistent but has not had a recurrence, we do not know whether they are cured or not cured. Their observed outcome information up to time t^*, however, can inform how likely we think they are cured or not cured. Given the subject is not persistent, predictions come directly from the multistate cure model in **Chapter IV**. After simplifying, we have

$$\begin{aligned}
P(T_r < T_d < t|X, G \neq 2, T_d > t^*, T_r > t^*) =& \frac{[\pi_1(t) - \pi_1(t^*)]\, e^{\alpha_0 + \alpha_1^T X}}{\pi_3(t^*)e^{\alpha_0 + \alpha_1^T X} + S_2(t^*)} \\
P(T_r < t < T_d|X, G \neq 2, T_d > t^*, T_r > t^*) =& \frac{[\pi_2(t) - \pi_2(t^*)]\, e^{\alpha_0 + \alpha_1^T X}}{\pi_3(t^*)e^{\alpha_0 + \alpha_1^T X} + S_2(t^*)} \\
P(T_d > t, T_r > t|X, G \neq 2, T_d > t^*, T_r > t^*) =& \frac{\pi_3(t)e^{\alpha_0 + \alpha_1^T X} + S_2(t)}{\pi_3(t^*)e^{\alpha_0 + \alpha_1^T X} + S_2(t^*)} \\
P(T_d < T_r, T_d < t|X, G \neq 2, T_d > t^*, T_r > t^*) =& \frac{S_2(t^*) - S_2(t) + [\pi_4(t) - \pi_4(t^*)]\, e^{\alpha_0 + \alpha_1^T X}}{\pi_3(t^*)e^{\alpha_0 + \alpha_1^T X} + S_2(t^*)}
\end{aligned}$$

5.6 Application to Head and Neck Cancer Data

In this section, we apply the proposed Bayesian and maximum likelihood estimation methods to fit the proposed multistate cure model with persistence to a dataset of 1692 subjects with head and neck cancer, of which 173 have persistent disease at baseline. Given that we have discussed this dataset in detail earlier in this dissertation, we will omit the details about the dataset. The reference category for cancer site is Oral Cavity. Cancer stage was excluded from the model for death among the persistent subjects due a lack of subjects in the lower stages.

We recall that the head and neck cancer dataset had both covariate missingness and unequal censoring. For the Bayesian estimation, we add a step to the MCMC algorithm in which we impute the missing covariate and outcome values (this time, single imputation within each iteration of the MCMC) using the imputation methods discussed in **Chapter IV**. For imputing covariates, we use the SMC-FCS approach of Bartlett et al. (2014) to perform the imputation incorporating the multistate model structure into the imputation. We handle the missingness in a similar fashion within a MCEM algorithm when performing the maximum likelihood estimation. For the Bayesian estimation, methods in Zhang and Little (2011) could also be applied to incorporate variable selection in the approach for dealing with the missing covariates. In fitting the multistate model to the data, we assume that the β's for the transitions to death from other causes are equal. For the MCEM estimation, we further assume that the corresponding baseline hazards are also equal.

5.6.1 Bayesian Estimation

We apply the proposed Bayesian methods to estimate parameters from the multistate cure model with persistence for the head and neck dataset. We use four different prior specifications: no selection/shrinkage, horseshoe priors, mixture of normals priors, and point mass at zero priors. **Table 5.3** provides more details about the prior distributions used for the MCMC estimation.

For all variable selection/shrinkage priors, we perform selection separately for parameters in different submodels. For the mixture of normals and the point mass at zero priors, we impose a grouped structure on the inclusion/exclusion indicators. We define

Table 5.3: Prior Distributions

Parameter	Prior Distribution	Selection/Shrinkage Prior
$\log(\lambda)$'s	$N(0,4)$	-
ρ's	Γ(2.5,0.4)	-
α_0	$N(0,16)$	-
ω_0	$N(0,16)$	-
θ_k	$N(0,4)$	No Shrinkage
$\theta_k \| \lambda_k$	$N(0,\lambda_k^2)$	Horseshoe
λ_k	$Cauchy^+(0,1)$	Horseshoe
$\theta_k \| \gamma_g$	$(1-\gamma_g)N(0,0.1^2)+\gamma_g N(0,4)$	Mixture of Normals
$\theta_k \| \gamma_g$	$(1-\gamma_g)\mathbb{I}(\theta_k=0)+\gamma_g N(0,4)$	Point Mass at Zero
γ_g	$Bernoulli(0.5)$	Point Mass at Zero, Mixture of Normals

*where θ_k is an element of β_{13}, β_{24}, β_{34}, $\beta_5 4$, α_1, or ω_1. γ_g represents that inclusion indicator for the parameter group containing θ_k parameter, where the groups are defined separately for each transition.

a group as a set of covariates that should be included/excluded jointly, and the same set of covariates are treated as separate groups in different submodels. Following Carvalho et al. (2009) and Carvalho et al. (2010), we choose $\sigma^2 = 1$ in the horseshoe prior. For both spike and slab priors, we choose $v_1 = 4$, which corresponds to the variance of the "slab" part of the distribution. In determining the prior variance of the "spike" distribution for the mixture of normals prior, we chose a value to represent effect sizes that can "safely" be replaced by zero (following George and McCulloch (1993) and Chipman et al. (2001)). We may often view odds ratios or hazard ratios between 0.9 and 1.1 to represent very small effect sizes (in terms of practical significance), and these correspond to roughly a change in 0.1 on the log scale, which we used as our choice for v_0. A prior inclusion probability of 0.5 was chosen to allow many covariates to be included in the model but still incorporate selection. We performed some minor sensitivity analysis to the choice of the prior inclusion probability, and we did not see much impact on inference. We note that the choice of the spike and slab priors results in inclusion/exclusion of covariates in a group jointly. The horseshoe prior we use does not impose any such grouping restriction, and covariates in the same group may have different amounts of shrinkage.

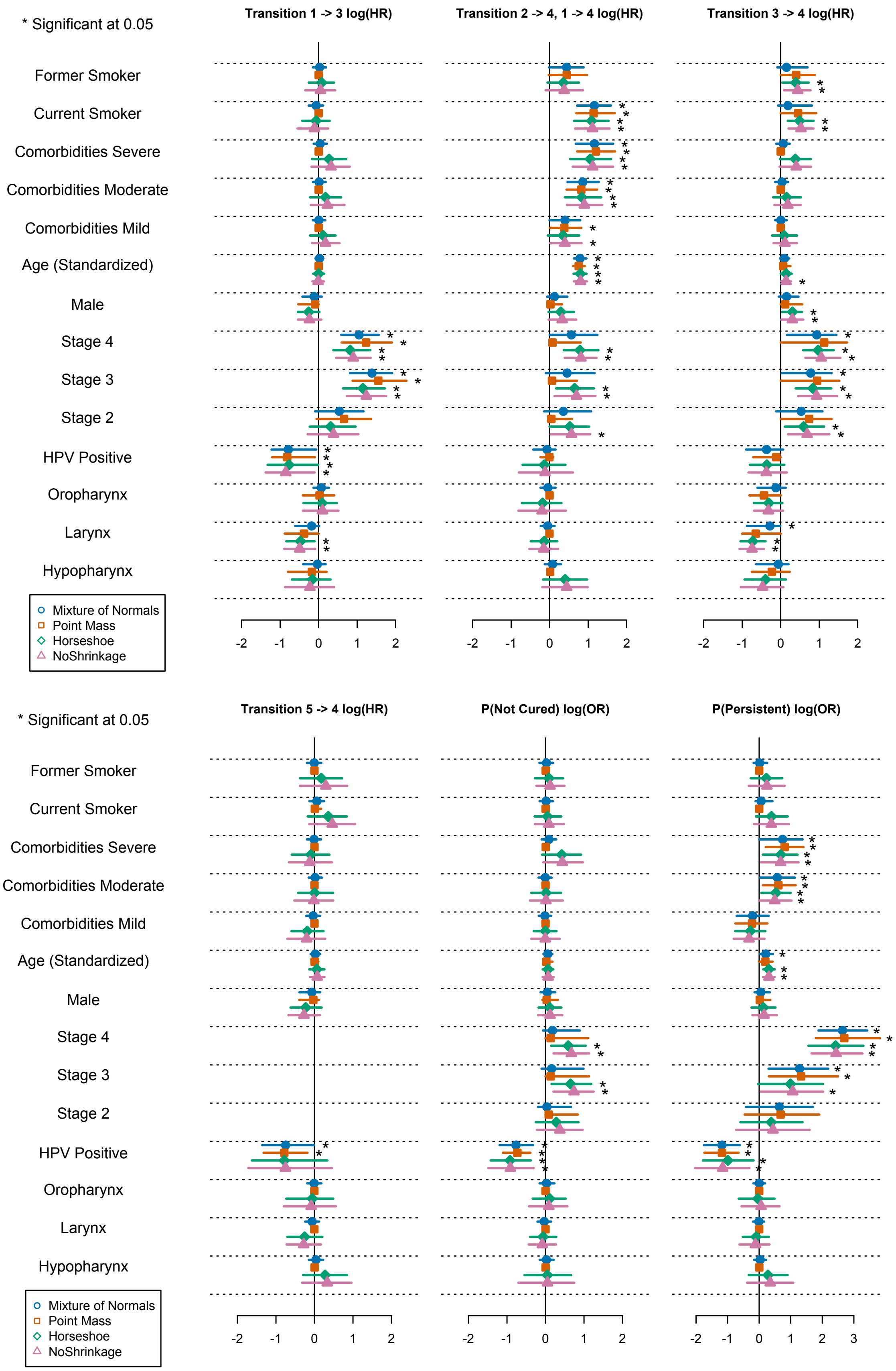
We run the Bayesian MCMC algorithm under each of the four prior specifications for 10,000 iterations, with the first 1000 iterations as burn-in. **Figure 5.2** shows the resulting posterior means and credible intervals using Bayesian model averaging. Recall, in Bayesian model averaging, we make inference using all draws of the parameter θ across iterations of the MCMC algorithm, which represents the posterior distribution of θ averaging across different model specifications. Credible intervals are determined based on posterior quantiles.

The four specifications of the prior distributions tend to give similar results with some exceptions. Generally, the two spike and slab priors (the point mass at zero and the mixture of normals priors) tend to have similar posterior means. Both priors result in strong shrinkage towards zero for covariates determined to be unimportant. One notable difference between the two priors is in the stage effects for the transitions to death from other causes. The point mass at zero prior suggests that there is not significant stage effect for this transition, while the mixture of normals prior, while not significant, does seem to suggest some relationship between stage and death from other causes. When both spike and slab methods determine that a covariate is unimportant (with strong shrinkage toward zero), the credible interval for the mixture of normals prior is wider than for the point mass at zero. This is a result of the fact that the point mass prior allows the parameter to be exactly zero while the mixture of normals prior assigns a value close to zero when a covariate is determined to be unimportant.

The horseshoe and no shrinkage fits give very similar results to each other except that the horseshoe prior tends to have slightly narrower credible intervals. One situation in which these two priors differ from the spike and slab priors is in the stage effect in the logistic regression for the probability of being non-cured given non-persistent. The horseshoe and no shrinkage priors indicate a significant effect of stage, while the other priors do not. However, the credible intervals for the spike and slab priors are large, so the results are not contradictory. Another difference is in the effect of severe (ACE27) comorbidities on the transition to death after recurrence ($3 \to 4$). The spike and slab priors indicate no effect, while the other priors suggest an increased transition rate compared to subjects with no comorbidities.

Generally, we estimate a significant effect of HPV status and stage on the transition to recurrence given not cured, where the HPV negative and higher stage subjects have higher transition rates to recurrence. Larynx cancer site may also be associated with a lower rate of recurrence compared to oral cavity. Higher stage, higher age, worse comorbidities, and increased smoking were related to higher rates of death from other causes in the non-persistent subjects. We may not expect cancer stage to be related to death from other causes, and this significant effect may be due to unmeasured confounding.

Figure 5.2: Multistate Model Fit using Bayesian Estimation



Higher cancer stage, male gender, and increased smoking were all generally associated with higher rates of death after recurrence. We also observe that subjects with larynx cancer had lower rates of death after recurrence compared to subjects with oral cavity cancer.

We observe that HPV positivity may be associated with lower rates of death among persistent subjects. HPV positivity was associated with higher rates of cure among the non-persistent subjects. Higher cancer stage was associated with lower rates of cure among non-persistent subjects. Higher comorbidities, older age, higher cancer stage, and HPV negativity were all associated with higher rates of persistence.

In **Appendix K**, we explore the posterior inclusion probabilities for the spike and slab priors, we compare the 5-year overall survival predictions across the four models, and we present the correlations of the drawn values for the γ inclusion indicators within and between submodels.

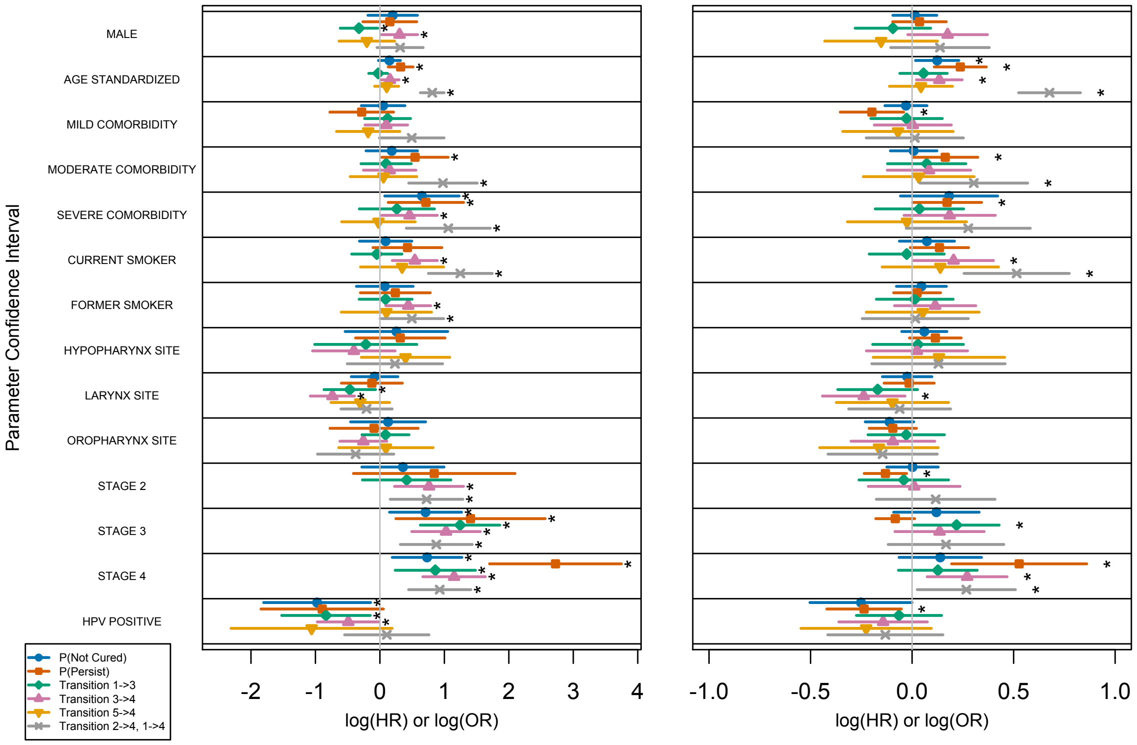
5.6.2 Maximum Likelihood Estimation

We apply the MCEM algorithm from **Chapter IV** incorporating two additional regressions to fit our proposed multistate model. We obtain two model fits. In the first fit, we perform estimation without imposing any parameter shrinkage. In the second fit, we impose ridge shrinkage on all parameters in β, ω_1, and α_1. In both cases, we run the MCEM for 100 iterations, and we use the Rubin's rules-based approach discussed in **Chapter IV** for variance estimation.

Figure 5.3 presents the results. We observe some very wide confidence intervals in the fit without any shrinkage. As expected, the inference for the part of the model not involving state 5 is very similar to the multistate cure model fit to the head and neck data in **Chapter IV**. However, the fit is not identical. This is because we are incorporating additional covariate information from the persistent subjects to do covariate imputation for the non-persistent subjects, which results in similar but not identical parameter estimates. In contrast, the magnitudes of the estimated effect sizes under ridge shrinkage are very different than in the fit without shrinkage. We note that the x-axis is different for the fit with ridge shrinkage. Nearly all of the parameter estimates are strongly shrunk towards zero. However, we still see some very intuitive covariates showing up as significant in the various submodels. For example, higher age, worse smoking status, and

Figure 5.3: Multistate Model Fit using Maximum Likelihood Estimation

(a) No Shrinkage (b) Ridge Shrinkage



Note: x-axis differs between plots

worse comorbidities are associated with higher rates of death from other causes.

There are some instances in which parameters are significant for the ridge fit and not for the fit without shrinkage. The most striking differences are in the model for the probability of persistence. For example, cancer stages 2 and 3 appear to be related to reduced probabilities of persistence compared to stage 1 in the ridge shrinkage fit. We see the opposite associations in the fit without shrinkage, and the opposite association is more intuitive.

The MLE-based fit without shrinkage and the Bayes-based fit without shrinkage are very similar, and they identify very similar or the same sets of covariates as significant for each submodel as shown in **Figure K.10** of **Appendix K**. In **Appendix K**, we compare the 5-year overall survival predictions for the two fits without shrinkage and for the ridge-penalized fit.

5.7 Discussion

In this paper, we propose a multistate model for time to recurrence and death that incorporates information about cancer-persistent and cured subpopulations after initial treatment. This model is a generalization of the multistate cure model with two additional regression models for the probability of being persistence and the rate of death in the persistent subjects. Our interest is in fitting the proposed model to a study of recurrence and death for patients with head and neck cancer. However, estimation presents some challenge.

As is common with large multistate models, the proposed multistate model has many model parameters, and we are interested in exploring variable selection/shrinkage methods for performing parameter estimation. One challenge for applying existing variable selection/shrinkage methods in the multistate modeling setting is that the same predictors can appear in multiple places in the model. It is known that highly correlated predictors can create problems in variable selection (George, 2010), and it is unclear how existing methods will perform for large multistate models. We therefore seek to compare the performance of various existing variable selection and shrinkage methods in our particular multistate model setting with the eventual goal of making inference for the head and neck cancer data. Our results can be used to guide estimation in other multistate modeling settings.

We consider two general strategies for estimation: 1) Bayesian estimation with three different prior distribution specifications (two spike and slab priors and the horseshoe) and 2) maximum likelihood estimation via EM and Monte Carlo EM algorithms with ridge penalization. With some small modifications to account for covariate grouping and submodel-specific selection, we apply the existing Bayesian methods to fit the proposed multistate model to the head and neck dataset. We also develop and apply methods to perform ridge penalization for the proposed multistate model. We then compare inference across the different estimation procedures.

We find that the two spike and slab priors considered (mixture of normals and point mass at zero priors) result in generally similar (Bayesian model averaged) credible intervals for the head and neck data. Greater differences between the two spike and slab priors can be seen in the estimated posterior inclusion probabilities. Additionally, we consider the number of iterations in which different combinations of covariates (different model formations) are chosen for each submodel in **Appendix K**. The model formulation with the highest posterior probability was different for many of the transitions.

If the goal is to obtain the Bayesian model-averaged credible intervals, we may recommend using the mixture of normals prior over the point mass at zero prior as the two priors resulted in similar inferences and the fit under a point mass at zero prior is much more difficult to implement (it involves a reversible jump algorithm). However, if the goal is to obtain the model with the highest posterior probability or using all predictors with a posterior inclusion probability over a particular threshold, the chosen prior may impact the resulting model, and we would recommend applying both priors and comparing the results.

In our example, the horseshoe prior obtained inference similar to the Bayesian fit without shrinkage but with slightly narrower credible intervals and is similar but somewhat different to the fits using spike and slab priors. However, the amount of shrinkage imposed by the horseshoe prior depends on the hyperparameters, and greater differences and similarities between the various fits may be seen with different values for the horseshoe hyperparameters. Both the horseshoe and the mixture of normals priors are straightforward to implement. The mixture of normals has a natural way to determine the values of hyperparameters based on how we determine “meaningful” effect sizes, while the specification of the hyperparameter for the horseshoe distribution is less clear. We could have specified hyperpriors for the hyperparameters to reduce the dependence on the choice of hyperparameters, but we ultimately chose not to do this for the current analysis for the sake of simplicity. Some sensitivity analysis was performed to evaluate the impact of the hyperparameters on the model inference, and very little impact was seen for small to moderate changes in the hyperparameters. If the variable inclusion indicators are not of primary interest, we would recommend applying both horseshoe and the mixture of normals priors and comparing the model inference. This can give the analyst a sense of the robustness of the model inference to choices about the selection/shrinkage method.

Ridge regression applied to the head and neck dataset resulted in strong shrinkage of the model parameters compared to maximum likelihood estimation without shrinkage. Many of the strong significant associations from the fit without shrinkage were preserved in the fit with shrinkage, although the effect sizes were attenuated. Like in the Bayesian approach, the ridge penalization procedure involves specification of tuning parameters, and these tuning parameters control the amount of shrinkage. In this analysis, we used software default methods for determining the tuning parameters, but other approaches could have been used. The ridge shrinkage approach has the attractive property of involving a small modification of the likelihood, and there is a large body of existing software for fitting model with ridge penalties. For usual multistate model formulations (without the missing data and latent variables), estimation often proceeds by fitting a single regression model to the data (as in *mstate* in R). Ridge penalization (along with other types of penalization such as LASSO) may often be easy to implement through the addition of ridge penalties to this regression model. However, we found that the parameters from ridge penalization resulted in different 5-year survival predictions compared to the other methods (**Appendix K**). In particular, the 5-year survival predictions seemed to be attenuated towards the population average. Therefore, if the goal is to obtain outcome predictions, we may recommend the Bayesian approach over ridge penalization. We note that the ridge penalization method involves using a single tuning parameter for each submodel. Therefore, different groups in the same submodel are given the same degree of shrinkage. Future work could explore a generalization of the ridge penalty that incorporates group-specific shrinkage.

In this chapter, we compared existing variable selection and shrinkage methods for a particular modeling setting and made some tentative recommendations for the application of these methods for general multistate modeling settings. Our evaluation was entirely based on an application of these methods to a particular head and neck dataset, and additional explorations with different datasets and model formulations are needed in order to determine how well these methods perform for general multistate models. Our results, however, provide some general intuition and *tentative* guidance for the application of these methods in other multistate modeling settings.

This study provides a thorough exploration of the proposed multistate model applied to the head and neck data. To our knowledge, this study is the first to explicitly model persistence for head and neck cancer, and it is therefore if great interest from a clinical point of view. By incorporating our study of persistence into the multistate model structure, we can use the model to inform the way we handle missing covariates. Our results suggest several predictors that may be related to the rate of persistence. Of note, we observe higher rates of persistence among HPV negative patients, patients with worse comorbidities, and older patients in addition to patients with higher cancer stage. It is well known that subjects with persistent disease have worse survival outcomes, and the results of this study may help to reveal predictors that can be used to identify patients at a higher risk of persistence for more careful observation after initial treatment. Future work could use the state occupancy probabilities derived in **Section 5.5** and the multistate model fits to obtain patient-specific predictions of the probability of persistence and other quantities of interest.

Chapter VI

Bayesian Variable Selection with Order Restrictions and Interactions

6.1 Introduction

In cancer modeling, we are often interested in including predictors for which there is a natural ordering of effect sizes. For example, we may expect subjects with higher cancer stage (e.g. AJCC stage III) to do worse in terms of overall survival compared to subjects with lower cancer stage (e.g. AJCC stage II). In the presence of strong prior beliefs about the order of effects, we can incorporate parameter order restrictions into the estimation procedure. By incorporating order restrictions, our goal is to improve efficiency for estimating the parameters of interest. Order restrictions can also ensure that nuisance parameters take reasonable values. For example, suppose our interest is in identifying a treatment effect on overall survival, and we are also adjusting for comorbidities. We may strongly believe that worse comorbidities would be related to similar or worse survival rates. Applying order restrictions to the parameters related to comorbidities may help us in determining the treatment effect.

In Bayesian estimation, these order restrictions are often imposed through prior distributions. Many authors have discussed methods for incorporating order restrictions into Bayesian parameter estimation (Dunson and Neelon, 2003; Gelfand et al., 1992). In this chapter, we suppose we are interested in performing variable selection under the order restrictions. Literature is relatively sparse in this setting. Kasim et al. (2012) and Otava et al. (2014) propose an approach for performing variable selection in the presence of order restrictions in the context of dose response modeling. Their approach generally involves specifying an order restriction including the possibility of equality of adjacent effects and performing variable selection to determine whether adjacent effects in the order restriction are equal or strictly ordered. This translates into variable selection for the predictors corresponding to these ordered effects. Their proposed approach incorporates variable selection through use of the variable selection prior proposed in Kuo and Mallick (1998). By incorporating variable selection, we can perform some regularization/shrinkage to help control the effective number of parameters.

We suppose we are also interested in incorporating interactions in the model. This can allow for increased flexibility in model specification. Many authors have discussed methods for performing variable selection in the presence of interactions. Variable selection is particularly important in the presence of interactions due to the large number of parameters. The methods for performing Bayesian variable selection with interactions usually involve defining heredity restrictions in which the interaction term is allowed to be nonzero only if one or both of the main effects are included in the model (Chipman, 1996; Farcomeni, 2010). Such heredity restrictions avoid model choices that include the interaction term without main effects. In this chapter, we will also refer to these heredity restrictions as "hierarchy constraints." Suppose we are interested in including interactions in which one or both of the interacted variables have order restrictions. This setting has not been explored in the Bayesian variable selection literature. Additional work is needed to explore how to incorporate both heredity and order restrictions into the variable selection procedure.

In this chapter, we develop methods for performing Bayesian variable selection with interactions incorporating both hierarchy constraints and (possibly two-way) order restrictions. In particular, we are interested in interactions between two categorical variables. The form of the proposed prior distribution depends on whether we impose ordering for one or both of the categorical variables in the interaction. We perform a simulation study to explore the performance of the proposed methods.

6.2 A Prior for Order-Restricted Selection with Interactions

In this section, we propose a prior distribution for incorporating order restrictions and hierarchy constraints for variable selection with interactions. First, we will clarify the types of order and hierarchy restrictions we are considering.

6.2.1 Order Restrictions

Let Y represent our outcome and A and B represent two categorical model predictors. Define both variables to take the value 1 for the reference category and integer values above 1 for all other categories. Let J be the number of possible values for A and K be the number of possible values for B.

Suppose we model Y using the linear model:

$$Y_{ijk} = \mu_{jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \tau^{-1}) \tag{6.1}$$

where μ_{jk} is the mean of Y in category $A = j, B = k$ and i indexes the subjects. Suppose we want to impose parameter restrictions such that μ_{jk} is nondecreasing with increasing values of B within categories of A. This is equivalent to the following:

$$\mu_{j1} \leq \mu_{j2} \leq \ldots \leq \mu_{jK} \qquad j = 1, \ldots, J$$

We will call this a *one-way restriction* because parameters are restricted in the B direction but not in the A direction. This type of order constraint is considered for main effects in Otava et al. (2014). Suppose we further want to impose order restrictions in the A direction such that μ_{jk} is also non-decreasing with increasing values of A within categories of B. This is equivalent to:

$$\begin{aligned} &\mu_{j1} \leq \mu_{j2} \leq \ldots \leq \mu_{jK} \qquad j = 1, \ldots, J \\ &\mu_{1k} \leq \mu_{2k} \leq \ldots \leq \mu_{Jk} \qquad k = 1, \ldots, K \end{aligned}$$

We will call this a *two-way restriction.*

Existing Method for Main Effects

In developing our own prior for performing selection with order restrictions, we will build off of a prior explored in Kasim et al. (2012), Otava et al. (2014), and Otava et al. (2017). This prior was developed for performing selection with order constraints for main effects variables. Suppose for now that A takes only one value, so the model in (6.1) is equivalent to a main effects only model for B. Suppose we want to impose the following (one-way) order restriction on B: $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_K$. Here, we suppress the index for A in the notation for simplicity.

Otava et al. (2014) considers a reparameterization of $\theta_k = \mu_{k+1} - \mu_k$ for $k = 1, \ldots, K-1$. We can re-write

$$\mu_k = \begin{cases} \mu_1 & k = 1 \\ \mu_1 + \sum_{t=1}^{k-1} \theta_t & k > 1 \end{cases}$$

The restrictions on μ are equivalent to the restrictions $\theta_k \geq 0$ for all $k = 1, \ldots, K-1$. Our goal is to impose the ordering on θ_k with added variable selection to determine whether we have equality or strict ordering between any μ_k and μ_{k+1}.

Define latent variables $\Gamma_k = \mathbb{I}(\theta_k \neq 0)$ and $\tilde{\theta}_k$ such that $\theta_k = \Gamma_k \tilde{\theta}_k$. We then re-write

$$\mu_k = \begin{cases} \mu_1 & k = 1 \\ \mu_1 + \sum_{t=1}^{k-1} \Gamma_t \tilde{\theta}_t & k > 1 \end{cases}$$

We now consider the model in terms of μ_1, Γ and $\tilde{\theta}$ rather than μ. The values of μ are just functions of μ_1, Γ and $\tilde{\theta}$. The prior in Otava et al. (2017) (an update of Otava et al. (2014)) can be written as

$$\begin{aligned} &\mu_1 \sim N(\eta_0, \tau_0^{-1}) \\ &\tilde{\theta}_k \sim N(\eta_k, \tau_k^{-1}) T(0, \infty) && k = 1, \ldots, K-1 \\ &\Gamma_k \sim Bernoulli(\pi_k) && k = 1, \ldots, K-1 \end{aligned}$$

with hyperpriors for the hyperparameters η_0, τ_0, η_k, τ_k, τ (from (6.1)), and π_k. Here, $T(a, b)$ indicates a distribution truncated at the left at a and the right at b. This prior makes uses of the variable selection prior in Kuo and Mallick (1998), which defines a parameter value (in this case $\tilde{\theta}_k$) that is then included or excluded from the model based on an inclusion indicator (in this case Γ_k). The prior for $\tilde{\theta}_k$ is truncated such that that the resulting values of θ and μ will satisfy the order restrictions.

6.2.2 Hierarchy Restrictions

Before discussing the heredity/hierarchy restrictions, we will first rewrite (6.1) in terms of the main effects and interactions of A and B. Let β be a set of parameters corresponding to the linear regression mean structure. We can write:

$$E(Y_i) = \beta_0 + \sum_{j=2}^{J} \beta_{Aj}\mathbb{I}(A=j) + \sum_{k=2}^{K} \beta_{Bk}\mathbb{I}(B=k) + \sum_{j=2}^{J}\sum_{k=2}^{K} \beta_{AjBk}\mathbb{I}(A=j, B=k)$$

This is a fully-saturated model, meaning that model has so many parameters that it can exactly fit the data. Through variable selection, our goal is to reduce the number of parameters. In terms of the original model parameters from (6.1), we have that

$$\begin{aligned}
\beta_0 &= \mu_{11} \\
\beta_{Aj} &= \mu_{j1} - \mu_{11} \\
\beta_{Bk} &= \mu_{1k} - \mu_{11} \\
\beta_{AjBk} &= \mu_{jk} - \mu_{1k} - \mu_{j1} + \mu_{11}
\end{aligned}$$

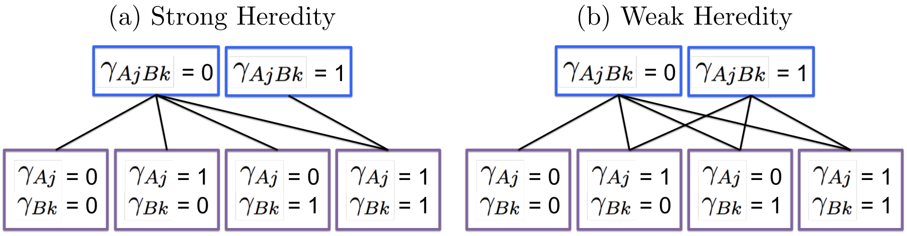
Let $\gamma_{Aj} = \mathbb{I}(\beta_{Aj} \neq 0)$, $\gamma_{Bk} = \mathbb{I}(\beta_{Bk} \neq 0)$, and $\gamma_{AjBk} = \mathbb{I}(\beta_{AjBk} \neq 0)$.

The heredity principle discussed in Chipman (1996) suggests making restrictions about the possible values of γ_{AjBk} given the values for γ_{Aj} and γ_{Bk}. Restated, the inclusion/exclusion of an interaction term depends on whether the corresponding main effects are included in the model. Weak heredity requires that at least one main effect is included in the model in order for the interaction term to be potentially included. Strong heredity requires both main effects to be included in order for the interaction term to be potentially included. **Figure 6.1** provides a visualization of the possible values of γ_{AjBk} given values for γ_{Aj} and γ_{Bk} under weak and strong heredity.

The heredity restrictions imply a hierarchy of variables for inclusion/exclusion, where the inclusion of interactions terms (the second level in the hierarchy) depends on the inclusion of main effects terms (the first level in the hierarchy). In this chapter, we will use “hierarchy constraints” to refer to heredity restrictions for the interaction terms.

Heredity constraints define the prior distribution of γ_{AjBk} based on the value of γ_{Aj}

Figure 6.1: Visualization of Heredity Restrictions



(a) Strong Heredity

(b) Weak Heredity

γ_{AjBk} = 0
γ_{AjBk} = 1
γ_{Aj} = 0 γ_{Bk} = 0
γ_{Aj} = 1 γ_{Bk} = 0
γ_{Aj} = 0 γ_{Bk} = 1
γ_{Aj} = 1 γ_{Bk} = 1
γ_{AjBk} = 0
γ_{AjBk} = 1
γ_{Aj} = 0 γ_{Bk} = 0
γ_{Aj} = 1 γ_{Bk} = 0
γ_{Aj} = 0 γ_{Bk} = 1
γ_{Aj} = 1 γ_{Bk} = 1

Connected boxes show the possible values of γ_{AjBk} given the values of γ_{Aj} and γ_{Bk} under different heredity constraints

and γ_{Bk} as follows:

$$P(\gamma_{AjBk} = 1) = \begin{cases} p_0 & \gamma_{Aj} = 0, \gamma_{Bk} = 0 \\ p_1 & \gamma_{Aj} = 1, \gamma_{Bk} = 0 \\ p_2 & \gamma_{Aj} = 0, \gamma_{Bk} = 1 \\ p_3 & \gamma_{Aj} = 1, \gamma_{Bk} = 1 \end{cases} \tag{6.2}$$

Weak hierarchy constraints require that at least one main effect is included in the model for the corresponding interaction to be included in the model (Chipman, 1996). This corresponds to an assumption that $p_0 = 0$. A strong hierarchy constraint will also require that $p_1 = p_2 = 0$. We will continue under the assumption of weak hierarchy constraints. However, we can approach strong hierarchy by choosing p_1 and p_2 small. If desired, we can restrict p_3 to be greater than both p_1 and p_2.

Table 6.4 explores what the weak hierarchy constraints on β imply in terms of constraints on μ. Note that the constraints on β are satisfied if and only if the constraints on μ are satisfied.

Table 6.1: Implications of Weak Hierarchy Constraints on μ

β Constraint	μ Constraint
$\beta_{Aj} = 0$	$\mu_{j1} = \mu_{11}$
$\beta_{Bk} = 0$	$\mu_{1k} = \mu_{11}$
$\beta_{AjBk} = 0$	$\mu_{jk} = \mu_{1k} + \mu_{j1} - \mu_{11}$

6.2.3 Incorporating Both Hierarchy and Order Constraints

In this section, we describe how we combine both types of constraints into a single modeling framework. In the setting where $J = K = 3$, we can imagine the following matrix for the mean of Y:

Table 6.2: Matrix of Means of Y

	$B = 1$	$B = 2$	$B = 3$
$A = 1$	μ_{11}	μ_{12}	μ_{13}
$A = 2$	μ_{21}	μ_{22}	μ_{23}
$A = 3$	μ_{31}	μ_{32}	μ_{33}

A one-way ordering restriction would require that the μ's be non-decreasing row-wise. A two-way ordering restriction would require that the μ's be non-decreasing row-wise and column-wise.

Developing the Notation

Similarly to the approach in Otava et al. (2014), we will define the differences between the values of μ within reference categories for the other variable (value $= 1$) as follows:

$$\begin{aligned} \theta_{Aj} &= \mu_{j+1,1} - \mu_{j1} \qquad & j = 1, \ldots, J-1 \\ \theta_{Bk} &= \mu_{1,k+1} - \mu_{1k} \qquad & k = 1, \ldots, K-1 \end{aligned}$$

We will denote the sets of parameters θ_A and θ_B respectively. Let $\theta_0 = \mu_{11}$. In Otava et al. (2014), indicators are introduced corresponding to whether each value of θ is nonzero. We introduce similar indicators defined as follows:

$$\begin{aligned} \Gamma_{Aj} &= \mathbb{I}(\theta_{Aj} \neq 0) \\ \Gamma_{Bk} &= \mathbb{I}(\theta_{Bk} \neq 0) \end{aligned}$$

The set of indicators Γ_A indicates whether the value of μ changes for consecutive values of A in the reference category of B. The set of indicators Γ_B indicates whether the value of μ changes for consecutive values of B in the reference category of A. Using the method of Kuo and Mallick (1998) and Otava et al. (2014), we define variables $\tilde{\theta}$ such that

$$\begin{aligned} \theta_{Aj} &= \Gamma_{Aj}\tilde{\theta}_{Aj} \\ \theta_{Bk} &= \Gamma_{Bk}\tilde{\theta}_{Bk} \end{aligned}$$

Here, we can view the set of parameters $\tilde{\theta}_A$ and $\tilde{\theta}_B$ as values of θ_A and θ_B without any selection to impose equality for adjacent categories. We define $\tilde{\theta}_0 = \theta_0$. To account for the inclusion/exclusion of the interaction terms, we define indicators

$$\Gamma_{AjBk} = \mathbb{I}(\mu_{jk} = \mu_{1k} + \mu_{j1} - \mu_{11}) = \mathbb{I}(\beta_{AjBk} \neq 0)$$

for $j > 1$ and $k > 1$. Γ_{AjBk} corresponds to whether the interaction term β_{AjBk} is nonzero. Denote the set of indicators Γ_{AjBk} as Γ_{AB}.

Define $\tilde{\theta}_{AjBk}$ such that $\mu_{jk} = \tilde{\theta}_{AjBk}$ when j and k are both greater than 1 and $\Gamma_{AjBk} = 1$. We can imagine $\tilde{\theta}_{AjBk}$ corresponds to the value of μ_{jk} when the hierarchy constraint is not imposed. Denote the set of $\tilde{\theta}_{AjBk}$ as $\tilde{\theta}_{AB}$.

We can re-write μ in terms of Γ and $\tilde{\theta}$ as follows:

$$\mu_{jk} = \begin{cases} \tilde{\theta}_0 & j = 1, k = 1 \\ \tilde{\theta}_0 + \sum_{s=1}^{j-1} \Gamma_{As}\tilde{\theta}_{As} & j > 1, k = 1 \\ \tilde{\theta}_0 + \sum_{t=1}^{k-1} \Gamma_{Bt}\tilde{\theta}_{Bt} & j = 1, k > 1 \\ \Gamma_{AjBk}\tilde{\theta}_{AjBk} + (1 - \Gamma_{AjBk})(\mu_{1k} + \mu_{j1} - \mu_{11}) & j > 1, k > 1 \end{cases} \tag{6.3}$$

Therefore, we can entirely re-parameterize our model for μ in terms of $\tilde{\theta} = (\tilde{\theta}_0, \tilde{\theta}_A, \tilde{\theta}_B, \tilde{\theta}_{AB})$ and $\Gamma = (\Gamma_A, \Gamma_B, \Gamma_{AB})$. We can re-write the hierarchy constraint from (6.2) in terms of Γ as follows:

$$P(\Gamma_{AjBk} = 1 | \Gamma_A, \Gamma_B, \tilde{\theta}_A, \tilde{\theta}_B, \tilde{\theta}_0) = \begin{cases} 0 & \sum_{s=1}^{j-1} \Gamma_{As} = 0, \quad \sum_{t=1}^{k-1} \Gamma_{Bt} = 0 \\ p_1 & \sum_{s=1}^{j-1} \Gamma_{As} > 0, \quad \sum_{t=1}^{k-1} \Gamma_{Bt} = 0 \\ p_2 & \sum_{s=1}^{j-1} \Gamma_{As} = 0, \quad \sum_{t=1}^{k-1} \Gamma_{Bt} > 0 \\ p_3 & \sum_{s=1}^{j-1} \Gamma_{As} > 0, \quad \sum_{t=1}^{k-1} \Gamma_{Bt} > 0 \end{cases} \tag{6.4}$$

We propose a modification of the usual hierarchy constraint (from (6.2) and (6.4)) for the order-restricted setting. There are some instances in which, in order to preserve the ordering, we must have an interaction between variables. For example, suppose we have $J = K = 3$ as in **Table 6.2** and we have current values of μ as follows:

Table 6.3: Example Values of μ

	$B = 1$	$B = 2$	$B = 3$
$A = 1$	$\mu_{11} = 0$	$\mu_{12} = 1$	$\mu_{13} = 2$
$A = 2$	$\mu_{21} = 3$	$\mu_{22} = 6$	$\mu_{23} =?$
$A = 3$	$\mu_{31} = 4$	$\mu_{32} =?$	$\mu_{33} =?$

If we are imposing (one- or two-way) order restrictions, we must have that $\mu_{22} \leq \mu_{23}$. When determining if $\Gamma_{A2B3} = 1$, what we are really determining is whether μ_{23} equals $\mu_{21} + \mu_{13} - \mu_{11}$. In this example, we have that $\mu_{21} + \mu_{13} - \mu_{11} = 3 + 2 - 0 = 5$. However, we know that $\mu_{23} \geq \mu_{22} = 6$. Therefore, it is necessary that μ_{23} is strictly greater than μ_{22} (so $\Gamma_{A2B3} = 1$) for the order restrictions to be satisfied.

Define $R_{jk} = \mathbb{I}(\mu_{j,k-1} > \mu_{1k} + \mu_{j1} - \mu_{11})$ for $j > 1, k > 1$ for one-way order restrictions and $R_{jk} = \mathbb{I}(\max(\mu_{j-1,k}, \mu_{j,k-1}) > \mu_{1k} + \mu_{j1} - \mu_{11})$ for two-way order restrictions. We propose the following modified hierarchy constraint:

$$P(\Gamma_{AjBk} = 1 | \Gamma_A, \Gamma_B, \tilde{\theta}_A, \tilde{\theta}_B, \tilde{\theta}_0) = \begin{cases} p_0 & \sum_{s=1}^{j-1} \Gamma_{As} = 0, \sum_{t=1}^{k-1} \Gamma_{Bt} = 0, R_{jk} = 0 \\ p_1 & \sum_{s=1}^{j-1} \Gamma_{As} > 0, \sum_{t=1}^{k-1} \Gamma_{Bt} = 0, R_{jk} = 0 \\ p_2 & \sum_{s=1}^{j-1} \Gamma_{As} = 0, \sum_{t=1}^{k-1} \Gamma_{Bt} > 0, R_{jk} = 0 \\ p_3 & \sum_{s=1}^{j-1} \Gamma_{As} > 0, \sum_{t=1}^{k-1} \Gamma_{Bt} > 0, R_{jk} = 0 \\ 1 & R_{jk} = 1 \end{cases} \quad (6.5)$$

Assuming $p_0 = 0$, the distribution in (6.5) will impose the hierarchy constraint while restricting the interaction to be present when it is needed in order to satisfy the order constraint.

Proposed Prior Distribution

We specify the variable selection prior for μ under the order and hierarchy constraints in terms of Γ and $\tilde{\theta}$.

$$\begin{aligned} f(\Gamma,\tilde{\theta}) &= f(\Gamma_A,\Gamma_B,\Gamma_{AB},\tilde{\theta}_0,\tilde{\theta}_A,\tilde{\theta}_B,\tilde{\theta}_{AB}) \\ &= f(\Gamma_{AB},\tilde{\theta}_{AB}|\tilde{\theta}_0,\Gamma_A,\tilde{\theta}_A,\Gamma_B,\tilde{\theta}_B)f(\Gamma_B,\tilde{\theta}_B|\tilde{\theta}_0,\Gamma_A,\tilde{\theta}_A)f(\Gamma_A,\tilde{\theta}_A|\tilde{\theta}_0)f(\tilde{\theta}_0) \end{aligned}$$

We will assume that the main effects parameters for A and B are a priori independent (so $(\Gamma_A,\tilde{\theta}_A)$ and $(\Gamma_B,\tilde{\theta}_B)$ are a priori independent). Following Kuo and Mallick (1998) and Otava et al. (2014), we will also assume that the two elements of each Γ and $\tilde{\theta}$ pair are a priori independent. This results in the following simplification

$$\begin{aligned} f(\Gamma,\tilde{\theta}) = f(\tilde{\theta}_{AB}|\tilde{\theta}_0,\Gamma_A,\tilde{\theta}_A,\Gamma_B,\tilde{\theta}_B)f(\Gamma_{AB}|\theta_0,\Gamma_A,\tilde{\theta}_A,\Gamma_B,\tilde{\theta}_B) \\ \times f(\Gamma_B|\tilde{\theta}_0)f(\tilde{\theta}_B|\tilde{\theta}_0)f(\tilde{\theta}_A|\tilde{\theta}_0)f(\Gamma_A|\tilde{\theta}_0)f(\tilde{\theta}_0) \end{aligned}$$

Intercept

We first specify the prior for $\tilde{\theta}_0$ as

$$\tilde{\theta}_0 \sim N(\eta_0,\tau_0^{-1})$$

where η_0 and τ_0 are hyperparameters. We will discuss hyperpriors later on.

Main effects of B

The priors for the main effects of B are

$$\begin{aligned} \Gamma_{Bk} &\sim Bernoulli(\pi_{Bk}) \\ \tilde{\theta}_{Bk} &\sim N(\eta_{Bk},\tau_{Bk}^{-1})T(0,\infty) \end{aligned}$$

This is the same prior as in Otava et al. (2017) for order-restricted variable selection.

Main Effects of A

The form of the prior for the main effects of A depends on whether we are imposing one- or two-way order restrictions. If we do not impose order restrictions in the A direction, we perform selection for the main effects of A using grouped selection (Chipman, 1996; Farcomeni, 2010). This means that we assume either all elements μ_{j1} are equal or unequal. We propose using grouped selection rather than separate selection for each main effect of A since we assume that A represents a single categorical variable taking different values. We could alternatively perform selection separately for each of the main effects of A. Using the grouped selection approach, we define $\Gamma_{A1} = \Gamma_{A2} = \ldots = \Gamma_{A,J-1}$ and use prior

$$\Gamma_{A1} \sim Bernoulli(\pi_{A1})$$
$$\tilde{\theta}_{Aj} \sim N(\eta_{Aj}, \tau_{Aj}^{-1})$$

In order to impose two-way order restrictions, we use the following prior

$$\Gamma_{Aj} \sim Bernoulli(\pi_{Aj})$$
$$\tilde{\theta}_{Aj} \sim N(\eta_{Aj}, \tau_{Aj}^{-1})T(0, \infty)$$

where $\tilde{\theta}_{Aj}$ is also restricted to be greater than zero as in the order-restricted selection prior in Otava et al. (2017).

Interaction Terms

We specify $f(\Gamma_{AB}|\Gamma_A, \Gamma_B)$ using a Bernoulli distribution with probability given by the hierarchy constraint prior in (6.5). Alternatively, we can perform the order-restricted selection without any sort of hierarchy constraint by setting $p_0 = p_1 = p_2 = p_3 > 0$. Now, we consider the specification of the prior for $\tilde{\theta}_{AB}$. We specify the prior for $j > 1$ and $k > 1$ as follows for one-way ordering:

$$\tilde{\theta}_{AjBk} \sim N(\mu_{1k} + \mu_{j1} - \mu_{11}, \tau_{AjBk}^{-1})T(\mu_{j,k-1}, \infty)$$

where the values of μ are functions of $\tilde{\theta}_A$, $\tilde{\theta}_B$, Γ_A, Γ_B, and $\tilde{\theta}_0$ as shown in (6.3). For two-way ordering, we have

$$\tilde{\theta}_{AjBk} \sim N(\mu_{1k} + \mu_{j1} - \mu_{11}, \tau_{AjBk}^{-1})T(\max(\mu_{j,k-1}, \mu_{j-1,k}), \infty)$$

Hyperparameters

The prior distributions depend on hyperparameters ϕ containing p_1, p_2, p_3, π_{Aj}, π_{Bk}, η_{Aj}, η_{Bk}, τ_{Aj}, τ_{Bk}, τ_{AjBk}, η_0, and τ_0. These hyperparameters can be pre-specified. However, we suggest using the following hyperpriors

$$\begin{aligned}
\pi_{Aj}, \pi_{Bk} &\sim U(0,1) \\
\eta_{Aj}, \eta_{Bk}, \eta_0 &\sim N(0, 10^a) \\
\tau_{Aj}, \tau_{Bk}, \tau_{AjBk}, \tau_0 &\sim Gamma(b, c) \\
p_1, p_2, p_3 &\sim U(0,1)
\end{aligned}$$

where a, b, and c are pre-specified constants. In our simulations, we choose $a = 2$, $b = 3$, and $c = 1$. These choices of hyperpriors are very similar to the hyperpriors used for order-restricted selection in Otava et al. (2017).

Posterior Distribution

We suppose our outcome of interest is the linear regression model

$$Y_{ijk} = \mu_{jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \tau^{-1})$$

We have the corresponding posterior distribution for Γ, $\tilde{\theta}$, and τ:

$$f(\Gamma, \tilde{\theta}, \tau | Y, A, B) \propto \prod_{i=1}^{n} f(Y_i | A_i, B_i; \Gamma, \tilde{\theta}, \tau) f(\Gamma, \tilde{\theta} | \phi) f(\tau) f(\phi)$$

The conditional posterior distributions for the elements of Γ are easy to derive since they are binary. We now consider the posterior distribution for $\tilde{\theta}$. We note that

$$\begin{aligned}
f(\tilde{\theta} | Y, A, B, \Gamma, \tau) \propto\ & f(Y | A, B; \Gamma, \tilde{\theta}, \tau) f(\tilde{\theta}_{AB} | \tilde{\theta}_0, \tilde{\theta}_A, \tilde{\theta}_B, \Gamma_A, \Gamma_B) f(\Gamma_{AB} | \tilde{\theta}_0, \tilde{\theta}_A, \tilde{\theta}_B, \Gamma_A, \Gamma_B) \\
& \times f(\tilde{\theta}_B | \tilde{\theta}_0) f(\tilde{\theta}_A | \tilde{\theta}_0) f(\tilde{\theta}_0)
\end{aligned}$$

These distributions are all normal, truncated normal, or Bernoulli. In the future, we will explore the posterior distributions for each parameter in more detail. For now, we will just mention that in the linear regression setting, the distributions take "nice" forms that should prove easy to sample from. In the meantime, we can implement our proposed prior distribution using the Gibbs sampling algorithm implemented in existing programs such as WinBUGS or JAGS.

6.3 Simulation Study

In this section, we perform a simulation study to explore the performance of our proposed methods in terms of posterior means, widths and coverage of credible intervals, and posterior inclusion probabilities. We compare the performance of our proposed methods with and without the hierarchy constraint. We call our approach the “collapsed” approach because it involves restricting neighboring values of μ to be equal (effectively collapsing two covariate values together). In our simulations, we denote the two collapsed approaches (without and without hierarchy restrictions) as *CollapsedHierarchy* and *CollapsedNoHierarchy*. We also compare our proposed methods to several other prior formulations including 1) order restriction with no variable selection (denoted *OrderNoSelection*), 2) hierarchy constrained-selection with no order restriction (denoted *Hierarchy*), and 3) no selection or order restrictions (denoted *None*).

6.3.1 Simulation Details

We perform simulations under two different outcome models as follows:

$$\begin{aligned}
&\text{Model 1: } Y_{ijk} = \mu_{jk} + e_{ijk}, \quad e_{ijk} \sim N(0, \tau^{-1}) \quad (\text{true } \tau = 1) \\
&\text{Model 2: } \text{logit}(P(Y_{ijk} = 1)) = \mu_{jk}
\end{aligned}$$

where μ_{jk} is a function of $E(Y)$ in category $A = j, B = k$ and i indexes the subjects. For each model, we simulate data under eight different values of μ. In each simulation setting, we simulate 200 datasets. We use 500 observations for Model 1 and 1000 observations for Model 2.

Table 6.4 presents the values of μ used for each of the eight simulation settings. The corresponding values for Γ and β are given in **Table 6.5**. The first four simulation settings correspond to main effect only models, and the last four models incorporate interactions. Setting 5 corresponds to situations in which weak hierarchy is violated. This means we have a nonzero interaction term with no corresponding main effects. All simulation settings satisfy a two-way order restriction across A and B.

Table 6.4: Values for μ in each Simulation Setting

Simulation Setting 1

	$B=1$	$B=2$	$B=3$
$A=1$	2	2	2
$A=2$	2	2	2
$A=3$	2	2	2

Simulation Setting 2

	$B=1$	$B=2$	$B=3$
$A=1$	2	2	3
$A=2$	2	2	3
$A=3$	3	3	4

Simulation Setting 3

	$B=1$	$B=2$	$B=3$
$A=1$	2	3	3
$A=2$	3	4	4
$A=3$	3	4	4

Simulation Setting 4

	$B=1$	$B=2$	$B=3$
$A=1$	2	3	4
$A=2$	3	4	5
$A=3$	4	5	6

Simulation Setting 5

	$B=1$	$B=2$	$B=3$
$A=1$	2	2	2
$A=2$	2	2	2
$A=3$	2	2	3

Simulation Setting 6

	$B=1$	$B=2$	$B=3$
$A=1$	2	2	3
$A=2$	2	2	4
$A=3$	3	4	5

Simulation Setting 7

	$B=1$	$B=2$	$B=3$
$A=1$	2	3	3
$A=2$	3	5	5
$A=3$	3	5	5

Simulation Setting 8

	$B=1$	$B=2$	$B=3$
$A=1$	2	3	4
$A=2$	3	5	6
$A=3$	4	6	7

Table 6.5: Corresponding Values for Γ and β

Setting	Γ_{A1}	Γ_{A2}	Γ_{B1}	Γ_{B2}	Γ_{A2B2}	Γ_{A2B3}	Γ_{A3B2}	Γ_{A3B3}
1	0	0	0	0	0	0	0	0
2	0	1	0	1	0	0	0	0
3	1	0	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0
5*	0	0	0	0	0	0	0	1
6	0	1	0	1	0	1	1	1
7	1	0	1	0	1	1	1	1
8	1	1	1	1	1	1	1	1

Setting	β_{A2}	β_{A3}	β_{B2}	β_{B3}	β_{A2B2}	β_{A2B3}	β_{A3B2}	β_{A3B3}
1	0	0	0	0	0	0	0	0
2	0	1	0	1	0	0	0	0
3	1	1	1	1	0	0	0	0
4	1	2	1	2	0	0	0	0
5*	0	0	0	0	0	0	0	1
6	0	1	0	1	0	1	1	1
7	1	1	1	1	1	1	1	1
8	1	2	1	2	1	1	1	1

*The hierarchy principle is violated.

As mentioned earlier, we compare our collapsed approaches to three other prior formulations. The prior *OrderNoSelection* involves imposing two-way order restrictions without any selection/shrinkage. This is done through specifying normal prior distributions for the μ_{jk} elements that are truncated to satisfy the order restrictions. The prior *None* uses normal priors for each μ_{jk} with no selection or order restrictions.

The prior *Hierarchy* involves variable selection without any order restriction. The prior involves seven indicators: one for whether the A main effects are nonzero (using grouped selection where the main effects of A are included/excluded jointly), one for whether the B main effects are nonzero (again using grouped selection), and one for each one of the four interaction terms. The priors for the inclusion of the interaction terms involve a weak hierarchy constraint. We would like to clarify what this prior assumes about the main effects. For example, we consider the main effects of A. This prior allows 1) all the main effects are equal to zero or 2) all the main effects are nonzero and unequal. In terms of μ, this prior allows 1) $\mu_{11} = \mu_{21} = \mu_{31}$ or 2) $\mu_{11} \neq \mu_{21} \neq \mu_{31}$ and $\mu_{11} \neq \mu_{31}$. Restated, we are using grouped selection to determine whether to include or exclude the main effects of A and B (Chipman, 1996; Farcomeni, 2010). In these simulations, we are assuming that A and B have some natural ordering, and it therefore makes sense to either include or exclude the main effects of a particular variable jointly.

In each simulation setting and model formulation, we use the program JAGS to fit the outcome model using each of five specifications of the prior distributions for Γ and $\tilde{\theta}$. For the methods assuming ordering, we impose a two-way order constraint (in both the A and B direction). For each prior distribution and simulated dataset, we run the Gibbs sampler (using JAGS) for 10,000 iterations with a burn-in of 1000 iterations. For each fit, we first compute the posterior mean and 95% quantiles for μ and β and the posterior probabilities for Γ. We then compute the average of these quantities across the 200 simulated datasets.

6.3.2 Simulation Results

Model 1: Linear Regression

Figure 6.2 shows the average posterior probability that each element of Γ equals one for each of the simulation settings. For this figure, we only consider three of the priors. The other two priors automatically impose $\Gamma = 1$. We recall that the values of Γ_A and Γ_B correspond to whether we have strict inequality for adjacent main effects in A and B. The values of Γ_{AB} correspond to whether we have a nonzero interaction term. The *CollapsedHierarchy* and *CollapsedNoHierarchy* methods do an excellent job at identifying nonzero effects when they are present.

In Simulation 5, weak hierarchy assumptions are violated, and as a consequence the *CollapsedHierarchy* method results in inflated posterior probabilities for some main effects in order to also include the interaction. The *CollapsedNoHierarchy* method retains low posterior probabilities for the main effects in this setting. The *Hierarchy* prior also runs into some trouble with Simulation 5, where the posterior probability of Γ_{A3B3} (which takes true value 1) is much smaller than for the collapsed methods. This is a result of the violated weak hierarchy constraint. The *CollapsedNoHierarchy* method is better able to account for the violation of weak hierarchy by making some but not all main effects nonzero. In contrast, the *Hierarchy* prior can only have a nonzero interaction if either all A main effects are nonzero or all B main effects are nonzero.

In Simulation 1, we have no main effects or interactions included in the model. In Simulation 4, we have all main effects included (with no equality for adjacent main effects), and in Simulation 8 we have all main effects included (with no equality for adjacent main effects) and all interactions included. These are the settings in which the *Hierarchy* constraint is well-suited. In these settings, the *Hierarchy* constraint does a reasonable job at determining which values of Γ should be nonzero. However, in Simulations 4 and 8, the collapsed methods do a better job at determining which values of Γ are nonzero.

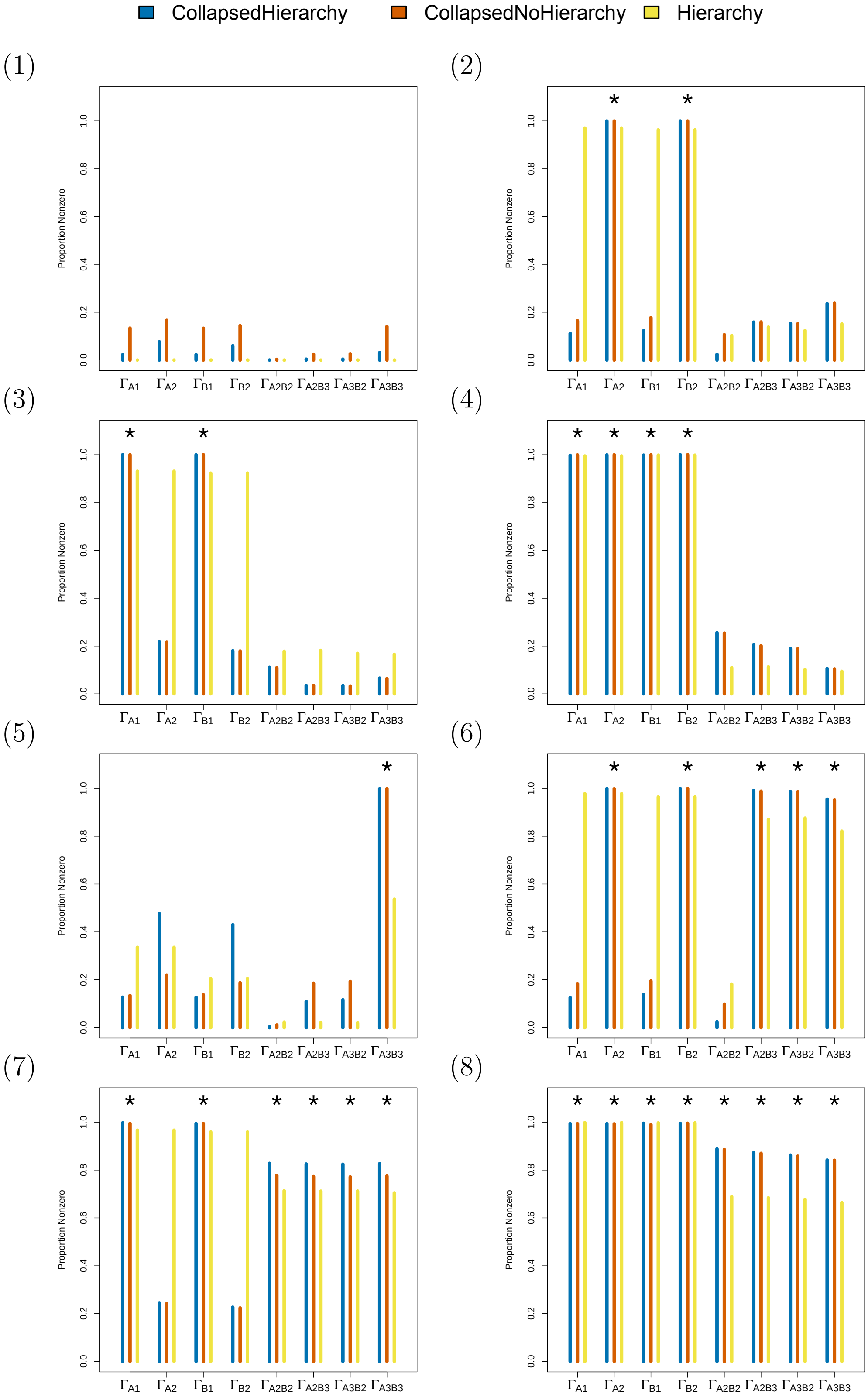
In Simulations 2, 3, 6, and 7, we have one but not both of the values in Γ_A and Γ_B that are nonzero. This is equivalent to having one equality and one strict inequality for the values of μ corresponding to the main effects for that variable. In this setting, the collapsed methods perform very well in terms of the posterior of Γ. The *Hierarchy* prior, which uses grouped selection for the main effects of A and B, performs poorly.

Figure 6.3 shows bias of the posterior mean of μ for each one of the methods through heat maps. This figure allows us to more clearly see the impact of the various priors on the resulting biases. Each three-by-three grouping corresponds to the posterior means of μ for one of the methods. The color in a particular cell corresponds to the magnitude of the bias. The *OrderedNoSelection* prior tends to result in greater bias compared to the other priors. The two collapsed methods produce very similar results. The collapsed methods do well in terms of bias except for Simulation 7. For this simulation, the collapsed methods struggle with estimating the interaction term related to the $A = B = 2$ cell. The *None* method (with no selection or order restrictions) results in essentially unbiased estimates for the posterior mean of μ. The *Hierarchy* prior also runs into problems with bias for Simulations 3, 5, and 8. The bias in Simulation 3 is likely due to the grouped selection used by the Hierarchy prior. The bias in Simulation 5 is due to the violated hierarchy constraint, which the *Hierarchy* prior is ill-suited to handle. In contrast, the *CollapsedHierarchy* prior performs well for this simulation in terms of bias. The bias of the *Hierarchy* prior for Simulation 8 appears to be related to too much spread for the main effects.

Figure 6.4 shows the estimated MSE for the estimate of the posterior mean of μ. The *None* and *OrderedNoSelection* perform poorly for all but Simulation 8. In this simulation, all main effects and interactions are present, and these two priors perform similarly or better than other methods. The good performance of the *None* prior in this setting may be attributed to the very strong effects chosen for the simulated data, and in this setting there may not be as much advantage to imposing order restrictions. The *Hierarchy* prior performs poorly for Simulation 8 due to bias, and the collapsed priors perform worse somewhat poorly in Simulations 7 and 8 prior due to some increased bias from too much shrinkage and very slightly inflated standard errors for μ_{11} in Simulation 8. The *Hierarchy* prior performs well in Simulations 1 and 4. In these simulations, the hierarchy constraint is satisfied, and either all or no main effects are included. In Simulations 1-6, the collapsed priors perform similarly or better than the other methods.

In **Appendix L**, we present additional results for this set of simulations including coverage and average credible interval widths for μ, the average credible intervals for μ and β, and heat maps for the posterior mean of μ.

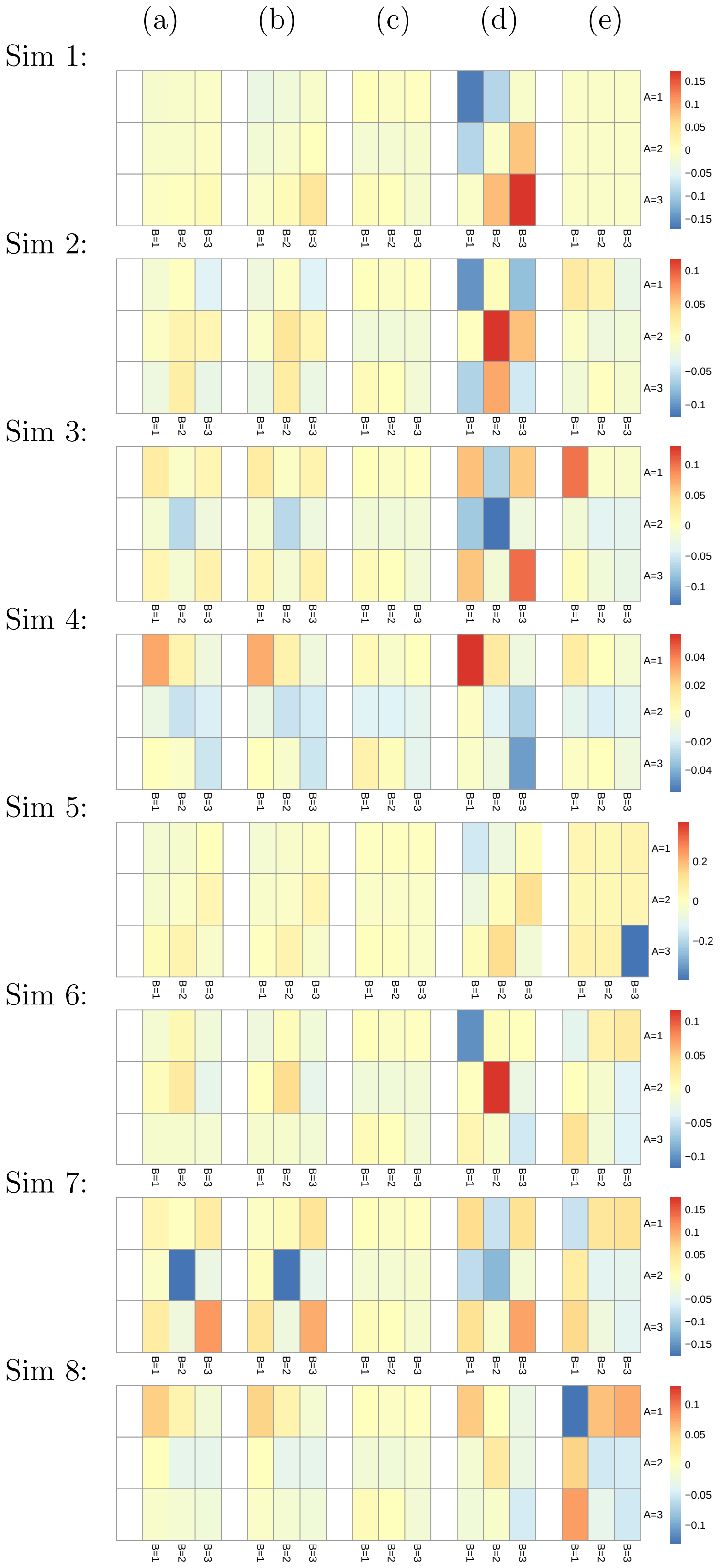
Figure 6.2: Average Proportion of Nonzero Γ (Linear Regression)



This figure shows the proportion of iterations in which each element of Γ is nonzero, averaged across 200 simulations. These proportions equal the posterior probability that each element of Γ equals one. '*' indicates a true nonzero value for Γ.

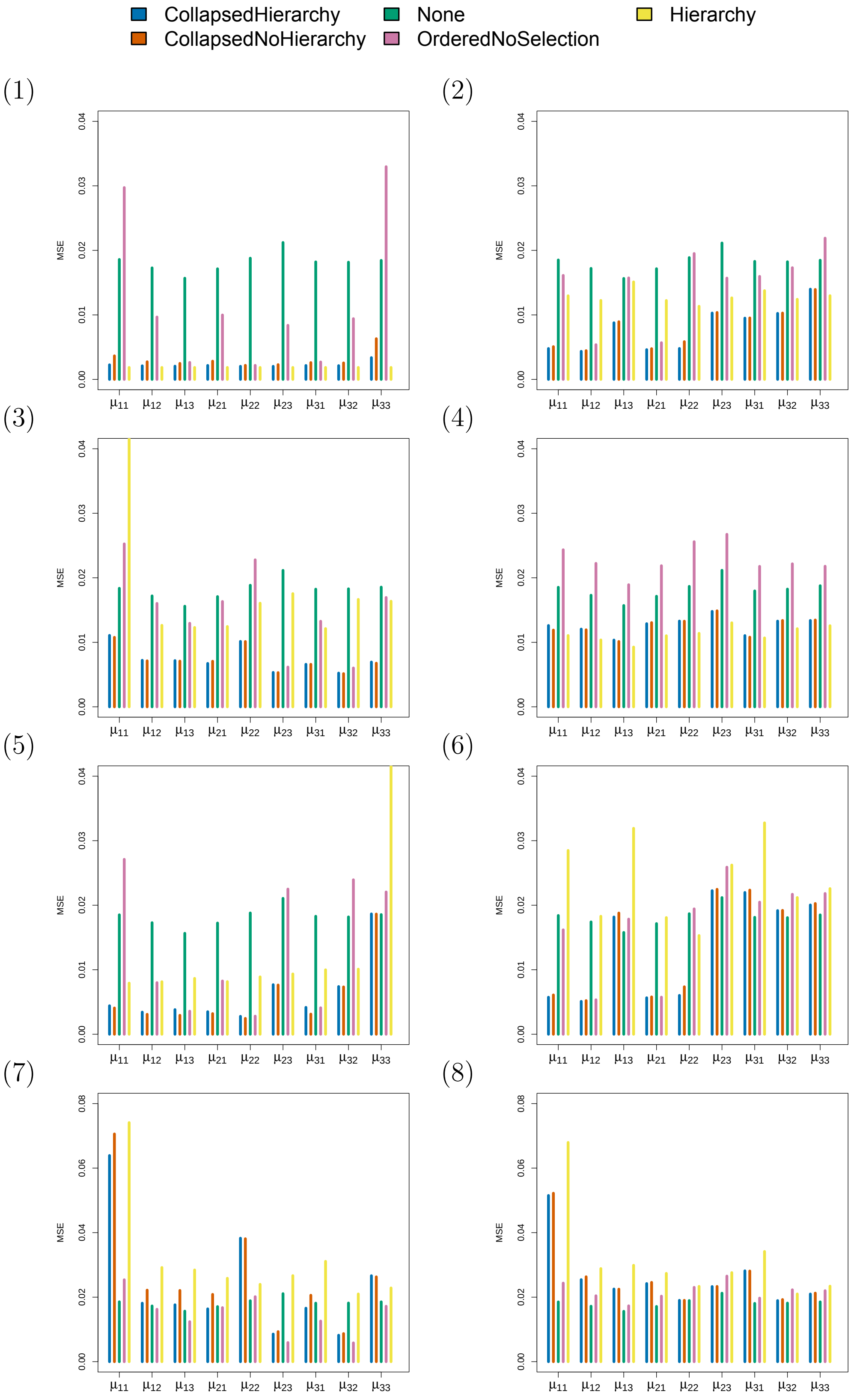
Figure 6.3: Heat Maps for Bias of Posterior Mean of μ (Linear Regression)

(a) CollapsedHierarchy (b) CollapsedNoHierarchy (c) None (d) OrderedNoSelection (e) Hierarchy



This figure shows the bias for the posterior mean values of μ for different combinations of A and B, averaged across 200 simulations

Figure 6.4: MSE for $\hat{\mu}$ (Linear Regression)



This figure shows the MSE of the posterior mean of μ. This was calculated as the squared bias of the posterior means across 200 simulations plus the variance of the posterior mean estimates across the 200 simulations.

Model 2: Logistic Regression

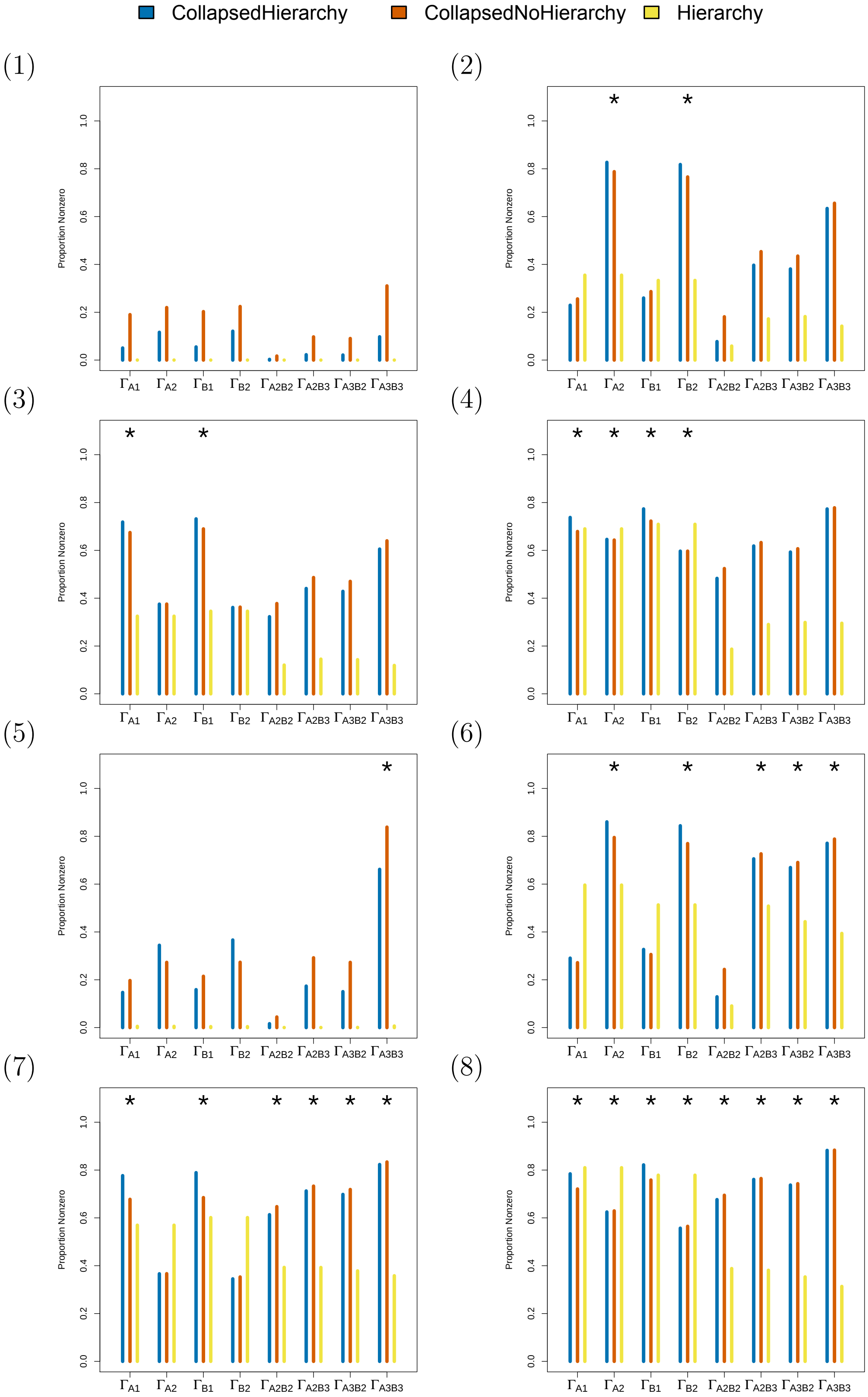
Figure 6.5 shows the average posterior probability that $\Gamma = 1$ for each of the simulation settings. We generally see greater difficulties in determining which values of Γ equal 1 than in the linear regression case. The collapsed methods appear to do a reasonable job at determining which main effect terms should be included, but they have a tendency to estimate large posterior probabilities for interaction terms even when there are none. The *Hierarchy* method, in contrast, estimates lower posterior probabilities for the interaction terms even when there are interaction terms present. The *Hierarchy* totally misses the interaction term in Simulation 5, where weak hierarchy is violated.

Figure 6.6 shows bias of the posterior mean of μ for each one of the methods through heat maps. The *OrderedNoSelection* prior tends to result in greater bias compared to the collapsed priors and the *Hierarchy* prior. The two collapsed methods produce very similar results. The collapsed methods generally do well in terms of bias, but they tend to overestimate μ_{33} in simulations in which Γ_{A1} and Γ_{B1} are nonzero. The *None* method (with no selection or order restrictions) shows evidence of numerical issues, as the biases for the interaction terms tend to be large. The *Hierarchy* prior also runs into problems with bias for Simulation 5, where the hierarchy constraint is violated. In contrast, the *CollapsedHierarchy* prior performs well for this simulation in terms of bias.

Figure 6.7 shows the estimated MSE for the estimate of the posterior mean of μ. The *None* and *OrderedNoSelection* perform poorly in all simulation settings. The *Hierarchy* prior performs well in Simulation 1, and the *CollapsedHierarchy* method performs only slightly worse than the Hierarchy prior for the simulation. The *CollapsedNoHierarchy* outperforms the *None* and *OrderedNoSelection* priors in Simulation 1. For other simulations, the *Hierarchy* prior tends to have larger MSE than the collapsed methods for parameter except μ_{33}. For this parameter, the *Hierarchy* prior often performs better than the collapsed methods due to an increased bias for this parameter for the collapsed methods. For all other settings and parameters, the collapsed methods tend to have smaller or similar MSE compared to the other methods. In general, the MSE values tend to be much larger for the interaction terms due to the large standard errors.

In **Appendix L**, we present additional results for this set of simulations including coverage and average credible interval widths for μ, the average credible intervals for μ and β, and heat maps for the posterior mean of μ.

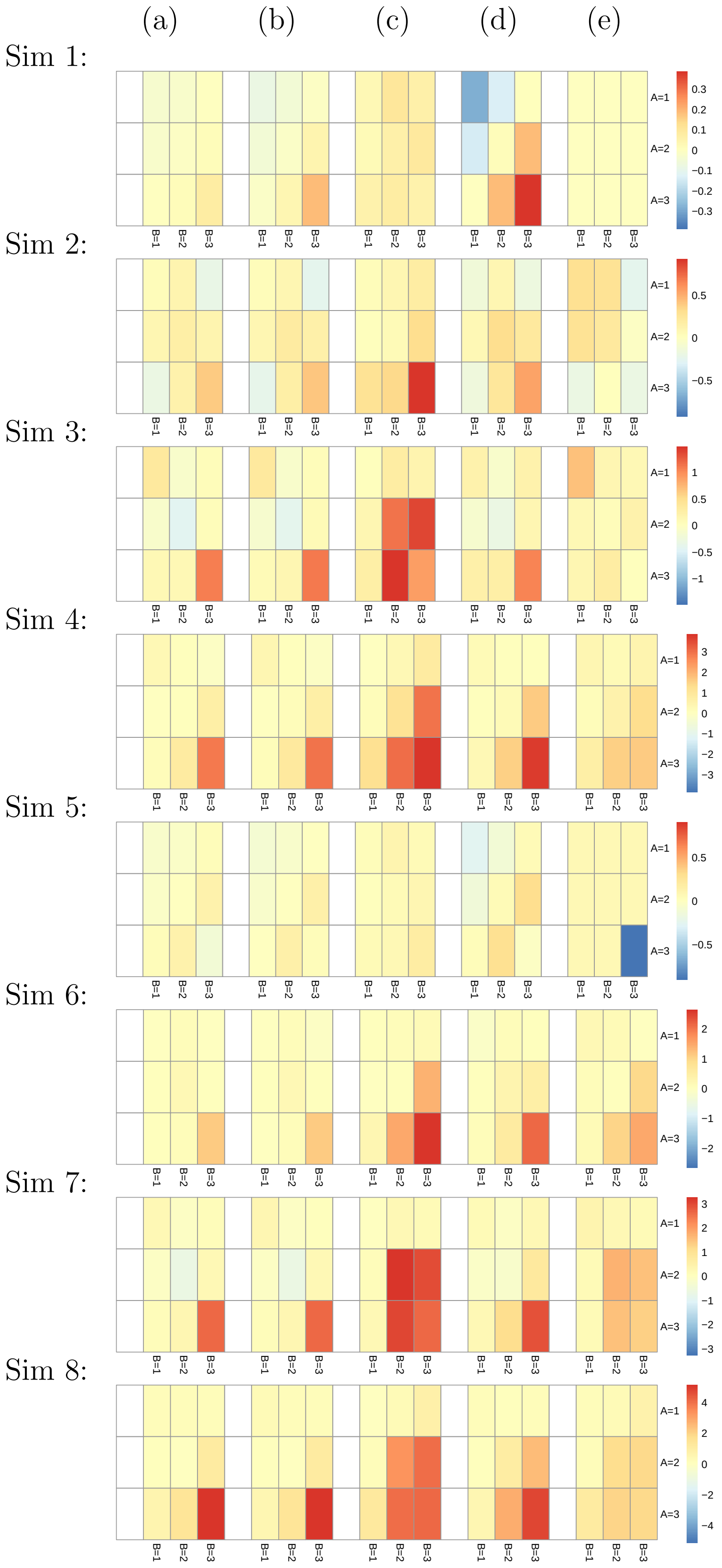
Figure 6.5: Average Posterior Probabilities of Γ (Logistic Regression)



This figure shows the proportion of iterations in which each element of Γ is nonzero, averaged across 200 simulations. These proportions equal the posterior probability that each element of Γ equals one. '*' indicates a true nonzero value for Γ.

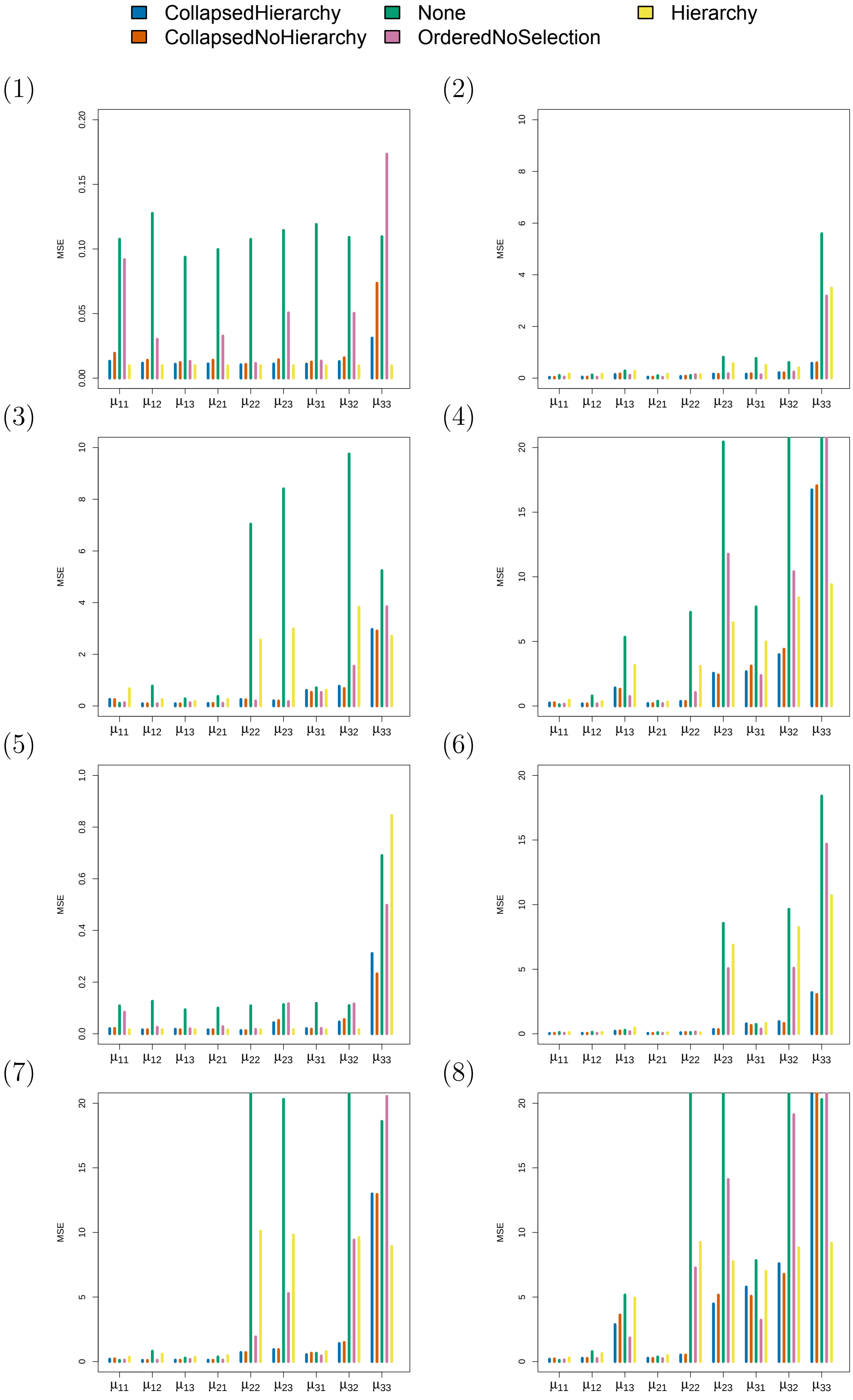
Figure 6.6: Heat Maps for Bias of Posterior Mean of μ (Logistic Regression)

(a) CollapsedHierarchy (b) CollapsedNoHierarchy (c) None (d) OrderedNoSelection (e) Hierarchy



This figure shows the bias for the posterior mean values of μ for different combinations of A and B, averaged across 200 simulations

Figure 6.7: MSE for $\hat{\mu}$ (Logistic Regression)



This figure shows the MSE of the posterior mean of μ. This was calculated as the squared bias of the posterior means across 200 simulations plus the variance of the posterior mean estimates across the 200 simulations.

6.4 Discussion

In this chapter, we developed a prior that can perform variable selection for interactions in the presence of one- or two-way order restrictions. This prior can also incorporate heredity restrictions (which we refer to as "hierarchy constraints") into the estimation.

Previous work has explored order-restricted variable selection for main effects models (Otava et al., 2014) and has explored hierarchy restrictions to improve variable selection with interactions (Chipman, 1996), but no previous work has explored order-restricted variable selection for models with interactions. Throughout the chapter, we casually refer to the proposed methods as "collapsed" methods because they involve merging adjacent categories in the order restriction together using variable selection.

It is well-known that inference under order restrictions can result in parameter estimates that are biased away from each other. One advantage of the proposed method is that this bias can be avoided or reduced through allowing some parameters in the order restriction to be equal. Unlike usual order-restricted inference, the proposed methods don't require strict order restrictions, which reduces the strength of the prior assumptions regarding ordering and allows greater flexibility in the resulting model. Additionally, it is often unappealing to include interactions of variables without first including the main effects in the model. The proposed prior can incorporate hierarchy constraints, allowing us to restrict our focus to models that "make sense." Through imposing order restrictions, we are able to observe gains in efficiency over inference without the order restrictions, and the variable selection component allows us some control over the parameter estimates and the size of the resulting model.

Simulations compare the bias and efficiency properties of the proposed methods over existing priors for linear regression and logistic regression models in a variety of simulation settings. In particular, we compare our proposed methods to priors with no selection or ordering, order restrictions without selection, and hierarchy-constrained selection without order restrictions. For linear regression, the proposed methods outperformed all other methods considered in terms of estimation of the "true" inclusion/exclusion of the main effect parameters and interaction terms. For logistic regression, the proposed methods did a reasonably good job at determining values for the main effects, but they struggled when estimating the interactions. This is because there is substantially less information in the observed outcome for logistic regression than for linear regression, and consequently it is more challenging to determine which parameters should be included and excluded from the model. Indeed, the proposed methods resulted in better estimates for the interactions than the other methods considered. In both the linear and logistic regression settings, the proposed methods generally produced parameter estimates with no or little bias, and the widths of the corresponding credible intervals were generally narrower for the other methods. Overall, simulations suggest that the proposed priors can improve flexibility by allowing for the merging of adjacent categories. The proposed priors can often improve efficiency and reduce bias compared to usual order-restricted methods through selection and may do better in the face of violations of hierarchy than the usual hierarchy-constrained selection priors.

This chapter focuses on the proposed prior itself and its properties in simulation, but our ultimate goal is to apply this prior to the head and neck data. There is a belief in the literature that different subsites for head and neck cancer (e.g. oral cavity, oropharynx, hypopharynx) are very distinct and may have different covariate effects. For example, the effect of HPV may differ across cancer subsites. We would like to add interactions between cancer subsite and other covariates in our modeling of the head and neck data. Given the large number of covariates, we also want to incorporate variable selection.

In the head and neck dataset, we have various covariates that have implied ordering. For example, we have comorbidities, AJCC stage, T stage, N stage, etc. For some parts of the model, we may want to impose order constraints for these variables. For example, suppose we want to model the head and neck data using a multistate cure model as in **Chapter IV**. We may want to constrain that the effects of comorbidities on the rate of death from other causes is non-decreasing with worsening comorbidities. If we also incorporate interactions between comorbidities and cancer subsite in the model, we can apply the proposed variable selection prior imposing a one-way order restriction. Additionally, suppose we wanted to include an interaction between T stage and N stage in the model for time to cancer recurrence. We may believe that increases in either T or N stage should result in greater or equal rates of recurrence. We could incorporate this assumption using variable selection with two-way order restrictions. In this way, the variable selection methods explored in this chapter are of great interest when modeling the head and neck cancer data.

One basic assumption made in this chapter is that we have some prior belief regarding the outcome model and the scale on which we want to assess ordering and interactions. For example, suppose that instead of modeling Y using a linear regression, we model $\sqrt{Y}$. Monotonic transformations of Y will preserve the parameter ordering, but they may not preserve the interactions, and modeling on different scales may change the posterior inclusion probabilities for the corresponding interaction terms. Additionally, in the linear regression case, our proposed methods also assume that we have constant variance across different combinations of A and B. We may have constant variance on the Y scale but not on the $\sqrt{Y}$ scale. We can generalize the proposed methods for the non-constant variance setting, but this issue is still worth some thought. For many applications, there may be a natural scale on which to explore interactions. In the head and neck cancer example, we are interested in exploring multistate modeling involving Cox regression models. In this modeling framework, we model the hazard $\lambda(t) = \lambda_0(t)e^{X\beta}$ where $\lambda_0(t)$ is the baseline hazard function. In this setting, it is natural to explore interactions incorporated linearly in the $X\beta$ term.

The work in this chapter is a first look at this proposed prior distribution, and additional work is needed. This analysis implemented estimation under the proposed priors automatically using the software JAGS. In the future, we will explore the structure of the posterior distributions of the parameters and develop an MCMC sampling scheme for parameter estimation. Additionally, we chose only 8 simulation settings in which to explore the proposed methods, but additional exploration is needed to determine how the proposed priors will perform in different scenarios. Finally, we may often tend to believe that parameters related to interaction terms are generally small even when they are included in the model. We can alter the specification of the prior distribution for θ_{AjBk} values to incorporate this prior belief. In the future, we ultimately plan to explore how to apply the proposed methods to multistate modeling of the head and neck data. Development of an MCMC sampling scheme is particularly important when it comes to applying the proposed methods to the head and neck dataset because JAGS is not well-suited to deal with the covariate and outcome missingness present in the head and neck dataset. Ultimately, we hope that by applying the proposed methods to the head and neck data (and incorporating our prior beliefs regarding parameter ordering), we can improve the efficiency of our estimation. The resulting modeling could be used to produce better outcome predictions, which could be extremely useful in clinical applications.

Chapter VII

Conclusion

With the increasing availability of patient information (from past medical records, new diagnostics, genetic testing, etc), there is a strong need to develop statistical methods to handle the challenges presented. This is particularly true for the large-scale observational data often used in cancer research. In this dissertation, we consider a study of recurrence and death in patients with head and neck cancer. Through this dissertation, we aim to address some of the statistical problems that arise for the head and neck cancer data, but the methods we develop can be applied to other diseases and different scientific questions. In particular, we consider the setting where a subset of the population is cured of their disease and can never experience a cancer recurrence, of which the head and neck cancer data is an example. Several frameworks exist for modeling recurrence with an underlying cured fraction of the population. We will consider two such models: the Cox proportional hazards mixture cure model (Kuk and Chen, 1992; Sy and Taylor, 2000) and the multistate cure model (Conlon et al., 2013). In this dissertation, we address issues of missing data, parameter estimation, and variable selection that arise in the application of these models to data.

Chapter II of this dissertation explores imputation-based methods for dealing with missing covariate values for the Cox proportional hazards cure model. We consider chained equations-type imputation strategies, which involve specifying a model for each variable with missingness. We first use an imputation strategy developed in Bartlett et al. (2014) that incorporates the structure of the cure model to guide the form of the imputation distributions. We then propose several regression model approximations that are easier to use for imputation in practice. We compare the proposed imputation methods to existing methods for imputing missing covariates for survival data without a cured fraction, and we apply the proposed methods to the head and neck cancer data. This work is the first to explore chained equations imputation for the cure model, and it therefore provides an extremely useful addition to the cure model literature. In the course of developing imputation methods for the cure model setting, we proposed a method we call "Outcome Binning" that performs fairly well in the cure model setting. As this method is not specific to the cure model setting, it would be interesting to explore the performance of this method for imputing missing covariates in the standard survival setting without a cured fraction in the future.

The second chapter of this dissertation considers covariate imputation under an assumption of missing at random (MAR), where missingness is assumed to be related to fully-observed information (Little and Rubin, 2002). However, this is a restrictive assumption, and there are many situations in which missingness may depend on unobserved information, called missing not at random (MNAR). In **Chapter III** of this dissertation, we consider a particular MNAR mechanism called latent ignorability or latent missing at random (LMAR), where missingness is allowed to depend on missing information through a latent or partially latent variable. We consider a modeling framework in which covariate or outcome missingness can depend on a latent variable that is part of the outcome model. In a mixture of normals model, for example, covariate missingness may be related to the underlying mixing variable. In the cure model setting, missingness may be related to cure status. We propose a sequential imputation algorithm for dealing with LMAR or MAR missingness in the covariates and/or outcomes. We derive the imputation distributions under joint modeling assumptions, and we then describe how we can use the results under a joint model to guide imputation when we do not assume a joint model. This allows for increased flexibility in the models used for imputation over standard joint modeling. One primary limitation of this work is the difficulty regarding parameter identifiability. In the chapter, we explore issues of identifiability and convergence for the proposed algorithm, but we do not present any theoretical results and instead address this problem through simulation and several examples. We provide guidance for how to apply the proposed methods in practice.

Chapters IV and V of this dissertation explore statistical issues arising for multistate cure models. Multistate models in general have many valuable uses in medical research. They provide a unified way to incorporate information from multiple event time outcomes, they allow us to study the impact of patient characteristics on different aspects of disease progression, and they are extremely useful for making patient-specific predictions. The multistate cure model is of particular interest in cancer research because it can allow us to study covariate effects on the cure rate, the rate of recurrence among non-cured subjects, and the death rates before and after recurrence. Additionally, incorporating the underlying cure structure into the multistate model may help us obtained better patient-specific predictions for recurrence and death rates. Despite the many advantages of using the multistate cure model, there are currently many statistical barriers that may make this model difficult to apply to data.

The first barrier is the lack of standard statistical software for fitting this model. The existing method in Conlon et al. (2013) for fitting the multistate cure model requires custom software and technical knowledge, and it can take a long time to converge. Additionally, there is no previous discussion of how to handle missing covariate data, which is extremely common in practice. In **Chapter IV** of this dissertation, we develop maximum likelihood-based methods for estimating multistate cure model parameters. In the setting with no missingness beyond the partially latent cure status, we propose an Expectation-Maximization (EM) algorithm for estimation. The proposed method can accommodate parametric or nonparametric baseline hazards and different assumptions regarding the rates of death from other causes in the cured and non-cured subjects. This provides a gain in the model flexibility over existing methods. We further propose a Monte Carlo Expectation-Maximization (MCEM) algorithm for estimating multistate cure model parameters in the presence of covariate missingness and/or unequal censoring of the outcomes. By unequal censoring, we refer to the setting in which we have longer follow-up for death than we have for recurrence, which often arises in practice. The proposed method involves imputing missing values for the covariates, underlying cure status, and outcome data when we have unequal censoring. We develop a novel imputation-based approach for dealing with unequal censoring, and this approach can be applied in general illness-death model settings. Simulations demonstrate good performance of the proposed methods when the modeling assumptions are sufficiently restrictive, and we apply the proposed methods to the head and neck cancer data. We develop an R package called *MultiCure* for fitting the multistate cure model using the proposed methods. We hope the developed methods and the corresponding R package can make multistate cure models more accessible to analysts performing data analysis. Future developments for this R package can incorporate ridge and LASSO penalization options and shrinkage-based methods to improve identifiability for the parameters in the transitions to death from other causes as explored in **Appendix J**. We would also like to develop a separate R package that can perform the imputation-based method for dealing with unequal censoring for general illness-death model settings. Ideally, this package would be easily combined with other multistate modeling software in R such as the package *mstate*.

A second barrier to the application of multistate models in general and the multistate cure model in particular is the large number of model parameters. When the number of model parameters is large, we can often run into numerical issues and overfitting. In **Chapter V**, we explore how we can apply existing Bayesian and maximum-likelihood-based variable selection methods (with some small modifications) in the multistate modeling setting. We restrict our attention to a particular multistate model. We propose a novel generalization to the multistate cure model that incorporates subjects with persistent disease. By subjects with persistent disease, we mean subjects that never appeared to clear their cancer through treatment. We expect these subjects to have different death rates than other subjects, and the developed multistate cure model with persistence can account for this. We apply this model to the head and neck dataset using several different Bayesian and maximum likelihood-based variable selection/shrinkage strategies in the literature and compare the resulting parameter estimates and credible/confidence intervals. We provide some tentative recommendations for the application of existing variable selection methods in general multistate modeling settings. Additional explorations comparing the different Bayesian variable selection methods for the head and neck data can explore different choices for hyperparameters and hyperpriors, which may change the comparative rates of shrinkage and resulting model inference across the methods.

In order to improve efficiency and control the number of model parameters, we may want to incorporate additional parameter restrictions into the variable selection procedure. For example, suppose that we have a strong prior belief regarding the ordering of parameters. For the head and neck cancer data, we may believe that worse comorbidities will be related to similar or greater rates of death from other causes. Otava et al. (2014) explores how we can incorporate order restrictions with Bayesian variable selection. Suppose we want to impose the restriction $\mu_1 \leq \mu_2 \leq \mu_3$, where the μ parameters correspond to different levels of a variable, A. The method in Otava et al. (2014) uses the Bayesian variable selection prior in Kuo and Mallick (1998) to determine whether adjacent values of μ are equal or strictly ordered. By setting adjacent values of μ to be equal, we are equivalently grouping (or collapsing) adjacent values of A together.

In **Chapter VI**, we explore a more general scenario where our goal is to perform Bayesian variable selection for a model with interactions and order restrictions for one or both of the interacted variables. Existing Bayesian variable selection methods for models with interactions often incorporate heredity restrictions, which determine the inclusion of interaction terms based on the inclusion of the corresponding main effects. We propose a Bayesian variable selection prior that can incorporate both heredity constraints and one- or two-way order restrictions. Simulations demonstrate the performance of the proposed prior in the linear and logistic regression settings. One drawback of the proposed approach is that it supposes that the model formulation and the scale for evaluating interactions are specified ahead of time. Changes to the model or the scale for the interactions (for example, by modeling $\sqrt{Y}$ instead of Y) may alter the importance of the interactions. Our plan for the future is to apply these methods to study the interaction between cancer site and order-restricted variables (such as comorbidities or cancer stage) for the head and neck cancer data.

In the future, we would like to develop additional Bayesian variable selection (BVS) methods to address problems for the head and neck data. For example, we note that we tend to see high autocorrelation across iterations of the MCMC algorithm when applying the BVS methods to the head and neck data, which negatively impacts mixing. This issue is common for many applications of BVS, and it requires us to perform a large number of MCMC iterations to estimate the posterior mean and credible intervals well. This autocorrelation is a result of the Metropolis-Hastings methods used to perform the various parameter draws. We hope to address the autocorrelation issue by performing parameter draws via other methods such as rejection sampling, which we do not expect to suffer from the same degree of inter-iteration correlation.

Using a multistate cure model fit to the head and neck cancer data, we would ultimately like to develop a web application that can be used by clinicians and researchers to estimate state occupancy probabilities given individual patient characteristics. Such a tool could be extremely useful for medical decision-making and for studying the aggregate effects of different covariates on prognosis. When the model parameters are estimated using BVS, it is not clear how to estimate the state occupancy probabilities. Do we use the Bayesian model-averaged posterior mean of the parameter, do we use the covariate combination with the highest posterior weight, or do we use some combination? Future work could explore this issue along with how to estimate corresponding standard errors.

In the design of medical studies, there is great interest in obtaining estimates of statistical power and sample size requirements. If the multistate cure model is to be more widely used in medical research, methods are needed to estimate these and related quantities. Additional generalizations of the multistate cure model may also widen the applicability. In all previous explorations of the multistate cure model, the underlying transition times between states are assumed to be independent within a subject and independent between subjects given covariates and baseline cure status. However, this may not always be the case. For example, in the head and neck cancer dataset, we have data from different hospitals. It may be that the recurrence and death times are correlated across individuals treated at the same hospital. Additionally, event times within an individual could be correlated even accounting for covariates. In literature for illness-death models, many authors have explored the inclusion of frailty terms to the state transition models to account for residual correlations within and between subjects (Bijwaard, 2014; de Castro et al., 2015). Future work can generalize the multistate cure model explored in this dissertation to incorporate different types of frailty terms and explore corresponding estimation methods. This will allow us to relax the independence assumptions, which can widen the scope of problems well-suited for the multistate cure model.

The ultimate goal of this dissertation is to provide statistical methods and guidance for dealing with common problems of missing data, parameter estimation, and variable selection in the cure model setting. In this dissertation, we have developed methods to handle missing data and variable selection for cure models and multistate models. This provides analysts with the statistical tools to apply cure models to messy data often seen in practice. We further developed imputation methods to handle latent ignorable missingness, which has not be previously explored in the cure model setting. In order to make multistate cure models easier to fit, we developed a more convenient estimation technique and provided an R package. This package can also be used to estimate state occupancy probabilities, which are of great clinical interest. Through this methodological work, we hope to improve the ability of analysts to apply cure models to "real data" both through providing solutions to common data problems and through creating software for fitting the models. In this dissertation, we focus our methodological development around problems observed in cure modeling of the head and neck cancer data, but the proposed approaches can be applied to many other scientific questions and modeling frameworks. This dissertation, therefore, provides methods to deal with issues of missing data and variable selection for a wide range of problems.

Appendix A

Performing Parameter Draws in LMAR-based Imputation

In this appendix, we provide more details regarding the univariate imputation steps for imputing missing values in D and L. In particular, we discuss distributions we can use to perform the parameter draws within the sequential imputation algorithm. Our proposed method for drawing model parameters within a given univariate imputation step will depend on whether we are performing imputation of the latent variable or a variable in D. The proposed method in **Section 3.3.3** assumes that imputation proceeds under a fully-specified joint model, and we generalize this algorithm in **Section 3.3.4** for settings in which the imputation distributions do not correspond to a valid joint distribution. Here, we will suppose that L is imputed from the kernel in (3.6) and that missing X and Y are imputed from *working* imputation models that may or may not correspond (3.7) and (3.8). Therefore, the following exploration can be applied when outcomes and covariates are imputed using (3.7) and (3.8) or using approximations.

First, we will review some notation. Define $D^{(p)}$ to be the p^{th} variable in D and $D^{(-p)}$ to be all variables in D except $D^{(p)}$. Parameter ν represents the parameters for the joint distribution, $f(D, L, R; \nu)$. We partition $\nu = (\phi, \rho)$ where ϕ represents the missingness model parameters and ρ represents all other model parameters. We assume that ρ and ϕ are distinct (a priori independent). Suppose that we specify $\tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \rho^p)$ to be the *working* conditional distribution of $D_i^{(p)}$ *used for imputation.* We can view $\tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \rho^p)$ as an approximation of $f(D_i^{(p)}|D_i^{(-p)}, L_i; \rho)$. If we use the form of the full conditional distribution as in (3.7) and (3.8) in the **Chapter III**, ρ^p will be a subset of ρ. If we impute using regression models, ρ^p may not be directly related to ρ. We suppose that we impute L from $f(L_i|D_i, R_i; \nu)$ as described in (3.6).

A.1 Imputing $D^{(p)}$

In **Chapter III**, we discuss how, when missingness is LMAR, we can impute $D^{(mis)}$ ignoring the contribution of R (assuming some distinctness properties). This is a result of the assumption that missingness is conditionally independent of $D^{(mis)}$. Rather than imputing $D^{(mis)}$ directly from $f(D^{(mis)}|D^{(obs)}, L)$, we instead obtain a draw of $D^{(mis)}$ by iteratively drawing missing values of each $D^{(p,mis)}$ from $f(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)$ or from an approximated version, $\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)$, treating the most recent imputations for the other variables as if they were observed data (including L).

At a given iteration, we want to draw missing values of $D^{(p)}$ under MAR and LMAR from its posterior predictive distribution:

$$\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L) = \int \tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L; \rho^p)\tilde{f}(\rho^p|D^{(p,obs)}, D^{(-p)}, L)d\rho^p$$

This integral suggests an approach for drawing from the posterior predictive distribution. Assuming that the data $D_i^{(p)}$ across subjects i are conditionally independent given L and $D_i^{(-p)}$, we can obtain a draw from the posterior predictive distribution by performing the following (Little and Rubin, 2002):

1) Draw ρ^p from $\tilde{f}(\rho^p|D^{(p,obs)}, D^{(-p)}, L)$
2) Draw missing $D_i^{(p)}$ from $\tilde{f}(D_i^{(p,mis)}|D_i^{(p,obs)}, D_i^{(-p)}, L_i; \rho^p) = \tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \rho^p)$.

We note that step 1) involves drawing ρ^p conditioning on $D^{(p,obs)}$ using only the observed part of $D^{(p)}$. This is consistent with chained equations imputation in which we draw parameter values using only the observed values of $D^{(p)}$ (Van Buuren et al., 2006). The step for drawing ρ^p conditioning only on the observed data can be accomplished by using the data with observed values for $D^{(p)}$ and prior $\tilde{f}(\rho^p)$. If we assume the prior distribution is proportional to 1, we can draw ρ^p by fitting model $\tilde{f}(D^{(p)}|D^{(-p)}, L; \rho^p)$ to a bootstrap sample of the data with observed values for $D^{(p)}$. We note that while this step for drawing ρ^p does not use the most recent imputation of $D^{(p)}$, it does use the imputed values for L.

An alternative to the above is to draw ρ^p using a Gibbs-type approach. In Gibbs sampling-type imputation algorithms, parameter values are drawn using all of the most recent imputed data, including imputed values for $D^{(p)}$ from the previous iteration. This approach is also used in SMC-FCS, a modified chained equations approach proposed in Bartlett et al. (2014). If preferred, we can obtain valid parameter draws using this approach as well. We note that we can write

$$\tilde{f}(\rho^p|D^{(p,obs)}, D^{(-p)}, L) = \int \tilde{f}(\rho^p|D^{(p)}, D^{(-p)}, L)\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)dD^{(p,mis)}$$

The above integral suggests that we can obtain a draw from $\tilde{f}(\rho^p|D^{(p,obs)}, D^{(-p)}, L)$ by drawing ρ^p from $\tilde{f}(\rho^p|D^{(p)}, D^{(-p)}, L)$ using the drawn value of $D^{(p,mis)}$ from the previous iteration, which was drawn from $\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)$. Rather than drawing parameter values using the complete case data as is in the usual implementation of chained equations, we can alternatively draw parameters conditioning on the imputed values of $D^{(p)}$ from the last iteration. We use this approach for drawing parameters in our simulations and in our presentation of the proposed method in **Chapter III**.

Rather than approximating the distributions for each variable with missingness with a regression model for imputation, suppose that we impute all variables using the kernel forms in (3.6), (3.7) and (3.8). In this case, ρ^p is is a subset of ρ. For simplicity, we might choose to perform only a single set of parameter draws per iteration of the sequential imputation algorithm and use that set of parameter draws for imputing all of the variables in that iteration. This approach is used in Gibbs sampling-type algorithms. In this case, we might perform a set of parameter draws for ρ in the step for imputing L, which involves drawing ρ using methods treating L as latent as described in the following section. Then, we can use that same drawn value for ρ for imputing the covariate/outcome values. We note that the above derivations above suggest that we should draw ρ conditioning on the imputed values of L when we are imputing covariates/outcomes. In our experience, however, a single draw of ρ using the above approach generally produces good results when we perform our final analysis using only the imputed values of D. When we perform our final analysis using the imputed values of D and L, drawing ρ before each imputation can sometimes produce improved parameter coverage.

A.2 Imputing the Latent Variable

In the imputation step for L at a given iteration of the sequential algorithm, we aim to draw missing values from the posterior predictive distribution:

$$f(L^{(mis)}|L^{(obs)}, D, R) = \int f(L^{(mis)}|L^{(obs)}, D, R; \nu) f(\nu|L^{(obs)}, D, R) d\nu$$

under LMAR and the posterior predictive distribution:

$$f(L^{(mis)}|L^{(obs)}, D) = \int f(L^{(mis)}|L^{(obs)}, D; \rho) f(\rho|L^{(obs)}, D) d\rho$$

under MAR. Here, we treat the most recent imputations for D as if they were the observed data. As before, this integral suggests an approach for drawing from the posterior predictive distribution. We can obtain a draw of the posterior predictive distribution by performing the following:

1) Under LMAR, draw ν from $f(\nu|L^{(obs)}, D, R)$.
   Under MAR, draw ρ from $f(\rho|L^{(obs)}, D)$.
2) Under LMAR, draw missing L_i from $f(L_i^{(mis)}|L_i^{(obs)}, D_i, R_i; \nu) = f(L_i|D_i, R_i; \nu)$.
   Under MAR, draw missing L_i from $f(L_i^{(mis)}|L_i^{(obs)}, D_i; \rho) = f(L_i|D_i; \rho)$

We note here that we are assuming that L_i values are conditionally independent across different values of i. Suppose our outcome model is a linear mixed model with a random intercept, L. Then i here would index the clusters (rather than the units within clusters), and a single value of L would be drawn for all units within the cluster.

Drawing ρ under MAR

When L is partially observed, we can draw ρ from $f(\rho|L^{(obs)}, D) \propto f(L^{(obs)}, D; \rho) f(\rho)$ using only the observed values of L and prior $f(\rho)$ using methods that treat L as latent or partially latent and ignoring R. For example, suppose our outcome model is a mixture of GLMs and we use $f(\rho) \propto 1$. Then, we can draw the parameter for the outcome model by fitting a latent class model to a bootstrap sample of the data treating L as fully latent.

Drawing ρ and ϕ under LMAR

We note that

$$f(\nu|L^{(obs)}, D, R) = f(\rho|L^{(obs)}, D, R, \phi)f(\phi|L^{(obs)}, D, R) \tag{A.1}$$

When L is partially latent (so it is partially observed), we can draw values of ν using only the subjects with L observed. When L is fully latent, however, drawing from (A.1) may not be so simple. Therefore, we will propose an alternative approach that can be applied for latent and partially latent L. We will consider how to draw ϕ and ρ separately using the factorization in (A.1).

We first consider how to draw values for ρ from $f(\rho|L^{(obs)}, D, R, \phi)$. We have that

$$\begin{aligned} f(\rho|L^{(obs)}, D, R, \phi) &\propto f(L^{(obs)}, D, R; \rho, \phi)f(\rho) \\ &\propto f(R|D, L^{(obs)}; \nu)f(L^{(obs)}, D; \rho)f(\rho) \end{aligned}$$

This kernel separates into two factors: one that depends on ϕ and R and one that does not. We note that L is treated as MCAR when L is fully latent and is assumed to be MAR when L is partially latent, so the missingness in L is ignorable given $D^{(obs)}$. When we condition on the imputed D, we can make valid inference about ρ (in a frequentist sense) without conditioning on R and ϕ (Little and Rubin, 2002). However, R does contain some information about the value of L under LMAR (ν and ρ are clearly not distinct) and therefore would contribute some information about ρ. Ignoring R when drawing ρ, therefore, may result in a loss of efficiency. We can validly (but with some potential loss of efficiency) ignore the contribution of R and ϕ to $f(\rho|L^{(obs)}, D, R, \phi)$ and instead draw ρ from $f(\rho|L^{(obs)}, D)$. This is important because it may be difficult to draw from $f(\rho|L^{(obs)}, D, R, \phi)$, but a draw from $f(\rho|L^{(obs)}, D)$ can be obtained using standard methods that treat L as latent or partially latent and ignoring R.

We now consider how to draw values for ϕ. The distribution $f(\phi|L^{(obs)}, D, R)$ may be difficult to draw from under LMAR assumptions since this distribution does not condition on $L^{(mis)}$. We instead use the integral decomposition:

$$f(\phi|L^{(obs)}, D, R) = \int f(\phi|L, D, R)f(L^{(mis)}|L^{(obs)}, D, R)dL^{(mis)}$$

We can obtain a valid draw from $f(\phi|L^{(obs)}, D, R)$ by instead drawing from $f(\phi|L, D, R)$ using the most recent imputation of L, which was drawn from $f(L^{(mis)}|L^{(obs)}, D, R)$. Therefore, we can draw values of ϕ directly using the most recent imputed values of L. This is easier than drawing from $f(\phi|L^{(obs)}, D, R)$ because it can directly incorporate the working LMAR model for the missingness mechanism without integrating out missing values of L. We do not choose to use this same integral decomposition approach for drawing ρ as our proposed approach (which does not condition on the most recent imputation of L) tends to result in more stable convergence properties in our experience (for fully latent L).

Appendix B

Identifiability under LMAR for Joint Normal Models (Example 1)

In **Chapter III**, we restrict applications of the proposed methods to cases in which the model parameters would be identified had the missing data been observed. Here, we present an example in which parameters identified in the LMAR-based model would not be identified if the missing data had been observed. In particular, we first explore assumptions required to achieve identifiability for a measurement error model. Then, we compare the measurement error model to linear mixed models and explain how the linear mixed model is able to attain identifiability of all outcome model parameters.

B.1 Example 1.1: Measurement Error Model with Covariates

Suppose we have a noisy version (Y) of an underlying variable of interest, L. L is never observed, and Y is observed at least for some subjects. We suppose Y and L are univariate and related to fully measured covariates, X. Suppose we model

$$Y_i = \alpha_0 + \alpha_1 L_i + \alpha_2 X_i + e_i, \quad L_i \sim N(\beta_0 + \beta_1 X_i, \Sigma_L), \quad e_i \sim N(0, \sigma^2), \quad e_i \perp L_i$$

This is an example of a measurement error model. This model contains 7 parameters. This implies the following:

$$\begin{pmatrix} Y_i \\ L_i \end{pmatrix} |X_i = N\left(\begin{pmatrix} \alpha_0 + \alpha_1\left(\beta_0 + \beta_1 X_i\right) + \alpha_2 X_i \\ \beta_0 + \beta_1 X_i \end{pmatrix}, \begin{pmatrix} \sigma^2 + \alpha_1^2 \Sigma_L & \alpha_1 \Sigma_L \\ \alpha_1 \Sigma_L & \Sigma_L \end{pmatrix}\right)$$

$$L_i|Y_i, X_i \sim N\left(\beta_0 + \beta_1 X_i + \frac{\alpha_1 \Sigma_L}{\sigma^2 + \alpha_1^2 \Sigma_L}\left[Y - \alpha_0 - \alpha_1\left(\beta_0 + \beta_1 X_i\right) - \alpha_2 X_i\right],\right.$$
$$\left.\Sigma_L - \frac{\alpha_1^2 \Sigma_L^2}{\sigma^2 + \alpha_1^2 \Sigma_L}\right)$$

Suppose we have no missingness in Y. In this case, the observed data likelihood can be expressed as follows:

$$Lik_{NoMissing}^{(obs)} = \prod_{i=1}^{n} f(Y_i|X_i) = \prod_{i=1}^{n} N\left(Y_i; \alpha_0 + \beta_0\alpha_1 + [\alpha_2 + \alpha_1\beta_1]X_i, \sigma^2 + \alpha_1^2\Sigma_L\right)$$

where $N(a; b, c)$ indicates the normal density evaluated at a with mean b and variance c. In order for the model to be identified, **we must fix 4 of the 7 parameters** in this model $(\alpha_0, \alpha_1, \alpha_2, \sigma^2, \beta_0, \beta_1, \Sigma_L)$, so we can identify the 3 remaining parameters.

Suppose instead that we have LMAR missingness in Y is follows: Probit$(P(R_i^Y = 1|L_i, Y_i, X_i)) = \phi_0 + \phi_1 L_i$, so we assume that missingness in Y only depends on L. This scenario is a simple case of the Heckman (1976) selection model if $\alpha_1 = 0$ with a modified missingness model (Little and Rubin, 2002; Heckman, 1976). The observed data likelihood can be expressed as follows:

$$Lik^{(obs)} = \prod_{i=1}^{n}\left[\int \Phi(\phi_0 + \phi_1 L_i) f(Y_i, L_i|X_i) dL_i\right]^{R_i^Y}\left[\int (1 - \Phi(\phi_0 + \phi_1 L_i)) f(L_i|X_i) dL_i\right]^{1-R_i^Y}$$
$$= \prod_{i=1}^{n}\left[f(Y_i|X_i)\int \Phi(\phi_0 + \phi_1 L_i) f(L_i|Y_i, X_i) dL_i\right]^{R_i^Y}\left[1 - \int \Phi(\phi_0 + \phi_1 L_i) f(L_i|X_i) dL_i\right]^{1-R_i^Y}$$
$$= \prod_{i=1}^{n}\left[f(Y_i|X_i) E_{L|Y,X}\left(\Phi(\phi_0 + \phi_1 L_i)\right) dL_i\right]^{R_i^Y}\left[1 - E_{L|X}\left(\Phi(\phi_0 + \phi_1 L_i)\right)\right]^{1-R_i^Y}$$

We will make use of the following identity:
Let $U \sim N(\mu_1, \sigma_1^2)$ and $V \sim N(\mu_2, \sigma_2^2)$ be independent random variables. Now, $U - V \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.

$$\Phi\left(\frac{-(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) = P(U \leq V) = \int \Phi\left(\frac{v - \mu_1}{\sigma_1}\right) f_V(v) dv = E_V\left(\Phi\left(\frac{v - \mu_1}{\sigma_1}\right)\right)$$

Using this identity and setting $\sigma_1 = 1/\phi_1$ and that $\mu_1 = -\phi_0/\phi_1$, we have that

$$Lik^{(obs)} = \prod_{i=1}^{n} \left[1 - \Phi\left(\frac{\phi_0 + \phi_1(\beta_0 + \beta_1 X_i)}{\sqrt{1 + \phi_1^2 \Sigma_L}} \right) \right]^{1-R^Y}$$

$$\times \left[f(Y_i|X_i)\Phi\left(\frac{\phi_0 + \phi_1(\beta_0 + \beta_1 X_i) + \phi_1\alpha_1\Sigma_L \left[\sigma^2 + \alpha_1^2\Sigma_L\right]^{-1}(Y_i - \alpha_0 - \alpha_1(\beta_0 + \beta_1 X_i) - \alpha_2 X_i)}{\sqrt{1 + \phi_1^2(\Sigma_L - \alpha_1^2\Sigma_L^2\left[\sigma^2 + \alpha_1^2\Sigma_L\right]^{-1})}} \right) \right]^{R^Y}$$

This expression contains 9 parameters, but we cannot simultaneously identify all parameters. Suppose we set

$$\begin{array}{llll} A = \phi_1\alpha_1\Sigma_L & B = \sigma^2 + \alpha_1^2\Sigma_L & C = \alpha_0 + \alpha_1\beta_0 & D = \alpha_1\beta_1 + \alpha_2 \\ E = \phi_0 + \phi_1\beta_0 & F = \phi_1\beta_1 & G = 1 + \phi_1^2\Sigma_L & \end{array}$$

Then we can rewrite the observed data likelihood as:

$$Lik^{(obs)} = \prod_{i=1}^{n} \left[N(Y_i; C + DX_i, B)\Phi\left(\frac{E + FX_i + \frac{A}{B}(Y_i - C - DX_i)}{\sqrt{G - \frac{A^2}{B}}} \right) \right]^{R^Y} \left[1 - \Phi\left(\frac{E + FX_i}{\sqrt{G}} \right) \right]^{1-R^Y}$$

Therefore, we can represent the 9 parameters as 7 parameters in the expression for the observed data likelihood, and the 7 parameters are estimable. We must fix 2 parameters in order for the remaining parameters to be (weakly) identified.

Suppose we fix ϕ_0 and ϕ_1. Then we can (weakly) identify all 7 remaining parameters under LMAR. However, suppose that we had observed Y for all subjects. In this case, we would need fix 4 parameters out of $(\alpha_0, \alpha_1, \alpha_2, \sigma^2, \beta_0, \beta_1, \Sigma_L)$ in order for the remaining 3 parameters to be identified. Therefore, the model fit without any outcome missingness requires some parameters to be fixed that do not need to be fixed in the LMAR-based model in order to achieve (weak) identifiability. Curiously, we have more information about the parameter set under LMAR than if we had observed Y for all subjects. It is worth noting that when we instead fix four parameters in $(\alpha_0, \alpha_1, \alpha_2, \sigma^2, \beta_0, \beta_1, \Sigma_L)$, the resulting parameters $A - G$ will be overidentified, but this should not present any problems.

It is important to note that we cannot verify the form of the missingness model, and here assumed missingness model results in additional parameters becoming identifiable under LMAR. Therefore, the identification is a direct result of unverifiable assumptions, and an analysis that relies on the missingness model being correct such that the outcome model parameters would not be identified if the model were incorrect seems untrustworthy. This provides further justification for excluding situations in which the parameters would not be identifiable if there was not covariate or outcome missingness.

While technically identified, our imputation algorithm leads to convergence problems when imputing under this LMAR model with only two fixed parameters (simulations not shown). If we fix additional parameters, the proposed imputation algorithm has better performance. In general, we do not expect our imputation algorithm to perform well in settings where the model would not be identified or would be very weakly identified if there were no covariate/outcome missingness. In such settings, we recommend fixing additional parameters to achieve good identification properties before performing the proposed imputation algorithm.

B.2 Example 1.2: Linear Mixed Model Example

We notice that the form of the measurement error model in the previous section is similar to the usual structure of a linear mixed model with a random intercept except that the outcome in the linear mixed model case is multivariate. Suppose we observe $K > 1$ values of Y for each subject and we assume that elements of Y within subjects are independent conditional that subject's covariates and the random intercept. We model:

$$Y_i|X_i, L_i \sim N_K(\alpha_0 + \mathbf{1}_K b_i + \alpha_2 X_i, \sigma^2 \mathbb{I}_K), \qquad b_i \sim N(0, \Sigma_L)$$

Here, $\mathbf{1}_K$ corresponds to α_1 in the previous measurement error model. Additionally, this model assumes that $\beta_0 = \beta_1 = 0$. Therefore, three parameters from the model in the previous section are fixed by design. The modeling assumptions imply the following joint distribution:

$$\begin{pmatrix} Y_i \\ L_i \end{pmatrix} |X_i = N\left(\begin{pmatrix} \alpha_0 + \alpha_2 X_i \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 \mathbb{I}_K + \mathbf{1}_K \Sigma_L \mathbf{1}_K^T & \mathbf{1}_K \Sigma_L \\ \mathbf{1}_K^T \Sigma_L & \Sigma_L \end{pmatrix}\right)$$

Suppose we have no missingness in Y. In this case, the observed data likelihood can be expressed as follows:

$$Lik_{NoMissing}^{(obs)} = \prod_{i=1}^{n} MVN_K\left(Y_i; \alpha_0 + \alpha_2 X_i, \sigma^2 \mathbb{I}_K + \mathbf{1}_K \Sigma_L \mathbf{1}_K^T\right)$$

We can identify all four of these model parameters. We compare this to the situation with the measurement error model with covariates in which 4 out of the 7 parameters needed to be fixed in order to achieve identifiability. In this case, three of the 7 parameters are fixed by design ($\alpha_2 = \mathbf{1}_K, \beta_0 = \beta_1 = 0$), and we can identify an additional parameter due to the compound symmetric structure of the variance for $Y|X$ resulting from the repeated measures within individuals. In this case, the model under no outcome or covariate missingness is well-identified, and the proposed imputation approach can perform well under some MAR and LMAR missingness scenarios.

Appendix C

Identifiability under LMAR for a Mixture of GLMs (Example 2)

In this section, we explore issues of identifiability for another simple modeling scenario. Unlike the measurement error example, this example demonstrates a situation in which the model is fully identified under no covariate/outcome missingness but has issues with identifiability under a simple LMAR missingness mechanism. We present simulations demonstrating evidence of identifiability-related numerical issues.

Suppose our model for outcome Y is a mixture of two GLMs and let C represent the fully latent mixing variable. Within each latent class, we model the relationship between Y and covariates X using a GLM. We will assume that $C \perp X$ with $P(C_i = 1|X_i) = \omega$. We first suppose there is no covariate/outcome missingness. The observed data likelihood takes the following form:

$$Lik^{(obs)}_{NoMissing} \propto \prod_{i=1}^{n} [\omega f(Y_i, |X_i, C_i = 1; \theta) + (1 - \omega) f(Y_i, |X_i, C_i = 2; \theta)]$$

Assuming the distribution of $Y|X, C$ depends on C and is an identifiable GLM in its own right, then θ and ω are both identifiable.

Suppose now that we have latent-dependent missingness in the outcome for some subjects. Let R^Y be a vector of indicators representing the response of Y. Let ϕ be the parameter attached to the missingness model. We define $p_j(\phi) = P(R_i = 1|X_i, C_i = j; \phi)$ for latent classes $j = 1, 2$. We can write the observed data likelihood as follows:

$$Lik^{(obs)}(\nu) \propto \prod_{i=1}^{n} \int \int f(R_i^Y|X_i, L_i; \phi) f(Y_i, |X_i, C_i; \theta) f(C_i|X_i; \omega) dY_i^{(mis)} dC_i^{(mis)}$$

$$\propto \prod_{i=1}^{n} [p_1(\phi)f(Y_i,|X_i,C_i=1;\theta)\omega + p_2(\phi)f(Y_i,|X_i,C_i=2;\theta)(1-\omega)]^{R_i^Y}$$
$$\times [(1-p_1(\phi))\omega + (1-p_2(\phi))(1-\omega)]^{1-R_i^Y}$$

C.1 Example 2.1: R^Y is Independent of X (Nonidentifiable Model)

First, we assume that R^Y is independent of X, so it only depends on C. Define $p_1(\phi) = \text{expit}(\phi_0+\phi_1)$ and $p_2(\phi) = \text{expit}(\phi_0)$. We can write the observed data likelihood as:

$$Lik^{(obs)}(\nu) \propto \prod_{i=1}^{n}\left[\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega f(Y_i,|X_i,C_i=1;\theta) + \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega)f(Y_i,|X_i,C_i=2;\theta)\right]^{R_i^Y}$$
$$\left[-\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega + 1 - \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega)\right]^{1-R_i^Y}$$

This likelihood can be reparameterized using $A = \frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega$ and $B = \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega)$, so we can represent three of the model parameters using just two parameters. Therefore, we will not be able to identify all three of ϕ_1, ϕ_0, and ω, but A and B can be identified. We suppose that θ is of primary interest. In this example, we can still identify θ even though we cannot identify ϕ_1, ϕ_0, and ω. We note that under MAR, $\phi_1 = 0$, and both ϕ_0 and ω are identified.

Under LMAR, we can identify A and B, but we cannot identify ϕ_1, ϕ_0, and ω. We want to know whether A and B are enough to perform the imputation of C and Y. In order to impute Y, we will draw from $f(Y_i|X_i,C_i)$, which does not involve ω or ϕ. We would impute C using:

$$P(C_i=1|X_i,Y_i) = \left[\frac{\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega f(Y_i|X_i,C_i=1)}{\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega f(Y_i|X_i,C_i=1) + \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega)f(Y_i|X_i,C_i=2)}\right]^{R_i^Y}$$
$$\times\left[\frac{\omega(1-\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}})f(Y_i|X_i,C_i=1)}{\omega(1-\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}})f(Y_i|X_i,C_i=1) + (1-\omega)(1-\frac{e^{\phi_0}}{1+e^{\phi_0}})f(Y_i|X_i,C_i=2)}\right]^{1-R_i^Y}$$

When we impute C and Y was observed, we are imputing using only functions of the parameters that ARE identifiable. However, imputation when Y is missing requires parameters that are not strictly identifiable. This may result in numerical issues within the imputation algorithm.

While we cannot identify all three of ϕ_1, ϕ_0, and ω, we can identify the other two parameters if we hold one parameter fixed. This provides a suggestion for imputation under this unidentifiable model. We can fix values of one of the parameters and then perform imputation. We can repeat this for different values of the fixed parameter and explore the impact of the fixed parameter on model inference.

C.2 Example 2.2: R^Y Depends on X (Identifiable Model)

Now, we assume that R^Y is not independent of X. Suppose we model $p_1(\phi) = \text{expit}(\phi_0 + \phi_1 X_i + \phi_2)$ and $p_2(\phi) = \text{expit}(\phi_0 + \phi_1 X)$. We can write

$$\begin{aligned} Lik(\nu) \propto \prod_{i=1}^{n} & \left[\frac{e^{\phi_0+\phi_1 X_i+\phi_2}}{1+e^{\phi_0+\phi_1 X_i+\phi_2}} \omega f(Y_i, |X_i, C_i = 1; \theta) + \frac{e^{\phi_0+\phi_1 X_i}}{1+e^{\phi_0+\phi_1 X_i}} (1-\omega) f(Y_i, |X_i, C_i = 2; \theta) \right]^{R_i^Y} \\ & \times \left[-\frac{e^{\phi_0+\phi_1 X_i+\phi_2}}{1+e^{\phi_0+\phi_1 X_i+\phi_2}} \omega + 1 - \frac{e^{\phi_0+\phi_1 X_i}}{1+e^{\phi_0+\phi_1 X_i}} (1-\omega) \right]^{1-R_i^Y} \end{aligned}$$

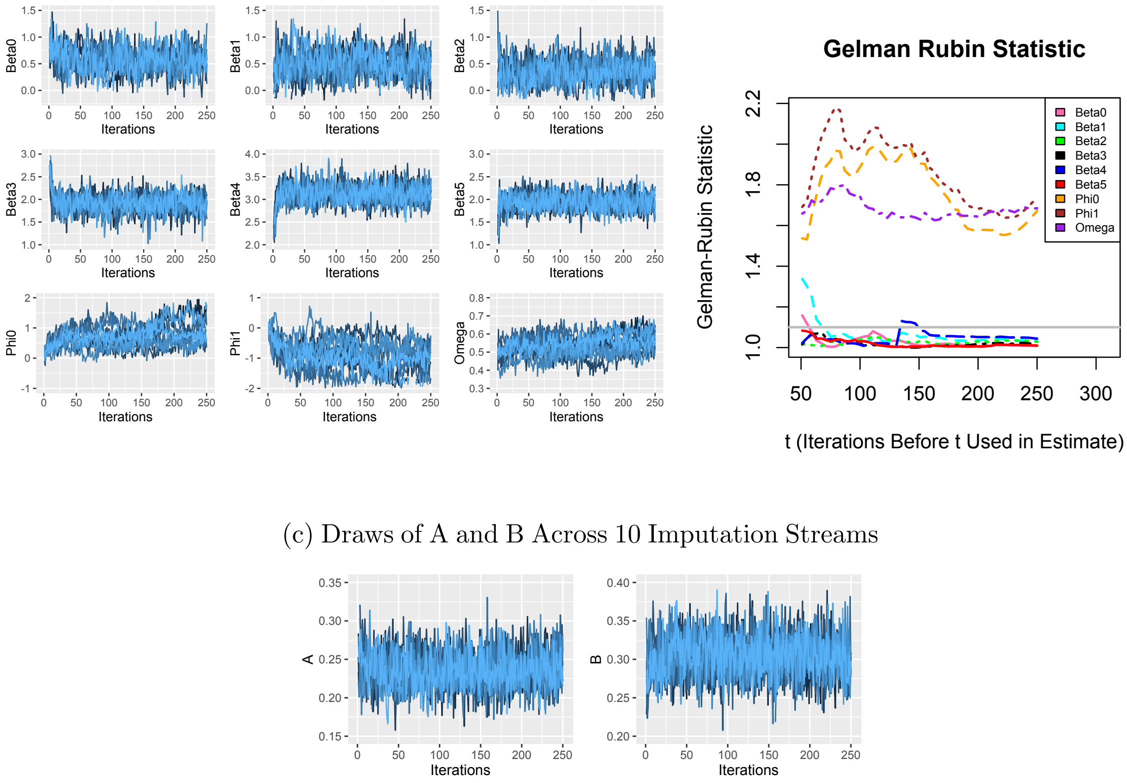
When ϕ_1 is nonzero, we can identify the model parameters. Therefore, additional complexity in the missingness mechanism results in an identifiable model.

C.3 Simulation using Nonidentifiable Model

We simulate a single dataset under a mixture of linear regressions model as in Simulation 3 in **Chapter III**. We impose outcome missingness using the relation $\text{logit}(P(Y \text{ is observed}|X_1, X_2, C, Y)) = \phi_0 + \phi_1 C$ where $\phi_0 = 1.1$ and $\phi_1 = -1.7$. Therefore, we have that $p_1(\phi) = \text{expit}(-0.6)$ and $p_2(\phi) = \text{expit}(1.1)$ (using notation from **Section C.1**). This is a LMAR mechanism. Define β to be the parameters of $f(Y|X, C)$ and $\omega = P(C = 1|X)$.

We first perform our imputation algorithm using a correct working model structure but without fixing values for ϕ_0 and ϕ_1. Previously, we showed in **Section C.1** that the parameters ϕ_0, ϕ_1, and ω are not all identifiable. However, at each iteration of the imputation algorithm, we can draw values of these three parameters. We perform 10 streams of our imputation algorithm in which we impute values of Y and L. **Figure C.1(a)** shows the parameter draws for each iteration of the imputation algorithm. Different imputation streams are shown with differently colored lines.

Figure C.1: Drawn Parameters in Nonidentifiable Model with No Fixed Parameters



(a) Parameter Draws Across 10 Imputation Streams

(b) Gelman-Rubin Diagnostics

(c) Draws of A and B Across 10 Imputation Streams

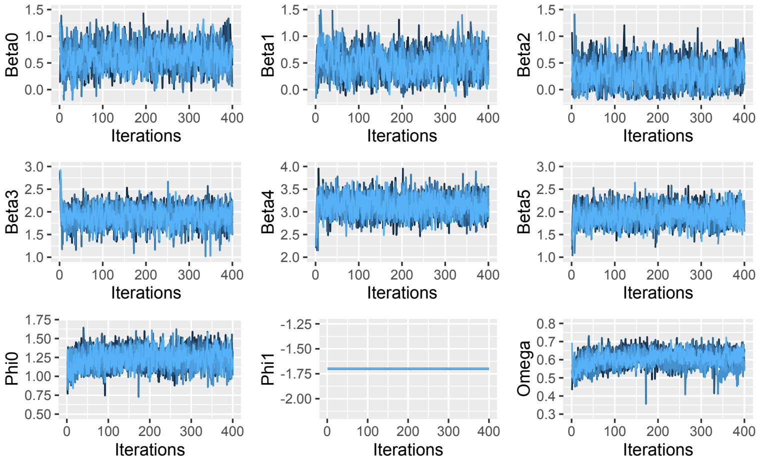
Visually, we can see that we have some issues with convergence for ϕ_1, ϕ_0, and ω.

However, the draws for the β parameters (the parameters ultimately of interest) appear to converge. One criterion for evaluating the convergence is the Gelman-Rubin statistic Gelman and Rubin (1992). This statistic is calculated by comparing the variation of the parameter draws within each stream to the variation between streams. For good algorithms, the value of this statistic should move toward 1 as the number of iterations increases, and values greater than 1.1 are generally considered to represent insufficient convergence. **Figure C.1(b)** shows the estimated Gelman-Rubin statistic for several model parameters across iterations of the imputation algorithm. We do not include the first 50 iterations in the calculations. The gray line represents a Gelman-Rubin statistic of 1.1. While the draws for the β parameters are converging, we do not see convergence for ϕ_0, ϕ_1, and ω. While we cannot identify ϕ_0, ϕ_1, and ω, we previously showed that functions A and B of these parameters are identifiable. **Figure C.1(c)** shows the parameter draws for A and B, and we can see that these parameters appear to converge nicely even though ϕ_0, ϕ_1, and ω do not.

Even though ϕ_0, ϕ_1, and ω are not all simultaneously identifiable, the parameter related to the outcome model can be identified. In terms of the practical implications of identifiability issues on inference, this hints that we may still be able to obtain reasonable inference about the outcome model parameter in some cases. In this simulation, the β parameters do appear to converge to values that are very close to the true values even in the presence of convergence issues for the other parameters.

While we cannot identify ϕ_0, ϕ_1, and ω simultaneously, we can identify two of the parameters if we fix values of the third. Fixing ϕ_1, we perform imputation drawing values for all other parameters. **Figure C.2** shows the resulting parameter draws across the 10 streams of imputation. When we fix ϕ_1, we see good numerical convergence properties for the other model parameters.

Figure C.2: Parameter Draws Across 10 Imputation Streams when ϕ_1 is Fixed



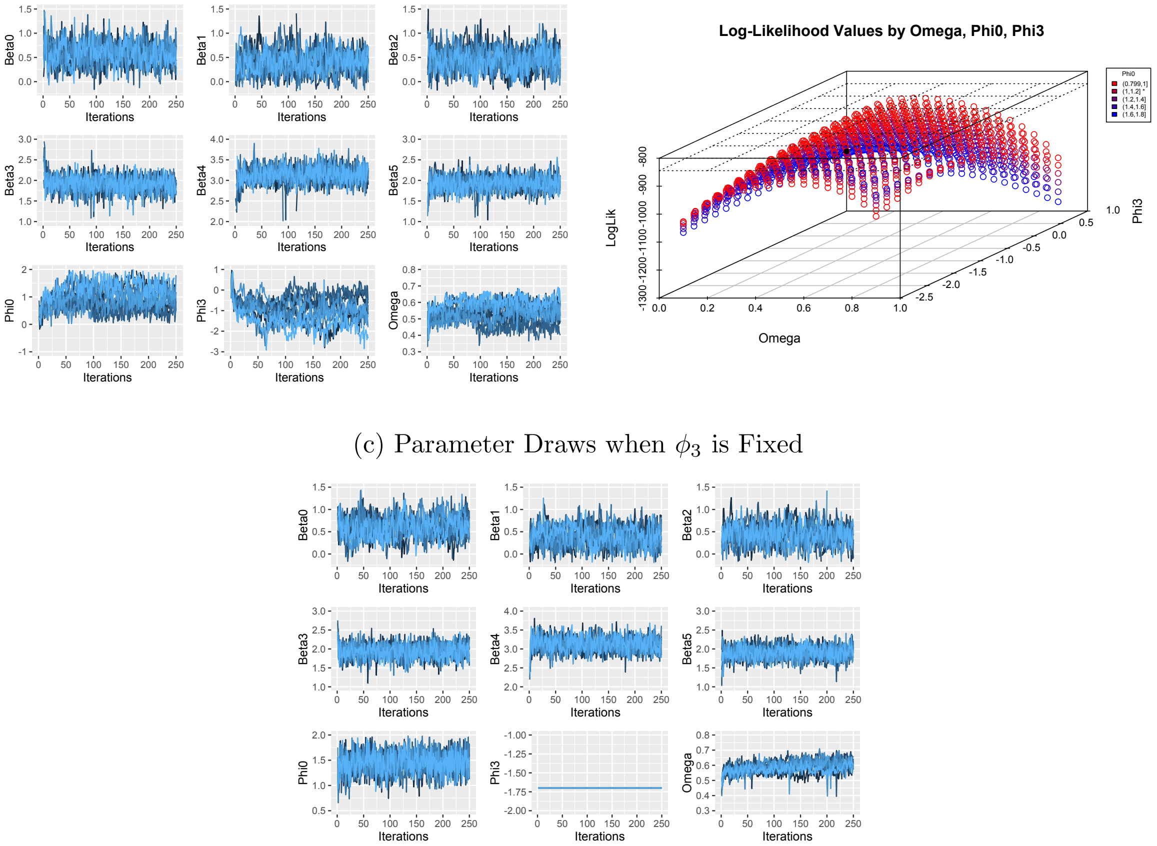
C.4 Simulation using Identifiable Model

We now consider the setting where missingness in the outcome is generated using the relation logit$(P(Y \text{is observed}|X_1, X_2, C, Y)) = \phi_0 + \phi_1 X_1 + \phi_2 X_2 + \phi_3 C$ where $\phi_0 = 1.1$, $\phi_1 = 0.5$, $\phi_2 = -0.5$, and $\phi_3 = -1.7$. Again, this is a LMAR mechanism.

We first perform imputation of Y and L using the correct working model without fixing any parameter values. **Figure C.3(a)** shows the parameter draws for the 10 imputation streams. We can see evidence of convergence issues for several model parameters. However, we still see that the parameters of interest in θ appear to converge nicely near their true values.

While the parameters may all be technically identifiable, we can sometimes run into problems when the observed data log-likelihood surface is nearly flat with respect to one or more parameters. **Figure C.3(b)** shows the value of the observed data log-likelihood for different values of ϕ_0, ϕ_3, and ω using the true values for all other parameters. The plotted plane indicates the maximum of the observed data log-likelihood and the black dot indicates the true values for the parameters. Fixing ϕ_3 and ϕ_0, we can see that the shape of the log-likelihood with ω is fairly concave. However, the log-likelihood surface as a whole is fairly flat across different combinations of ϕ_0, ϕ_3, and ω. When we fix the value of ϕ_3, however, we can do a better job at estimating ω and ϕ_0, resulting in improved convergence performance as shown in **Figure C.3(c)**.

Figure C.3: Drawn Parameters in Identifiable Model



(a) Parameter Draws with No Fixed Parameters

(b) Log-Lik Surface with Respect to ϕ_0, ϕ_3, and ω

(c) Parameter Draws when ϕ_3 is Fixed

Appendix D

Implementation of LMAR-based Imputation Algorithm under Various Outcome Models

In this section, we provide specifics for how we can implement the proposed imputation algorithm for the three examples of latent ignorability considered in **Chapter III**. In each case, we will use notation defined in **Chapter III** and use R^{-S} as defined in *Lemma 3*. We will assume we are using flat priors for all model parameters. This assumption allows us to draw parameter values using maximum likelihood methods on bootstrap samples of the data.

D.1 Drawing from a Distribution Known up to Proportionality

In **Chapter III**, we present distributions we can use to impute missing values for latent variables, but in some cases these distributions may only known up to proportionality. We call the form of the distribution known up to proportionality the "kernel" of the distribution. Many methods exist in the literature for drawing from a distribution knowing only the kernel. In this section, we will *briefly* describe two such methods.

Rejection Sampling

The strategy of rejection sampling is to determine a easy-to-draw-from distribution that dominates a hard-to-draw-from distribution. We can then draw values from the hard-to-draw-from distribution by instead drawing from the easy-to-draw-from distribution distribution many times and accepting the first draw that satisfies a simple inequality. In more concrete terms, rejection sampling algorithms involve determining a simple density, $g(v)$, that dominates the distribution known up to proportionality, $k(v)$, such that we can write

$$k(v) \leq Kg(v) \;\forall\; v$$

where K is a constant greater than or equal to 1. Once we have specified a density $g(v)$ that dominates $k(v)$, we can obtain a draw V from $k(v)$ by performing the following:

1) Generate V from $g(v)$ and U from $U(0, 1)$

2) Accept draw V if $U \leq \frac{k(V)}{Kg(V)}$. Otherwise, we reject draw V and return to 1) (Robert and Casella, 2004).

If $Kg(v)$ is much larger than $k(v)$, the rejection sampling algorithm may require many repetitions in order to accept a draw. Therefore, the choice of $g(v)$ and K is important to the efficiency of the imputation algorithm. In the following sections, we propose possible choices for K and $g(v)$ in specific settings, but more efficient choices may be available.

Rejection sampling methods for imputation knowing the distribution only up to proportionality were considered in Bartlett et al. (2014), which used dominating function $f(X_i^{(t)}|X_i^{(-t)}; \psi)$ for covariate imputation. We can use a similar approach for covariate imputation as discussed below.

Metropolis-Hastings

Like the rejection sampling algorithm, the goal of the Metropolis-Hastings algorithm is to obtain a draw values of variable V from a distribution known only up to proportionality, $k(v)$. The strategy is to first specify a proposal distribution, $p(v|u)$, from which we propose new values for the variable $V = v$ given the most recent drawn value of V, u. We can obtain a draw V from $k(v)$ by performing the following:

1) Generate v^* from $p(v|u)$. Generate $U \sim U(0, 1)$

2) Define acceptance probability $\alpha = \min\left(1, \frac{p(u|v^*)k(v^*)}{p(v^*|u)k(u)}\right)$. Accept draw $V = v^*$ if $U \leq \alpha$.

Otherwise, we reject draw $V = v^*$ and keep $V = u$ (Robert and Casella, 2004).

One popular choice of proposal distributions is a normal distribution centered at the most recent imputation u and with variance as a tuning parameter.

D.2 Linear Mixed Model with Random Intercept

Suppose our outcome model is a linear mixed model with a latent random intercept, b_i. Let outcome Y_i be a vector of $K > 1$ normal outcomes and X_i be a $K \times d$ matrix containing a column of 1's and covariates for subject i. We model

$$Y_i | X_i, b_i \sim N_K(X_i\theta + 1_K b_i, \Sigma) \quad \text{and} \quad b_i | X_i \sim N(0, \omega^2)$$

We have the following joint distribution:

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} | X_i = N\left(\begin{pmatrix} X_i\theta \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma + \mathbf{1}_K\omega^2\mathbf{1}_K^T & \mathbf{1}_K\omega^2 \\ \mathbf{1}_K^T\omega^2 & \omega^2 \end{pmatrix}\right)$$

In this modeling framework, random intercept b_i is missing for all subjects. Suppose we also have missingness in Y and X that may be MAR or LMAR. We also suppose that $\Sigma = \sigma^2\mathbb{I}_K$, so the outcomes are independent across subjects given b and X. We can use the imputation algorithm described below to impute missing values in b_i, X, and Y. We can initialize the missing values of the covariates by drawing from the observed values with equal probability. We can initialize the latent random intercept using the Best Linear Unbiased Predictors (BLUPs) from a complete case fit.

Imputation of Latent Variable

Assuming MAR

Under MAR and using (3.6), we want to impute missing b_i from

$$f(b_i|X_i, Y_i; \nu) \propto f(Y_i|X_i, b_i; \theta) f(b_i|X_i; \omega) = f(b_i|X_i, Y_i; \rho)$$

Using properties of multivariate normal random variables, we have that

$$f(b_i|X_i, Y_i; \rho) = N(\mathbf{1}_K^T\omega^2 \left[\Sigma + \mathbf{1}_K\omega^2\mathbf{1}_K^T\right]^{-1}(Y_i - X_i\theta), \omega^2 - \mathbf{1}_K^T\omega^2\left[\Sigma + \mathbf{1}_K\omega^2\mathbf{1}_K^T\right]^{-1}\mathbf{1}_K\omega^2)$$

We can draw values of Σ, ω^2, and θ by fitting a linear mixed model to a bootstrap sample of the most recently imputed data and then draw missing b_i from $f(b_i|X_i, Y_i; \rho)$.

Assuming LMAR

Under LMAR and using (3.6), we want to impute missing b_i from

$$f(b_i|X_i, Y_i, R_i^{-S}; \nu) \propto f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, b_i; \phi^{-S}) f(b_i|X_i, Y_i; \rho) \tag{D.1}$$

This distribution depends on R_i^{-S}, the subset of R_i corresponding to variables that are LMAR. We must specify a model for R_i^{-S} given $Y_i^{(obs)}, X_i^{(obs)}$, and b_i. When R_i^{-S} contains missingness indicators for multiple variables (e.g. outcome at different time-points), this may be a challenging task. Several authors have discussed specification of this missingness model in the context of missingness dependent on random effects, and we will not discuss this choice further here (Wu and Carroll, 1988; Yang et al., 2008).

The distribution in (D.1) is only known up to proportionality, but we can use one of the two above methods for drawing from a distribution knowing only the kernel. For example, we may use Metropolis-Hastings methods to draw values of b_i with a normal proposal distribution centered at the most recent imputed value of b_i and with some small variance, τ, which will be a tuning parameter. Given τ, the most recent imputation of D, and draws of ρ and ϕ, we can use the above kernel to impute b_i under LMAR.

Another option is to use rejection sampling. We note that $f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, b_i; \phi^{-S})$ is a probability, so it is less than or equal to 1. We define

$$k(b_i) = f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, b_i; \phi^{-S}) f(b_i|X_i, Y_i; \rho)$$

and can define dominating function $g(b_i) = f(b_i|X_i, Y_i; \rho)$ with $K = 1$. $g(b_i)$ is a normal distribution with mean and variance as functions of model parameters, so this distribution is easy to draw from. We can then perform the following algorithm to impute b_i:

1) Generate V from $g(b_i) = f(b_i|X_i, Y_i; \rho)$ and U from $U(0, 1)$

2) Accept draw $V = b_i$ if $U \leq \; f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, V; \phi^{-S})$.

Otherwise, we reject draw V and return to 1).

Under LMAR, we can obtain a draw of ρ using the same approach as under MAR. We can obtain a draw of ϕ by fitting our specified model for the missingness given $Y_i^{(obs)}, X_i^{(obs)}$, and b_i to a bootstrap sample of the data and using the most recent imputation of b_i.

Imputation of Missing Covariates and Outcomes

Covariates

We also note that X_i as defined in the above equation is a matrix. In the notation developed in **Section 3.2**, covariate set X_i represents a vector. Therefore, we have some notation mismatch that we will need to rectify in order to apply (3.8) for imputation. Let $Z_i^{(t)}$ represent the vector of elements corresponding to covariate t for subject i and $Z_i^{(-t)}$ be a stacked vector containing the remaining elements of X_i that are not in $Z_i^{(t)}$. We note that by assumption, $b_i|X_i$ does not depend on X_i. Using this notation, we can impute missing $Z_i^{(t)}$ (and therefore the missing values for the t^{th} variable in X_i) using:

$$\begin{aligned} f(Z_i^{(t)}|Z_i^{(-t)}, Y_i, b_i; \rho) &\propto f(Y_i|X_i, b_i; \theta) f(b_i|X_i; \omega) f(Z_i^{(t)}|Z_i^{(-t)}; \psi) \\ &\propto f(Y_i|X_i, b_i; \theta) f(Z_i^{(t)}|Z_i^{(-t)}; \psi) \end{aligned}$$

In this case, $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ is a *multi-dimensional* distribution. For example, $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ may be multivariate normal.

We can obtain imputations of $Z_i^{(t)}$ by performing a block-wise Metropolis-Hastings draw. In settings with where $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ is not easy to draw from, we recommend this approach. Alternatively, we could perform the following rejection sampling procedure. Define $k(Z_i^{(t)}) = f(Y_i|X_i, b_i; \theta) f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ and $g(Z_i^{(t)}) = f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$. We want to find a constant that dominates $f(Y_i|X_i, b_i; \theta)$ across different values of $Z_i^{(t)}$. We note that $f(Y_i|X_i, b_i; \theta)$ is multivariate normal by assumption, and its maximum value across all covariate values will occur when $Y_i = X_i\theta + 1_K b_i$, at which point $f(Y_i|X_i, b_i; \theta) = \frac{1}{\sqrt{|2\pi\Sigma|}}$. Define $K = \frac{1}{\sqrt{|2\pi\Sigma|}}$. We can then impute $Z_i^{(t)}$ jointly using the following rejection sampling algorithm:

1) Generate V from $g(Z_i^{(t)}) = f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ and U from $U(0, 1)$

2) Accept draw $V = Z_i^{(t)}$ if

$$U \leq \frac{f(Y_i|X_i, b_i; \theta)}{K}|_{Z_i^{(t)}=V} = e^{-\frac{1}{2}(Y_i - X_i\theta - 1_K b_i)^T \Sigma^{-1} (Y_i - X_i\theta - 1_K b_i)}|_{Z_i^{(t)}=V}$$

Otherwise, return to 1).

We note that the above imputation algorithm allows the elements of $Z_i^{(t)}$ to take different values. Suppose the covariate represented by $Z_i^{(t)}$ is time-independent. Then we would want the elements of $Z_i^{(t)}$ to be equal. We can impose this property by defining $f(Z_i^{(t)}|Z_i^{(-t)};\psi)$ such that it requires all of the elements of $Z_i^{(t)}$ to be equal. In this case, the rejection sampling algorithm would be simple to perform.

Imputation using the above approach requires draws of Σ, θ, and ψ. We can use the drawn values of Σ and θ from the step for imputing the random intercept. However, suppose we want to draw new values for the parameters conditional on the imputed values of b. Since we assumed that $\Sigma = \sigma^2\mathbb{I}_K$ (so the elements of Y_i are independent given b_i), we can draw Σ and θ by fitting a linear regression model to Y treating the elements of Y_i as independent and using offset term b_i for all elements in Y_i (to a bootstrap sample of the data). We can draw ψ by fitting a model for $Z_i^{(t)}|Z_i^{(-t)}$ to a bootstrap sample.

Outcomes

We note that Y_i is a vector in this case. We can impute the t^{th} element of Y_i using:

$$f(Y_i^{(t)}|Y_i^{(-t)}, b_i; \rho) \propto f(Y_i|X_i, b_i; \theta)$$

Since the elements of Y_i are multivariate normal by assumption, we can easily work out this conditional distribution. This distribution simplifies further when we assume that the elements of Y_i are independent given b_i and X_i. In this case, we can impute $Y_i^{(t)}$ from a normal distribution with mean equal to the t^{th} element of $X_i\theta$ and variance σ^2. We can draw θ and σ^2 as we do for covariate imputation.

Final Analysis

We can use the above imputation method to obtain M imputed datasets. We can then fit a model to each of the imputed datasets and use Rubin's combining rules to obtain a single set of parameter estimates and standard errors. As discussed in **Chapter III**, there are several different ways we can perform the final analysis for any given imputed dataset. If we choose to use the imputed random intercept values, we can estimate θ by fitting a linear regression with offset term b_i. For this fit, we can either use or ignore the imputed D. We can estimate ω^2 as the sample variance of the imputed b_i. Alternatively, we can ignore the imputed random intercept values and fit linear mixed model using the imputed values for D. This approach may be simpler and more stable in practice, but it may not be fully efficient in the LMAR setting as shown in *Lemma 4*.

Brief Comparison to Some Existing Methods

Imputation-based approaches for dealing with missing linear mixed model outcome data under MAR have been explored extensively in the literature. The proposed approach under MAR is very similar to existing Gibbs Sampler-based approaches (e.g. Schafer and Yucel, 2002). Unlike other Gibbs Sampling approaches, our method for imputing b_i involves drawing parameters from a distribution that does not condition on the imputed values for b_i and imputes missing data sequentially rather than jointly. Additionally, in our application of the proposed methods, we assume flat priors for all model parameters. This assumption substantially simplifies the step for drawing model parameters in practice.

Yang et al. (2008) describes a two-stage imputation approach for linear mixed models with intermittent MAR outcome missingness and LMAR dropout. Unlike Yang et al. (2008), we propose performing imputation of all outcome missingness (from different causes) in a single stage. Missing outcome values are imputed under the same model regardless of the mechanism generating the missingness, and information about different sources of missingness can be incorporated into the missingness model used to impute the latent variable. Additionally, Yang et al. (2008) takes a Gibbs Sampling approach, and the steps for drawing the parameter values can be complicated and themselves require methods for sampling from distributions known only up to proportionality. In the proposed algorithm, parameter draws under uniform priors can be obtained my fitting models using MLE methods to a bootstrap sample of the data. This substantially simplifies the parameter drawing.

D.3 Cox Proportional Hazards Cure Model Algorithm

We define indicator G_i that takes the value 1 if subject i is not cured and 0 if subject i is cured. Let T_i be the observed event or censoring time and δ_i be the event indicator. We have $Y_i = (T_i, \delta_i)$. Let X_i be a set of covariates. The CPH mixture cure model consists of 1) a logistic regression for the probability of being "not cured" [$\text{logit}(P(G_i = 1|X_i)) = \omega_0 + \omega_1 X_i$] and 2) a Cox proportional hazards model for the event hazard in the "not cured" group $\left[\lambda(t) = \lambda_0(t)e^{\theta X_i}\right]$.

We recall that non-cure status, G_i, is partially latent. For subjects with observed events ($\delta_i = 1$), we know that $G_i = 1$. We may also assume that subjects still at risk by a certain time t are cured ($G_i = 0$). For all other subjects, G_i is unknown. In addition to missingness in cure status, suppose we have ignorable or latent ignorable missingness in covariates X. We can use the imputation algorithm proposed in **Chapter III** to iteratively impute values for the latent variable and the covariates. Below, we present some details for the approach for imputing the latent variable and covariates. We can initialize the missing values of the latent variable and the covariates from drawing from the observed values with equal probability.

Imputation of Latent Variable

Assuming MAR

We will first assume that missingness in X_i is MAR. In this case, we can impute G_i using the following relation derived from (3.6):

$$\text{logit}(P(G_i = 1|X_i, T_i, \delta_i = 0; \rho)) = \omega_0 + \omega_1 X_i - \Lambda_0(T_i)e^{\theta X_i}$$

This imputation distribution depends on the most recent imputed values for X_i, parameters ω and θ, and the cumulative baseline hazard function, $\Lambda_0(t)$. An identical imputation distribution was proposed in **Chapter II** for imputing cure status in the Cox proportional cure model setting under MAR. In **Chapter II**, $\Lambda_0(t)$ is estimated using a weighted Breslow-type estimator at each iteration of the imputation algorithm, and we can use the same estimation approach here. We can draw values for ρ by fitting a Cox proportional cure model to a bootstrap sample of the most recent imputed data or by fitting a cure model to the most recent imputed data and draw ρ from a multivariate normal distribution with mean and variance from the cure model fit.

Assuming LMAR

Now, we assume missingness in X_i is LMAR. From (3.6), we can impute G_i using

$$\begin{aligned}\text{logit}(P(G_i = 1|X_i, T_i, \delta_i = 0, R_i; \nu)) =& \omega_0 + \omega_1 X_i - \Lambda_0(T_i)e^{\theta X_i} \\ &+ \log\left[\frac{f(R_i^{-S}|T_i, \delta_i = 0, X_i^{(obs)}, G_i = 1; \phi^{-S})}{f(R_i^{-S}|T_i, \delta_i = 0, X_i^{(obs)}, G_i = 0; \phi^{-S})}\right]\end{aligned}$$

This distribution differs from the one used under MAR by an offset term on the logit scale. When the difference in the missingness distribution by cure status is small, the offset term will be near zero. This distribution again depends on the cumulative baseline hazard function, $\Lambda_0(t)$, which can be estimated as in the MAR case. It also depends on ω, θ, and ϕ. We also must specify a model for missingness of the set of indicators that are conditionally dependent on L_i, R_i^{-S}.

We can draw θ and ω using the same approach as in the MAR case (ignoring the most recent imputations of L). We can draw ϕ by fitting a model for R_i^{-S} to a bootstrap sample of the data using the most recent imputation of cure status.

Imputation of Missing Covariates

By (3.8), we can impute missing values for covariate $X^{(t)}$ using:

$$\begin{aligned}f(X_i^{(t)}|X_i^{(-t)}, Y_i, G_i; \rho) &\propto [P(G_i = 1|X_i; \omega) f(Y_i|X_i, G_i; \theta)]^{G_i} P(G_i = 0|X_i; \omega)^{1-G_i} f(X_i^{(t)}|X_i^{(-t)}; \psi) \\ &\propto \left[\frac{e^{\omega_0 + \omega_1 X_i}}{1 + e^{\omega_0 + \omega_1 X_i}} \left(\lambda_0(T_i)e^{\theta X_i}\right)^{\delta_i} e^{-\Lambda_0(T_i)e^{\theta X_i}}\right]^{G_i} \left[\frac{1}{1 + e^{\omega_0 + \omega_1 X_i}}\right]^{1-G_i} f(X_i^{(t)}|X_i^{(-t)}; \psi) \qquad \text{(D.2)}\end{aligned}$$

When $X_i^{(t)}$ is categorical, we can easily use the above expression to derive the full form of the distribution used for imputation. For example, imputation of a binary covariate. Then imputation can proceed using the following relation:

$$P(X_i^{(t)} = 1|X_i^{(-t)}, Y_i, G_i; \rho) = \frac{(D.2)|_{X_i^{(t)}=1}}{(D.2)|_{X_i^{(t)}=1} + (D.2)|_{X_i^{(t)}=0}}$$

When $X_i^{(t)}$ has continuous structure, the imputation distribution may only be known up to proportionality. We can use Metropolis-Hastings methods to draw missing $X_i^{(t)}$ from (D.2) using a proposal distribution centered at the most recent imputation of $X_i^{(t)}$. Alternatively, we could use the following rejection sampling algorithm: Define $k(X_i^{(t)})$ =(D.2). We note that

$$\begin{aligned} k(X_i^{(t)}) &\leq \left[\left(\lambda_0(T_i)e^{\theta X_i}\right)^{\delta_i} e^{-\Lambda_0(T_i)e^{\theta X_i}}\right]^{G_i} f(X_i^{(t)}|X_i^{(-t)};\psi) \\ &\leq \left[f(T_i|X_i, G_i = 1)\right]^{\delta_i} f(X_i^{(t)}|X_i^{(-t)};\psi) \end{aligned}$$

Suppose we define

$$K = (1-\delta_i) + \delta_i \max_{X_i^{(t)}} f(T_i|X_i, G_i = 1)$$

so K takes the value 1 if $\delta_i = 0$ and takes the maximum of the event time distribution function across $X_i^{(t)}$ if $\delta_i = 1$. This maximum can usually be easily calculated given parameter values when the baseline hazard is parametric. We further define $g(X_i^{(t)}) = f(X_i^{(t)}|X_i^{(-t)};\psi)$. Then we have that $k(X_i^{(t)}) \leq Kg(X_i^{(t)})$. Then we can obtain a draw of $X_i^{(t)}$ from $k(X_i^{(t)})$ through the following algorithm:

1) Generate V from $g(X_i^{(t)}) = f(X_i^{(t)}|X_i^{(-t)};\psi)$ and U from $U(0,1)$

2) Accept draw $V = X_i^{(t)}$ if $U \leq \frac{\left[\frac{e^{\omega_0+\omega_1 X_i}}{1+e^{\omega_0+\omega_1 X_i}}\left(\lambda_0(T_i)e^{\theta X_i}\right)^{\delta_i} e^{-\Lambda_0(T_i)e^{\theta X_i}}\right]^{G_i}\left[\frac{1}{1+e^{\omega_0+\omega_1 X_i}}\right]^{1-G_i}}{K}$.

Otherwise, return to 1).

Imputation by (D.2) requires draws of ω, θ, and ψ. We can either use the draws of ω and θ obtained in the imputation step for the latent variable or draw new values. If we draw new values, we should use methods that use the most recent imputation of L. We can then draw θ by fitting a Cox regression to a bootstrap sample of the subjects with imputed $G = 1$. We can draw ω by fitting a logistic regression to G for a bootstrap sample of the entire dataset. We can draw ψ by fitting a model for $X_i^{(t)}|X_i^{(-t)}$ to a bootstrap sample.

Final Analysis

We can use the above imputation method to obtain M imputed datasets. We can then fit a model to each of the imputed datasets and use Rubin's combining rules to obtain a single set of parameter estimates and standard errors. There are several different ways we can perform the final analysis for any given imputed dataset. If we choose to use the imputed G, we can estimate θ by fitting a Cox regression to the subjects with imputed $G = 1$, and we can estimate ω by fitting a logistic regression for G. For these fits, we can either use or ignore the imputed D. Alternatively, we can ignore the imputed G and fit cure model using the imputed values for D. We recommend this last approach.

Brief Comparison to Some Existing Methods

Chapter II explores covariate imputation for the Cox proportional hazards cure model under MAR assumptions, and our proposed algorithm under MAR is very similar with some small differences in the methods for drawing parameters. We believe we are the first to explore covariate imputation for the Cox proportional hazards cure model under LMAR assumptions.

D.4 Mixture of GLMs

Suppose our outcome Y is generated from a mixture of K generalized linear models (GLMs) where K is known. Let C_i be a fully latent mixing variable indicating which element of the mixture distribution generated the observation for subject i. Missingness in C_i can be viewed as MCAR with probability 1. We suppose the distribution of $Y_i|X_i, C_i = j$ is modeled using a GLM (e.g. normal, logistic, Poisson) for $j = 1, \ldots, K$ and that the distribution for $C_i|X_i$ is independent of X_i.

We suppose that we have ignorable or latent ignorable missingness in Y and/or X. We can use the proposed methods for imputation. We can initialize the missing values of the covariates from drawing from the observed values with equal probability. We can initialize C based on the estimated probabilities $P(C_i = 1)$ obtained by fitting a latent class model to the complete case data.

Imputation of Latent Variable

Assuming MAR

The imputation distribution for the latent mixing variable C_i under MAR can be easily worked out based on the kernel in (3.6) to be multinomial with corresponding probabilities as follows:

$$P(C_i = j|X_i, Y_i, R_i; \nu) = \frac{f(Y_i|X_i, C_i = j; \theta)P(C_i = j; \omega)}{\sum_{l=1}^{K} f(Y_i|X_i, C_i = l; \theta)P(C_i = l; \omega)}$$

We can obtain a draw of θ and ω by fitting a latent class model to a bootstrap sample the most recently imputed data. In R, we can perform this latent class model fit using the package *flexmix* (Leisch, 2004). This package will estimate θ and ω for a specified number of latent classes, but it cannot differentiate between the different class labels. Therefore, we will need to impose a restriction to relate the latent classes identified by *flexmix* to values of C.

Assuming LMAR

Under LMAR, we can impute missing values of C_i using:

$$P(C_i = j|X_i, Y_i, R_i; \nu) = \frac{f(R_i^{-S}|X_i^{(obs)}, C_i = j, Y_i^{(obs)}; \phi^{-S})f(Y_i|X_i, C_i = j; \theta)P(C_i = j|X_i; \omega)}{\sum_{l=1}^{K} f(R_i^{-S}|X_i^{(obs)}, C_i = l, Y_i^{(obs)}; \phi^{-S})f(Y_i|X_i, C_i = l; \theta)P(C_i = l|X_i; \omega)}$$

This imputation distribution requires us to model R_i^{-S}. Draws of θ and ω can be obtained as in the MAR case. We can obtain a draw of ϕ by fitting a model for R_i^{-S} to a bootstrap sample of the data using the most recent imputation of C.

Imputation of Missing Covariates and Outcome

Covariates

By (3.8) and since $f(C_i|X_i; \omega) = f(C_i; \omega)$ by assumption, we can impute missing values for covariate $X^{(t)}$ using:

$$f(X_i^{(t)}|X_i^{(-t)}, Y_i, C_i; \rho) \propto f(Y_i|X_i, C_i; \theta)f(X_i^{(t)}|X_i^{(-t)}; \psi)$$

When $X_i^{(t)}$ is categorical, we can easily use the above expression to derive the full form of the distribution used for imputation. Otherwise, we can use methods to draw from the above distribution known only up to proportionality. For example, we can use the following rejection sampling algorithm: Define $k(X_i^{(t)}) = f(Y_i|X_i, C_i; \theta)f(X_i^{(t)}|X_i^{(-t)}; \psi)$ and $g(X_i^{(t)}) = f(X_i^{(t)}|X_i^{(-t)}; \psi)$. Define

$$K = \max_{X_i^{(t)}} f(Y_i|X_i, C_i; \theta)$$

Then we have that $k(X_i^{(t)}) \leq Kg(X_i^{(t)})$. Then we can obtain a draw of $X_i^{(t)}$ from $k(X_i^{(t)})$ through the following algorithm:

1) Generate V from $g(X_i^{(t)}) = f(X_i^{(t)}|X_i^{(-t)}; \psi)$ and U from $U(0, 1)$

2) Accept draw $V = X_i^{(t)}$ if $U \leq \frac{f(Y_i|X_i,C_i;\theta)}{K}$.

Otherwise, return to 1).

Imputation using the above method requires draws of ω, θ, and ψ. We can draw θ by fitting a GLM (or multiple GLMS) to a bootstrap sample of subjects using the most recent imputation of C. We can draw ω by looking at the proportion of subjects with

$C = j$ for each j in a bootstrap sample of the data. We can draw ψ by fitting a model for $X_i^{(t)}|X_i^{(-t)}$ to a bootstrap sample.

Outcome

We will assume here that Y is univariate. By (3.7), we can impute missing values for outcome Y using:

$$f(Y_i|X_i, C_i; \rho) = f(Y_i|X_i, C_i; \theta)$$

We can obtain a draw for θ as in covariate imputation and then draw missing values of Y simply using the GLM corresponding to the most recent imputed value for C_i.

Final Analysis

We can use the above imputation method to obtain M imputed datasets. We can then fit a model to each of the imputed datasets and use Rubin's combining rules to obtain a single set of parameter estimates and standard errors. There are several different ways we can perform the final analysis for any given imputed dataset. If we choose to use the imputed C, we can estimate θ by fitting a GLM for $f(Y|C, X)$ using the imputed C and either using or ignoring the imputed D. Alternatively, we can ignore the imputed C and fit a latent class model (e.g. using *flexmix*) using the imputed values for D. This second approach would require us to use an identifying assumption to determine which cluster identified by the latent class modeling corresponds to which value of C. We recommend this second approach.

Brief Comparison to Some Existing Methods

Many authors have explored similar imputation approaches for mixtures of GLMs under MAR assumptions, but comparatively little work has been done exploring LMAR missingness in this setting (e.g. Vidotto et al., 2015). Jung (2007) considers the case of a multivariate outcome related to a categorical latent mixing variable and proposes a MCMC imputation scheme that iteratively imputes missing values of the outcome and C. Additionally, Jung (2007) assumes the outcome is independent of the covariates given C. Our proposed approach can be viewed as a generalization of the approach in Jung (2007) that can handle LMAR missingness in the outcome and covariates while also allowing for conditional dependence between Y and X.

Appendix E

Estimation of Baseline Hazards for EM Algorithm

In this section, we use profile likelihood to derive estimators for the nonparametric baseline hazard functions used within the EM algorithm. In order to perform the E-step of the EM algorithm, we require estimates for $\lambda_{14}(t)$, $\lambda_{24}(t)$, and $\lambda_{13}(t)$. The form of the estimators for $\lambda_{14}(t)$ and $\lambda_{24}(t)$ depends on the assumptions imposed on the baseline hazards. For these derivations, we will assume that we use the same set of predictors, X_i in the model for each of the transitions. However, these estimators are easily generalized to allow the covariate sets to differ across transitions.

$2 \to 4$ and $1 \to 4$ Baselines Unrestricted

Suppose we have estimates of θ from the previous M-Step and an estimated p_i from the previous E-Step. We would like to maximize the expected log-likelihood with respect to the baseline hazard functions. Suppose we do not assume any relationship between $\lambda^0_{14}(t)$ and $\lambda^0_{24}(t)$. We consider the contributions of each of the baseline hazards to the (expected) log-likelihood. For each hazard, the contribution is:

$$\begin{aligned}
\lambda^0_{14}(t): &\qquad \sum_{i=1}^{n} p_i \log\left[\lambda_{14}(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}(Y_{ir})\}\right] \\
\lambda^0_{24}(t): &\qquad \sum_{i=1}^{n} (1-p_i) \log\left[\lambda_{24}(Y_{id})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{24}(Y_{id})\}\right] \\
\lambda^0_{13}(t): &\qquad \sum_{i=1}^{n} p_i \log\left[\lambda_{13}(Y_{ir})^{\delta_{ir}}\exp\{-\Lambda_{13}(Y_{ir})\}\right] \\
\lambda^0_{34}(t): &\qquad \sum_{i=1}^{n} \delta_{ir} \log\left[\lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}}\exp\{-\Lambda_{34}(Y_{id}-Y_{ir})\}\right]
\end{aligned}$$

In obtaining the above equations, we used that $\delta_{ir} = 0$ for all subjects with nonzero $1 - p_i$ and that, under equal follow-up, $Y_{ir} = Y_{id}$ for all subjects with $\delta_{ir} = 0$. Let T be the event time, D be the event indicator, and W be the weight in one of the above cases. Each of the contributions take the form

$$\begin{aligned}\tilde{Q} =& \sum_{i=1}^{n} W_i \text{log} \left[\lambda(T_i)^{D_i} \exp\{-\Lambda(T_i)\}\right] \\ =& \sum_{i=1}^{n} W_i \text{log} \left[\lambda^0(T_i)^{D_i} \exp\{\beta X_i\}^{D_i} \exp\{-\Lambda^0(T_i) e^{\beta X_i}\}\right]\end{aligned}$$

This is the form of the log-likelihood for a weighted Cox regression model. In the literature, many authors have discussed profile-likelihood estimators for the baseline (cumulative) hazard in such a setting, and the resulting estimator resembles a weighted Breslow estimator. We derive its form below.

The function $\tilde{Q}$ will be maximized for $\lambda^0(t)$ taking value 0 when there is no event and takes non-zero values when there is an event. Suppose $t_1, \dots, t_K$ are the (unique) ordered event times T_i such that $D_i = 1$. Let R_k be the subjects at risk just before t_k and E_k be the subjects with events at t_k. Define λ_k^0 to be the value of $\lambda^0(t)$ at t_k. We can re-write $\tilde{Q}$ as

$$\begin{aligned}\tilde{Q} =& \sum_{i=1}^{n} W_i \text{log} \left[\lambda^0(T_i)^{D_i} \exp\{\beta X_i\}^{D_i}\right] - \sum_{i=1}^{n} W_i \exp\{\beta X_i\} \sum_{(t_k \leq T_i)} \lambda_k^0 \\ \propto& \sum_{k=1}^{K} \left[\log\left(\lambda_k^0\right) \sum_{i \in E_k}^{n} W_i\right] - \sum_{k=1}^{K} \left[\lambda_k^0 \sum_{i \in R_k} W_i \exp\{\beta X_i\}\right]\end{aligned}$$

If we take the first derivative and set it equal to zero, we have

$$\frac{\partial \tilde{Q}}{\partial \lambda_k^0} = \frac{\sum_{i \in E_k}^{n} W_i}{\lambda_k^0} - \sum_{i \in R_k} W_i \exp\{\beta X_i\} = 0 \implies \hat{\lambda}_k^0 = \frac{\sum_{i \in E_k}^{n} W_i}{\sum_{i \in R_k} W_i \exp\{\beta X_i\}}$$

Table E.1 shows the resulting baseline hazard estimates for each transition. We then define $\hat{\Lambda}^0(t) = \sum_{k: t_k \leq t} \hat{\lambda}_k^0$ for each transition.

Table E.1: EM Baseline Hazard Estimates (Unrestricted)

Transition	D_i	T_i	W_i	$\hat{\lambda}_k^0$
$1 \to 3$	δ_{ir}	Y_{ir}	p_i	$\frac{\sum_{i\in E_k}^n p_i}{\sum_{i\in R_k} p_i \exp\{\beta_{13}X_i\}}$
$2 \to 4$	$\delta_{id}(1-\delta_{ir})$	Y_{id}	$1-p_i$	$\frac{\sum_{i\in E_k}^n (1-p_i)}{\sum_{i\in R_k} (1-p_i) \exp\{\beta_{24}X_i\}}$
$1 \to 4$	$\delta_{id}(1-\delta_{ir})$	Y_{ir}	p_i	$\frac{\sum_{i\in E_k}^n p_i}{\sum_{i\in R_k} p_i \exp\{\beta_{14}X_i\}}$
$3 \to 4$	δ_{id}	$Y_{id}-Y_{ir}$	δ_{ir}	$\frac{\sum_{i\in E_k}^n \delta_{ir}}{\sum_{i\in R_k} \delta_{ir} \exp\{\beta_{34}X_i\}}$

We note that under the proposed estimator, $\hat{P}(T_r > t|X_i) = e^{-\hat{\Lambda}_{13}^0(t)\exp(\beta_{13}X_i)}$ will never be exactly zero for any t, although it will get close. Previous work studying baseline hazard estimators in the usual Cox proportional hazards cure model setting suggests that model-fitting properties (via EM) may be slightly improved when we use a product-limit type estimator for the baseline survival function directly (which can go exactly to zero) rather than a Breslow-type estimator for the hazard function (which can have estimated $\hat{P}(T_r > t|X_i)$ near zero but never exactly zero) (Sy and Taylor, 2000). In simulations (not shown), we did not see much impact of using a product-limit-type estimator rather than the proposed Breslow-type estimators, but future work could explore the impact of the different baseline hazard estimators.

$2 \to 4$ and $1 \to 4$ Baselines Assumed Equal or Proportional

Suppose we assume that $\lambda_{14}^0(t) = \lambda_{24}^0(t)$ for all $t \in (0, \tau]$, where τ is the last event time of any type observed. In this case, the estimators for $\lambda_{13}^0(t)$ and $\lambda_{34}^0(t)$ do not change, but we do modify the estimator for $\lambda_{14}^0(t) = \lambda_{24}^0(t)$. We can rewrite the expected log likelihood contribution of $\lambda_{14}^0(t)$ as:

$$\begin{aligned}&\sum_{i=1}^n p_i \log\left[\lambda_{14}(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}(Y_{ir})\}\right] \\ &+\sum_{i=1}^n (1-p_i)\log\left[\lambda_{24}(Y_{id})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{24}(Y_{id})\}\right]\end{aligned} \tag{E.1}$$

Using that $Y_{ir} = Y_{id}$ if $\delta_{ir} = 0$ and $C_{ir} = C_{id}$ and applying the equality assumption, we can rewrite (E.1) as

$$\begin{aligned}
&\sum_{i=1}^{n} p_i \log\left[\lambda_{14}^0(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}^0(Y_{ir})\exp(\beta_{14}X_i)\}\right] \\
&+ (1-p_i)\log\left[\lambda_{14}^0(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}^0(Y_{ir})\exp(\beta_{24}X_i)\}\right] + C
\end{aligned}$$

where C is a constant with respect to $\lambda_{14}^0(t)$. Again, this function is going to be maximized with respect to $\lambda_{14}^0(t)$ when the baseline hazard is nonzero only at event times. Suppose $t_1, \ldots, t_K$ are the (unique) ordered values of Y_{ir} such that $\delta_{id}(1-\delta_{ir}) = 1$ (death without recurrence). Let R_k be the subjects at risk at just before t_k and E_k be the subjects with events at t_k. Define λ_k^0 to be the value of $\lambda_{14}^0(t)$ at t_k.

$$\begin{aligned}
Q_{14} + Q_{24} =& \sum_{i=1}^{n} p_i \log\left[\lambda_{14}^0(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\exp(\beta_{14}X_i)\sum_{t_k \le Y_{ir}}\lambda_k^0\}\right] \\
&+ \sum_{i=1}^{n}(1-p_i)\log\left[\lambda_{14}^0(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\exp(\beta_{24}X_i)\sum_{t_k \le Y_{ir}}\lambda_k^0\}\right] + C \\
=& \sum_{k=1}^{K}\left[\log\left(\lambda_k^0\right)\sum_{i \in E_k} p_i\right] - \sum_{i=1}^{n}\left[p_i \exp\{\beta_{14}X_i\}\sum_{t_k \le Y_{ir}}\lambda_k^0\right] \\
&+ \sum_{k=1}^{K}\left[\log\left(\lambda_k^0\right)\sum_{i \in E_k}(1-p_i)\right] - \sum_{i=1}^{n}\left[(1-p_i)\exp\{\beta_{24}X_i\}\sum_{t_k \le Y_{ir}}\lambda_k^0\right] + C \\
=& \sum_{k=1}^{K}\left[\log\left(\lambda_k^0\right)\sum_{i \in E_k} p_i\right] - \sum_{k=1}^{K}\left[\lambda_k^0\sum_{i \in R_k} p_i \exp\{\beta_{14}X_i\}\right] \\
&+ \sum_{k=1}^{K}\left[\log\left(\lambda_k^0\right)\sum_{i \in E_k}(1-p_i)\right] - \sum_{k=1}^{K}\left[\lambda_k^0\sum_{i \in R_k}(1-p_i)\exp\{\beta_{24}X_i\}\right] + C
\end{aligned}$$

$$\frac{\partial Q_{14} + Q_{24}}{\partial \lambda_k^0} = \frac{\sum_{i \in E_k} p_i}{\lambda_k^0} - \sum_{i \in R_k} p_i \exp\{\beta_{14}X_i\} + \frac{\sum_{i \in E_k}(1-p_i)}{\lambda_k^0} - \sum_{i \in R_k}(1-p_i)\exp\{\beta_{24}X_i\} = 0$$

$$\begin{aligned}
\implies \lambda_k^0 &= \frac{\sum_{i \in E_k} p_i + \sum_{i \in E_k}(1-p_i)}{\sum_{i \in R_k} p_i \exp\{\beta_{14}X_i\} + \sum_{i \in R_k}(1-p_i)\exp\{\beta_{24}X_i\}} \\
&= \frac{\sum_{i \in E_k} 1}{\sum_{i \in R_k} p_i \exp\{\beta_{14}X_i\} + \sum_{i \in R_k}(1-p_i)\exp\{\beta_{24}X_i\}}
\end{aligned}$$

We therefore obtain the estimators in **Table E.2**

Table E.2: EM Baseline Hazard Estimates (Equal)

Assumption	$\hat{\lambda}_k^0$
$\beta_{24} \neq \beta_{14}$	$\frac{\sum_{i \in E_k} 1}{\sum_{i \in R_k} p_i \exp\{\beta_{14} X_i\} + \sum_{i \in R_k} (1-p_i) \exp\{\beta_{24} X_i\}}$
$\beta_{24} = \beta_{14}$	$\frac{\sum_{i \in E_k} 1}{\sum_{i \in R_k} \exp\{\beta_{14} X_i\}}$

When $\beta_{14} = \beta_{24}$, the estimator for the baseline hazard is just a traditional Breslow estimator with event/censoring time Y_{ir} and event indicator $\delta_{id}(1 - \delta_{ir})$ using parameter β_{14}. We note that if the β's are equal, the multistate cure model would reduce to a CPH cure model with two additional Cox regressions for death before and after recurrence.

Suppose instead that we want to assume proportional baseline hazards, where $\lambda_{14}^0(t) = \lambda_{24}^0(t) \exp\{\beta_0\}$ for all $t \in (0, \tau]$. We can model:

$$
\begin{aligned}
\lambda_{24}(t) &= \lambda_{24}^0(t) \exp\{\beta_{24}^T X_i\} \\
\lambda_{14}(t) &= \lambda_{14}^0(t) \exp\{\beta_{14}^T X_i\} = \lambda_{24}^0(t) \exp\{\beta_0 + \beta_{14}^T X_i\}
\end{aligned}
$$

This situation is easily handled by including an intercept in the covariate set for the model for $\lambda_{14}(t)$ and then assuming that the resulting baseline hazards are equal. The resulting estimates for the baseline hazard jumps λ_k^0 for $\lambda_{24}^0(t)$ at event times $t_1, \ldots, t_K$ are expressed in **Table E.3**.

Table E.3: EM Baseline Hazard Estimates (Proportional)

Assumption	$\hat{\lambda}_k^0$
$\beta_{24} \neq \beta_{14}$ or $\beta_{24} = \beta_{14}$	$\frac{\sum_{i \in E_k} 1}{\sum_{i \in R_k} p_i \exp\{\beta_0 + \beta_{14} X_i\} + \sum_{i \in R_k} (1-p_i) \exp\{\beta_{24} X_i\}}$

Appendix F

Estimation of Baseline Hazards for MCEM Algorithm

In this section, we present estimators of the nonparametric baseline hazards for the Monte Carlo EM Algorithm. The proposed estimators are different from those used in the conventional EM algorithm. Let $l(\theta|\mathbb{D})$ be the complete data log-likelihood with complete data $\mathbb{D}$. In iteration t of the Monte Carlo EM Algorithm, we obtain M imputed values of $\mathbb{D}$, $\mathbb{D}^{(t,1)}, \mathbb{D}^{(t,2)}, \ldots, \mathbb{D}^{(t,M)}$, and an estimate of the parameter θ. At the t^{th} iteration, updated estimates of the baseline hazard can be obtained by maximizing $\frac{1}{M}\sum_{m=1}^{M} l(\theta|\mathbb{D}^{(t,m)})$ with respect to the baseline hazards we want to estimate, treating the imputed data and the parameter estimate as fixed.

This time, we require estimates of all four baseline hazards. We note that the proposed imputation-based algorithm for handling unequal follow-up has poor performance when we use baseline hazard estimators that are nonzero only at event times. Instead, we will restrict our estimators of the baseline hazards to be step functions that change value at the observed event times for the event corresponding to the baseline hazard of interest. For λ_{14}^{0}, for example, this would be observed death times without prior recurrence.

First, we will use the profile likelihood method to derive the form of the estimators assuming $\mathbb{D}$ is known. Then, we will generalize this estimators for use in the Monte Carlo EM Algorithm.

For these derivations, we will assume that we use the same set of predictors, X_i in the model for each of the transitions. However, these estimators are easily generalized to allow the covariate sets to differ across transitions.

$2 \rightarrow 4$ and $1 \rightarrow 4$ Baselines Unrestricted

Assuming $\mathbb{D}$ is known

We would like to maximize $l(\theta|\mathbb{D})$ with respect to the baseline hazard functions. We assume θ is fixed at the estimated value from the previous M-Step. Suppose we do not assume any relationship between $\lambda_{14}^0(t)$ and $\lambda_{24}^0(t)$. We consider the contributions of each of the baseline hazards to the log-likelihood. For each hazard, the contribution is:

$$\begin{aligned}
\lambda_{14}^0(t): &\qquad \sum_{i=1}^{n} G_i \text{log}\left[\lambda_{14}(Y_{ir})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{14}(Y_{ir})\}\right] \\
\lambda_{24}^0(t): &\qquad \sum_{i=1}^{n} (1-G_i) \text{log}\left[\lambda_{24}(Y_{id})^{\delta_{id}(1-\delta_{ir})}\exp\{-\Lambda_{24}(Y_{id})\}\right] \\
\lambda_{13}^0(t): &\qquad \sum_{i=1}^{n} G_i \text{log}\left[\lambda_{13}(Y_{ir})^{\delta_{ir}}\exp\{-\Lambda_{13}(Y_{ir})\}\right] \\
\lambda_{34}^0(t): &\qquad \sum_{i=1}^{n} \delta_{ir} \text{log}\left[\lambda_{34}(Y_{id}-Y_{ir})^{\delta_{id}}\exp\{-\Lambda_{34}(Y_{id}-Y_{ir})\}\right]
\end{aligned}$$

In each case, we can construct censored outcome T and event indicator D to represent the outcome of interest with a corresponding weight term W. Let $t_1, \ldots, t_K$ be the ordered event times. Let R_k be the subjects at risk just before t_k and E_k be the subjects with events at t_k. We want to maximize the log-likelihood:

$$l(\lambda^0(t)) = \sum_{i=1}^{N} W_i D_i log(\lambda_0(T_i)) - W_i \exp\{\beta X_i\} \sum_{t \leq T_i} \lambda_0(T_i) + Constant$$

Suppose that the baseline hazard takes the form of a step function, so within each interval $[t_j, t_{j+1})$, the baseline hazard is constant. We also restrict the hazard to be zero for all $t > t_K$. We define $t_0 = 0$ and t_{K+1} to be just after the maximum time on study (so that the maximum time is contained in $[t_K, t_{K+1})$). Let λ_j^0 represent the baseline hazard in the interval $[t_j, t_{j+1})$. We define $\lambda_0^0 = 0$. For individual i, let L_i represent the value of j such that T_i is contained in $[t_j, t_{j+1})$. We can rewrite the log-likelihood as

$$l(\lambda^0(t)) = \sum_{i=1}^{N} W_i D_i log[\lambda_{L_i}^0] - W_i \exp\{\beta X_i\} \left[\sum_{j=0}^{L_i - 1} \lambda_j^0 (t_{j+1} - t_j) + \lambda_{L_i}^0 (T_i - t_{L_i})\right]$$

$$
\begin{aligned}
\frac{\partial l}{\partial \lambda_k^0} &= \frac{1}{\lambda_k^0}\sum_{i=1}^{N}\mathbb{I}(L_i = k)W_i D_i - \sum_{i=1}^{N} W_i \mathbb{I}(L_i > k)\exp\{\beta X_i\}(t_{k+1} - t_k) \\
&- \sum_{i=1}^{N} W_i \mathbb{I}(L_i = k)\exp\{\beta X_i\}(T_i - t_k) \\
&= \frac{1}{\lambda_k^0}\sum_{i \in E_k} W_i - \sum_{i=1}^{N} W_i \mathbb{I}(L_i > k)\exp\{\beta X_i\}(t_{k+1} - t_k) \\
&- \sum_{i=1}^{N} W_i \mathbb{I}(L_i = k)\exp\{\beta X_i\}(T_i - t_k) \\
&\implies \hat{\lambda}_k^0 = \frac{\sum_{i \in E_k} W_i}{\sum_{i \in R_k} W_i \exp\{\beta X_i\}(\min(T_i, t_{k+1}) - t_k)} \qquad \text{(F.1)} \\
&\implies \hat{\Lambda}^0(t) = \sum_{j: t > t_j} \hat{\lambda}_j^0(\min(t_{j+1}, t) - t_j)
\end{aligned}
$$

where R_k is the set of people at risk at just before t_k and E_k is the group of subjects with events in $[t_k, t_{k+1})$. This is the maximizer of $l(\theta|\mathbb{D})$. We note that a subject having an event at t_k contributes nothing to the denominator in (F.1). **Table F.1** shows the baseline hazard estimator for each transition.

Table F.1: MCEM Baseline Hazard Estimates (Unrestricted)

Transition	D_i	T_i	W_i	$\hat{\lambda}_k^0$
$1 \to 3$	δ_{ir}	Y_{ir}	G_i	$\frac{\sum_{i \in E_k}^{n} G_i}{\sum_{i \in R_k} G_i \exp\{\beta_{13} X_i\}(\min(Y_{ir}, t_{k+1}) - t_k)}$
$2 \to 4$	$\delta_{id}(1 - \delta_{ir})$	Y_{id}	$1 - G_i$	$\frac{\sum_{i \in E_k}^{n} (1 - G_i)}{\sum_{i \in R_k} (1 - G_i) \exp\{\beta_{24} X_i\}(\min(Y_{id}, t_{k+1}) - t_k)}$
$1 \to 4$	$\delta_{id}(1 - \delta_{ir})$	Y_{ir}	G_i	$\frac{\sum_{i \in E_k}^{n} G_i}{\sum_{i \in R_k} G_i \exp\{\beta_{14} X_i\}(\min(Y_{ir}, t_{k+1}) - t_k)}$
$3 \to 4$	δ_{id}	Y_{id}	δ_{ir}	$\frac{\sum_{i \in E_k}^{n} \delta_{ir}}{\sum_{i \in R_k} \delta_{ir} \exp\{\beta_{34} X_i\}(\min(Y_{id}, t_{k+1}) - t_k)}$

We note that by this definition, we could have that the baseline hazard for the 1 → 3 transition does not go exactly to zero at the last event time. If there is truly a cure structure to the data, however, we would like our estimator to go to zero at some point. Therefore, we suggest defining $\hat{\lambda}_k^0$ to be equal to zero at the last event time. This restriction is equivalent to assuming that all subjects at risk for recurrence after the last observed recurrence event are cured. A similar restriction is often made when fitting Cox proportional hazards mixture cure models, and it is associated with improved performance in that setting (Sy and Taylor, 2000).

Additionally, we note that the estimator for the 3 → 4 transition will be exactly zero if the subject with longest follow-up for this transition has an event. Unlike for the 1 → 3 transition, we would like our estimator to allow for events after the last observed event time. This subtle issue makes a difference for our imputation approach for unequal follow (see **Appendix G** for details). Therefore, we will define $t_1, \dots, t_K$ for this transition such that t_K is the next to last observed event (rather than the last observed event). This will result in a nonzero value for the hazard rate at t_K.

Using imputed $\mathbb{D}$

The estimator in (F.1) is the maximizer of $l(\theta|\mathbb{D})$ for some fully-observed dataset $\mathbb{D}$. In the Monte Carlo EM Algorithm, we impute several (M) versions of the dataset $\mathbb{D}$. Suppose we create a stacked version of the dataset, $\mathbb{D}^{(t)}$, by stacking the imputed versions of the data at iteration t, $\mathbb{D}^{(t,1)}, \mathbb{D}^{(t,2)}, \dots, \mathbb{D}^{(t,M)}$. Then we can maximize $\frac{1}{M}\sum_{m=1}^{M} l(\theta|\mathbb{D}^{(t,m)})$ by instead maximizing $l(\theta|\mathbb{D}^{(t)})$. The resulting estimators for the baseline hazards take the same form as in (F.1) except T, D, W, E_k, and R_k are defined and indexed by the elements of the stacked dataset $\mathbb{D}^{(t)}$ (so individual subjects enter the estimator M times). It is worth noting that this estimator may result in very large estimates for the step heights when the event times are very close together. In this case, we may choose to set cutpoints $t_1, \dots, t_K$ to be a subset of the event times.

2 → 4 and 1 → 4 Baselines Assumed Equal or Proportional

Assuming $\mathbb{D}$ is known

We assume that $\lambda_{14}^0(t) = \lambda_{24}^0(t)$ for all $t \in (0, \tau]$, where τ is the last event time of any type observed. In this case, the estimators for $\lambda_{13}^0(t)$ and $\lambda_{34}^0(t)$ do not change, but we do modify the estimator for $\lambda_{14}^0(t) = \lambda_{24}^0(t)$. We can rewrite the log likelihood contribution of as $\lambda_{14}^0(t)$ as:

$$\sum_{i=1}^{n} G_i \log\left[\lambda_{14}^0(Y_{ir})^{\delta_{id}(1-\delta_{ir})} \exp\{-\Lambda_{14}^0(Y_{ir})\exp(\beta_{14}X_i)\}\right]$$
$$+ (1-G_i)\log\left[\lambda_{14}^0(Y_{ir})^{\delta_{id}(1-\delta_{ir})} \exp\{-\Lambda_{14}^0(Y_{ir})\exp(\beta_{24}X_i)\}\right] + C$$

where C is a constant. Suppose $t_1, \ldots, t_K$ are the (unique) ordered values of Y_{ir} such that $\delta_{id}(1-\delta_{ir}) = 1$ (death without recurrence) and R_k be the subjects at risk just before t_k and E_k be the subjects with events at t_k.

As before, suppose that the baseline hazard takes the form of a step function, so within each interval $[t_j, t_{j+1})$, the baseline hazard is constant. We also restrict the hazard to be zero for all $t > t_K$. We define $t_0 = 0$ and t_{K+1} to be just after the maximum time on study (so that the maximum time is contained in $[t_K, t_{K+1})$). Let λ_j^0 represent the baseline hazard in the interval $[t_j, t_{j+1})$. We define $\lambda_0^0 = 0$. For individual i, let L_i represent the value of j such that T_i is contained in $[t_j, t_{j+1})$. We can rewrite the log-likelihood as

$$\begin{aligned} l(\lambda^0(t)) = \sum_{i=1}^{N} \delta_{id}(1-\delta_{ir}) log[\lambda_{L_i}^0] - G_i \exp\{\beta_{14}X_i\} \left[\sum_{j=0}^{L_i-1} \lambda_j^0 (t_{j+1} - t_j) + \lambda_{L_i}^0 (T_i - t_{L_i}) \right] \\ - (1-G_i)\exp\{\beta_{24}X_i\} \left[\sum_{j=0}^{L_i-1} \lambda_j^0 (t_{j+1} - t_j) + \lambda_{L_i}^0 (T_i - t_{L_i}) \right] \end{aligned}$$

$$\begin{aligned} \frac{\partial l}{\partial \lambda_k^0} &= \frac{1}{\lambda_k^0} \sum_{i=1}^{N} \mathbb{I}(L_i = k)\delta_{id}(1-\delta_{ir}) - \sum_{i=1}^{N} G_i \mathbb{I}(L_i > k) \exp\{\beta_{14}X_i\}(t_{k+1} - t_k) \\ &- \sum_{i=1}^{N} G_i \mathbb{I}(L_i = k) \exp\{\beta_{14}X_i\}(T_i - t_k) - \sum_{i=1}^{N} (1-G_i)\mathbb{I}(L_i > k) \exp\{\beta_{24}X_i\}(t_{k+1} - t_k) \\ &- \sum_{i=1}^{N} (1-G_i)\mathbb{I}(L_i = k) \exp\{\beta_{24}X_i\}(T_i - t_k) \\ &= \frac{1}{\lambda_k^0} \sum_{i \in E_k} 1 - \sum_{i=1}^{N} G_i \mathbb{I}(L_i > k) \exp\{\beta_{14}X_i\}(t_{k+1} - t_k) \\ &- \sum_{i=1}^{N} G_i \mathbb{I}(L_i = k) \exp\{\beta_{14}X_i\}(T_i - t_k) \\ &- \sum_{i=1}^{N} (1-G_i)\mathbb{I}(L_i > k) \exp\{\beta_{24}X_i\}(t_{k+1} - t_k) \\ &- \sum_{i=1}^{N} (1-G_i)\mathbb{I}(L_i = k) \exp\{\beta_{24}X_i\}(T_i - t_k) \end{aligned}$$

$$\begin{aligned} \hat{\lambda}_k^0 &= \frac{\sum_{i \in E_k} 1}{\sum_{i \in R_k} G_i e^{\beta_{14}X_i}(\min(T_i, t_{k+1}) - t_k) + \sum_{i \in R_k} (1-G_i) e^{\beta_{24}X_i}(\min(T_i, t_{k+1}) - t_k)} \\ \hat{\Lambda}^0(t) &= \sum_{j: t > t_j} \hat{\lambda}_j^0 (\min(t_{j+1}, t) - t_j) \end{aligned}$$

Suppose instead we want to assume the baseline hazards are proportional with $\lambda_{14}^0(t) = \lambda_{24}^0(t)e^{\beta_0}$ for all $t \in (0, \tau]$. We can use the same trick as in the conventional EM algorithm estimators to transform the proportional baseline hazards situation into the equal baseline hazards situation (by adding an intercept to the model for the $1 \rightarrow 4$ transition). Then, we can use the estimator for equal baseline hazards above with $\beta_{14}X_i$ replaced by $\beta_0 + \beta_{14}X_i$.

Using imputed $\mathbb{D}$

As before, we can obtain the Monte Carlo EM estimate of the baseline hazard by creating an stacked version of the dataset (created by stacking $\mathbb{D}^{(t,1)}, \mathbb{D}^{(t,2)}, \ldots, \mathbb{D}^{(t,M)}$) and applying the above estimator to the stacked version of the data.

Appendix G

Derivation of Imputation Approach for Handling Unequal Follow-up

G.1 General Approach

In this section, we derive the imputation approach used to handle unequal follow-up in the outcomes. As before, we let T_{ir} and T_{id} be the underlying recurrence and death times for subject i. Let C_{ir} be the censoring time for recurrence and C_{id} be the censoring time for death, but this time we assume that $C_r \leq C_d$ with $C_{ir} < C_{id}$ for at least some subjects. For all subjects, we observe $Y_{ir}^0 = \min(T_{ir}, C_{ir}, T_{id})$, $\delta_{ir}^0 = \mathbb{I}(Y_{ir} = T_{ir})$, $Y_{id} = \min(T_{id}, C_{id})$ and $\delta_{id} = \mathbb{I}(Y_{id} = T_{id})$.

When recurrence is sometimes censored before death, we can run into the issue where recurrence status is unknown for part of the follow-up time for death. In this case, we do not know to which transition we should attribute the time at risk for death after censoring of recurrence. This setting is similar to issues of interval censoring and panel data for standard illness-death models (Jackson, 2011). One difference between our setting and usual panel data is that with panel data, subject's states are known only at discrete time points. However, for subjects with unequal censoring in our setting, we know that the subject is in the death state at Y_{id} if they have an observed death at Y_{id}, but if they are censored for death at Y_{id}, we do not know whether they are in State 1, State 2, or State 3 at the time Y_{id}.

In Conlon et al. (2013), unequal censoring is handled by directly incorporating what was observed for each subject into the likelihood (conditional on cure status). For example, for non-cured subjects with unequal censoring that had observed deaths at Y_{id}, the likelihood contribution of the outcome data would be P(in state 4 at $Y_{id}|$ state 1 at Y_{ir}^0

and non-cured). For non-cured subjects with unequal censoring that had censored deaths at Y_{id}, the likelihood contribution of the outcome data would be P(in state 1 at $t|$ state 1 at u and non-cured) + P(in state 3 at $t|$ state 1 at u and non-cured). This leads to the following likelihood contribution for subjects with unequal censoring (Conlon et al., 2013):

$$\left[P(G_i = 0)S_2(Y_{id})\lambda_{24}(Y_{id})^{\delta_{id}}\right]^{1-G_i} \times$$
$$\left[P(G_i = 1)S_1(Y_{id})\lambda_{14}(Y_{id})^{\delta_{id}} + P(G_i = 1)\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(r)S_1(r)\lambda_{34}(Y_{id} - r)^{\delta_{id}} S_3(Y_{id} - r)dr\right]^{G_i}$$

We note that this likelihood contribution involves an integral for non-cured subjects. Rather than attempting to maximize a likelihood involving an integral in the M-Step of the MCEM algorithm, we instead propose the following imputation strategy to handle the unequal censoring.

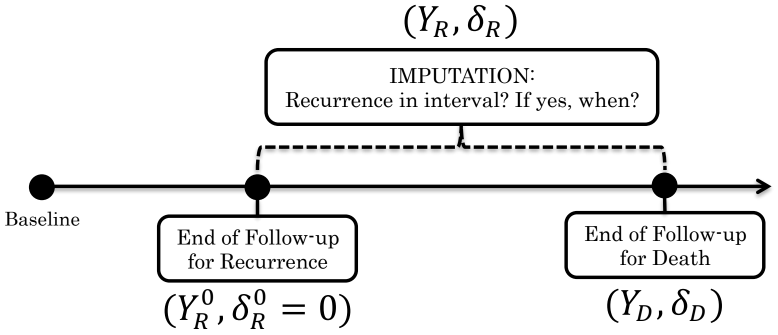
Suppose we had followed all subjects for recurrence as long as we followed them for death. Define $Y_{ir} = \min(T_{ir}, C_{id}, T_{id})$ and $\delta_{ir} = \mathbb{I}(Y_{ir} = T_{ir})$ to be the resulting values for recurrence event/censoring time and event indicator under the later censoring time. These versions of the outcomes do not suffer from the same problem regarding unknown recurrence status. Define $P_i = \mathbb{I}(Y_{ir}^0 < Y_{id} \text{ and } \delta_{ir}^0 = 0)$. For subjects with $P_i = 0$, $Y_{ir} = Y_{ir}^0$ and $\delta_{ir} = \delta_{ir}^0$. For subjects with $P_i = 1$, however, Y_{ir} is only known to be greater than Y_{ir}^0. Our goal is to impute values of Y_{ir} and δ_{ir} for subjects with $P_i = 1$. We will perform this imputation within the Monte Carlo EM Algorithm in which we also impute values for G.

Suppose that we have already imputed $G_i = 0$ (cured) for subject i. By imputing $G_i = 0$, we claim that this subject will never experience a recurrence. Therefore, we can automatically set $\delta_{ir} = 0$ and $Y_{ir} = Y_{id}$.

Suppose, however, that we imputed $G_i = 1$ (non-cured). By imputing $G_i = 1$, we claim that this subject will eventually experience a recurrence, and therefore we can ignore the contribution of the 2 → 4 part of the multistate model and focus instead on the semi-competing risks model for recurrence and death in the non-cured subjects. We can impute (Y_{ir}, δ_{ir}) jointly from their posterior predictive distribution, $f(Y_{ir}, \delta_{ir}|X_i, \delta_{id}, Y_{id}, Y_{ir}^0, \delta_{ir}^0, G_i = 1, P_i = 1)$ (Little and Rubin, 2002). In practice, however, we can obtain a draw from the predictive distribution by drawing from

$f(Y_{ir}, \delta_{ir}|X_i, \delta_{id}, Y_{id}, Y_{ir}^0, \delta_{ir}^0, G_i = 1, P_i = 1; \theta^{(t)})$ using the most recent estimate of θ. However, drawing from the joint distribution of Y_{ir} and δ_{ir} parameterized by $\theta^{(t)}$ may still be difficult. Instead, we propose to first impute δ_{ir} from $f(\delta_{ir}|X_i, \delta_{id}, Y_{id}, Y_{ir}^0, \delta_{ir}^0, G_i = 1, P_i = 1; \theta^{(t)})$ and then impute Y_{ir} from $f(Y_{ir}|X_i, \delta_{id}, Y_{id}, Y_{ir}^0, \delta_{ir}^0, G_i = 1, P_i = 1, \delta_{ir}; \theta^{(t)})$. This approach is equivalent to first imputing whether a recurrence occurred in the time between Y_{ir}^0 and Y_{id} (value of δ_{ir}) and, if so, when the recurrence occurred. If not, then $Y_{ir} = Y_{id}$. **Figure G.1** provides a visualization of the imputation approach.

Figure G.1: Diagram of Unequal Follow-up Scenario


(Y_R, δ_R)
IMPUTATION:
Recurrence in interval? If yes, when?
Baseline
End of Follow-up for Recurrence
$(Y_R^0, \delta_R^0 = 0)$
End of Follow-up for Death
(Y_D, δ_D)

Step 1: Imputation of δ_{ir}

First, we note that $P_i = 1$ implies $T_{ir} > Y_{ir}^0$. We can draw δ_{ir} with

$$\begin{aligned} & P(\delta_{ir} = 1|X_i, \delta_{id}, Y_{id}, Y_{ir}^0, \delta_{ir}^0, G_i = 1, P_i = 1; \theta^{(t)}) \\ & = P(\delta_{ir} = 1|X_i, \delta_{id}, Y_{id}, Y_{ir}^0, G_i = 1, T_{ir} > Y_{ir}^0; \theta^{(t)}) \\ & = \frac{\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(t) S_1(t) S_3(Y_{id} - t) \lambda_{34}(Y_{id} - t)^{\delta_{id}} dt}{\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(t) S_1(t) S_3(Y_{id} - t) \lambda_{34}(Y_{id} - t)^{\delta_{id}} dt + \lambda_{14}^{\delta_{id}}(Y_{id}) S_1(Y_{id})} \end{aligned} \tag{G.1}$$

For parametric baseline hazards, we can substitute the parameter estimates from the M Step into the (G.1) in order to impute δ_{ir}. Suppose, however, that we want to use nonparametric baseline hazards. Weighted Breslow-type estimators of the baseline hazard function take nonzero values only at event times, which will create a problem when imputing using (G.1). Instead, we will require our baseline hazard estimators to be step functions as discussed in **Appendix F**.

Step 2: Imputation of Y_{ir}

If the imputed value of $\delta_{ir} = 0$, then automatically set $Y_{ir} = Y_{id}$. Otherwise, we know that $Y_{ir} = T_{ir} < Y_{id}$. We note that $\{Y_{id} > Y^0_{ir}, \delta^0_{ir} = 0, \delta_{ir} = 1, G_i = 1\}$ implies $\{Y^0_{ir} < T_{ir} < Y_{id}\}$. Therefore, we can draw T_{ir} from:

$$\begin{aligned} &f(T_{ir} = t | X_i, \delta_{id}, Y_{id}, Y^0_{ir}, Y^0_{ir} < T_{ir} < Y_{id}; \theta^{(t)}) \\ &\propto \lambda_{13}(t) S_1(t) S_3(Y_{id} - t) \lambda_{34}(Y_{id} - t)^{\delta_{id}} \mathbb{I}(Y^0_{ir} < t < Y_{id}) \end{aligned} \tag{G.2}$$

We note that we actually know the full form of this distribution since we calculate the proportionality constant for (G.2) in Step 1 (integral in expression (G.1)). Several options exist to draw T_{ir} from distribution (G.2), and we explore several approaches under Weibull and nonparametric assumptions for the baseline hazards.

We note that, under nonparametric baseline hazards, the estimator for $\lambda_{13}(t)$ will be exactly zero for t greater than the last observed recurrence time by construction. Therefore, the distribution in (G.2) will be zero for t greater than the last observed recurrence time. Additionally, in **Appendix F**, we restrict the hazard for the 3 → 4 transition to be nonzero at the last event time. This avoids the possibility that the distribution in (G.2) evaluated at the proposed baseline hazard estimates will be zero across all t for a particular subject.

G.2 Implementation of T_r Draw

In this section, we propose various methods to accomplish the imputation in Step 2. One proposed method is rejection sampling, which can be applied under parametric or nonparametric baseline hazard assumptions. In our experience, this method has good performance when that baseline hazards are parametric, but may have poor performance in some nonparametric baseline settings. We therefore propose three additional methods that tend to have better performance under nonparametric baseline assumptions.

Method 1: Rejection Sampling

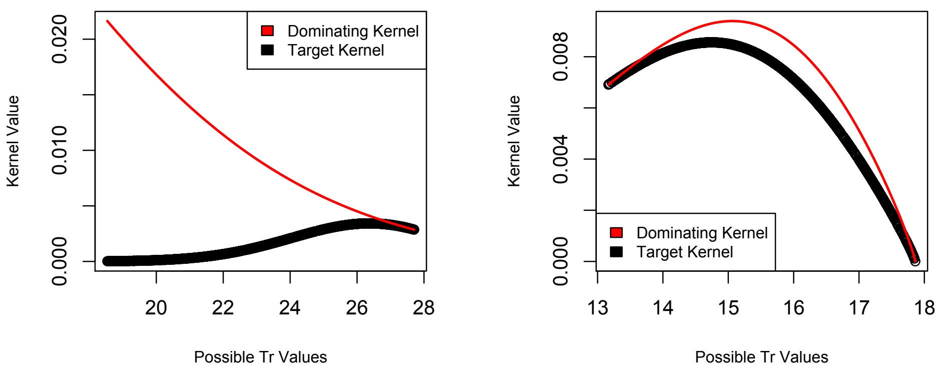
We can draw from (G.2) using a rejection sampling algorithm assuming Weibull baseline hazards. Rejection sampling algorithms involve determining a simple density, $g(t)$, that dominates the kernel of interest, $k(t)$, such that we can write $k(t) \leq Kg(t) \ \forall \ t, K \geq 1$. We:

1) generate T from $g(t)$ and U from $U(0,1)$

2) Accept draw T if $U \leq \frac{k(T)}{Kg(T)}$. Otherwise, reject draw T and return to 1) (Robert and Casella, 2004).

We define $k(t) = \lambda_{13}(t)S_1(t)S_3(Y_{id} - t)\lambda_{34}(Y_{id} - t)^{\delta_{id}}\mathbb{I}(Y_{ir}^0 < t < Y_{id})$, which is equal to the kernel in (G.2). We can obtain a draw of T_{ir} from (G.2) using one of the following two rejection sampling algorithms. Option 1 below can be applied to impute T_{ir} for $\delta_{id} = 0$ or $\delta_{id} = 1$, but it may not be very efficient when $\delta_{id} = 1$ and $\lambda_{34}(t)$ has a large range over (Y_{ir}^0, Y_{id}). We therefore propose a second approach, Option 2, for drawing from (G.2) for subjects with $\delta_{id} = 1$. **Figure G.2** shows examples of the target and dominating kernel for each approach.

Figure G.2: Example of Target and Dominating Kernels in Rejection Sampling


(a) Kernels from Option 1 (with $\delta_{id} = 0$)

(b) Kernels from Option 2 (with $\delta_{id} = 1$)

Option 1:

Let $g(t) \propto \lambda_{13}(t)e^{-\Lambda_{13}(t)}\mathbb{I}(Y_{ir}^0 < t < Y_{id})$. This is a truncated failure time distribution.

A draw from $g(t)$ can be obtained much more easily than a draw from $k(t)$. We can obtain a draw from $g(t)$ using

$$P(T > t|Y_{ir}^0 < T < Y_{id}; \theta^{(t)}) = \frac{e^{-\Lambda_{13}(t)} - e^{-\Lambda_{13}(Y_{id})}}{e^{-\Lambda_{13}(Y_{ir}^0)} - e^{-\Lambda_{13}(Y_{id})}}\mathbb{I}(Y_{ir}^0 < t < Y_{id}) \sim U(0,1)$$

Draw $M \sim U(0,1)$. Then we can obtain a draw for T by solving

$$\frac{e^{-\Lambda_{13}(T)} - e^{-\Lambda_{13}(Y_{id})}}{e^{-\Lambda_{13}(Y_{ir}^0)} - e^{-\Lambda_{13}(Y_{id})}} = M \implies T = \Lambda_{13}^{-1}\left(-\log\left[M\left\{e^{-\Lambda_{13}(Y_{ir}^0)} - e^{-\Lambda_{13}(Y_{id})}\right\} + e^{-\Lambda_{13}(Y_{id})}\right]\right)$$

Let T be a draw from $g(t)$. We accept the draw $T_{ir} = T$ if $U \leq \frac{1}{K}e^{-\Lambda_{14}(T)}S_3(Y_{id} - T)\lambda_{34}(Y_{id} - T)^{\delta_{id}}$ where constant $K = \max_{Y_{ir}^0 < t < Y_{id}} e^{-\Lambda_{14}(t)}S_3(Y_{id} - t)\lambda_{34}(Y_{id} - t)^{\delta_{id}}$. We continue drawing T until the corresponding inequality is satisfied.

Option 2:

This option is only appropriate for imputation if $\delta_{id} = 1$. Let $g(t) \propto \lambda_{34}(Y_{id} - t)e^{-\Lambda_{34}(Y_{id}-t)}\mathbb{I}(Y_{ir}^0 < t < Y_{id})$. This is again a truncated failure time distribution.

Again, a draw from $g(t)$ can be obtained much more easily than a draw from $k(t)$. Define $S = Y_{id} - T$. S has a truncated survival distribution $f_{34}(t)$. We can obtain a draw from $g(t)$ using

$$P(T > t|Y_{ir}^0 < T < Y_{id}; \theta^{(t)}) = P(Y_{id} - S > t|Y_{ir}^0 < Y_{id} - S < Y_{id})$$

$$= P(Y_{id} - t > S | 0 < S < Y_{id} - Y_{ir}^0) = \frac{P(0 < S < Y_{id} - t)}{P(0 < S < Y_{id} - Y_{ir}^0)} \mathbb{I}(Y_{ir}^0 < t < Y_{id})$$
$$= \frac{1 - e^{-\Lambda_{34}(Y_{id} - t)}}{1 - e^{-\Lambda_{34}(Y_{id} - Y_{ir}^0)}} \mathbb{I}(Y_{ir}^0 < t < Y_{id}) \sim U(0,1)$$

Draw $M \sim U(0,1)$. Then we can obtain a draw for T by solving

$$\frac{1 - e^{-\Lambda_{34}(Y_{id} - T)}}{1 - e^{-\Lambda_{34}(Y_{id} - Y_{ir}^0)}} = M \implies T = Y_{id} - \Lambda_{34}^{-1} \left(-\log \left[1 - M \left\{ 1 - e^{-\Lambda_{34}(Y_{id} - Y_{ir}^0)} \right\} \right] \right)$$

Let T be a draw from $g(t)$. We accept the draw $T_{ir} = T$ if $U \leq \frac{1}{K} \lambda_{13}(T) S_1(T)$ where constant $K = \max\limits_{Y_{ir}^0 < t < Y_{id}} \lambda_{13}(t) S_1(t)$. We continue drawing T until the corresponding inequality is satisfied. We note that this approach should not be used if T_{ir} is part of the covariate set for the 3 → 4 transition as drawing from $g(t)$ may be difficult in this case (since the covariate set also would depend on t).

Method 2: Metropolis-Hastings

Like the rejection sampling algorithm, the goal of the Metropolis-Hastings algorithm is to obtain a draw of variable V from a distribution known only up to proportionality, $k(v)$. The strategy is to first specify a proposal distribution, $p(v|u)$, from which we propose new values for the variable $V = v$ given the most recent drawn value of V, u. We can obtain a draw V from $k(v)$ by performing the following:

1) Generate v^* from $p(v|u)$. Generate $U \sim U(0,1)$

2) Define acceptance probability $\alpha = \min \left(1, \frac{p(u|v^*) k(v^*)}{p(v^*|u) k(u)} \right)$. Accept draw $V = v^*$ if $U \leq \alpha$.

Otherwise, we reject draw $V = v^*$ and keep $V = u$ (Robert and Casella, 2004). Under parametric baseline hazards, we suggest using a Uniform(Y_{ir}^0, Y_{id}) proposal distribution. However, this proposal distribution can run into problems under nonparametric baseline hazards when $Y_{id} > \tau_R$, the last observed recurrence time in the dataset. In this case, the $k(v)$ evaluated at the baseline hazard estimators could be zero for some values of t in (Y_{ir}^0, Y_{id}). To avoid this issue, we suggest using a Uniform(Y_{ir}^0, min(Y_{id}, τ_R)) proposal distribution under nonparametric baseline hazards.

Method 3: Inversion

We can use the fact that we have already calculated the proportionality constant for the distribution in (G.2) when imputing G to help draw from (G.2). Let $\tilde{S}(t)$ be the survival function corresponding to the distribution in (G.2). Then we can obtain a draw from (G.2) by solving the following equation for T:

$$\tilde{S}(T) = \frac{\int_{T}^{Y_{id}} \lambda_{13}(t) S_1(t) S_3(Y_{id} - t) \lambda_{34}(Y_{id} - t)^{\delta_{id}} dt}{\int_{Y_{ir}^0}^{Y_{id}} \lambda_{13}(t) S_1(t) S_3(Y_{id} - t) \lambda_{34}(Y_{id} - t)^{\delta_{id}} dt} = U$$

where U is drawn from $U(0,1)$. This equation can be solved using a root solver such as *uniroot* in R. In the case of Cox baseline hazards, this approach to drawing from (G.2) may be faster than the rejection-sampling methods. This approach is convenient because it is fairly simple, but it does require many numerical calculations of an integral within the root solver.

Method 4: Nested Parametric Model

This fourth method is applicable only when we have nonparametric baseline hazard assumptions. In this approach, we make parametric assumptions for the baseline hazards only for performing the imputation to deal with the unequal follow-up. For all other imputations and estimation, we assume nonparametric baselines.

We first obtain parameter estimates from a parametric multistate model fit to the data using the imputed outcome data from the previous iteration. Using these parameter estimates, we perform imputation of Y_{ir} and δ_{ir} under parametric assumptions for the baseline hazards.

Some Additional Comments

The proposed imputation-based approach for dealing with unequal follow-up can perform well when some subjects have longer follow-up for recurrence such that each interval (Y_{ir}^0, Y_{id}) we are imputing over contains follow-up for recurrence for at least a few subjects. Suppose no subjects have follow-up time for recurrence in the interval (Y_{ir}^0, Y_{id}). When we make parametric assumptions about the baseline hazard, the imputed outcome information will be entirely dependent on the parametric assumptions. When we have no subjects with later follow-up for recurrence and we have non-parametric baseline hazards, we might recommend the nested parametric model method for imputation, as the other methods may have poor performance. In our simulations, we supposed that some subset of the subjects had long follow-up for recurrence, allowing us to estimate the baseline hazard for recurrence over each interval (Y_{ir}^0, Y_{id}).

Suppose we want to use nonparametric baseline hazards in a multistate cure model fit to data with unequal follow-up and all subjects have early censoring for recurrence. In this case, several options exist. We could assume parametric baseline hazards just for the unequal follow-up imputation step (Method 4 above). However, this approach would be entirely dependent on the parametric assumptions. Another approach would be to censor the death data back to the follow-up time for recurrence for subjects with $P_i = 1$ and then fit the multistate cure model to the data with the modified death information. This may *substantially* decrease the death information included in the model fit, but it would remove the issue of unequal follow-up from the dataset.

Appendix H

Additional Comments on MCEM Variance Estimation Method

In this section, we provide some comments justifying our proposed approach to variance estimation after the MCEM algorithm. Our approach can be directly motivated by well-understood properties of multiple imputation.

Suppose D represents the complete data. In multiple imputation, we obtain M draws from $f(D^{(mis)}|D^{(obs)})$ (proper imputations). We then fit the desired model for $f(D)$ (in our case, the multistate cure model) using each one of the imputed datasets. For each imputed dataset, we obtain a set of parameter estimates and variances, and we use Rubin's combining rules to obtain a single estimate of the standard errors that correctly accounts for the uncertainty due to the imputation. This approach is well-understood and justified from a Bayesian perspective (Little and Rubin, 2002).

In our proposed method for estimating the standard errors, our goal is to obtain M approximate draws from $f(D^{(mis)}|D^{(obs)})$ and then apply Rubin's rules. At the end of the Monte Carlo EM algorithm, we obtain M draws from $f(D^{(mis)}|D^{(obs)};\theta^{(t)})$ where $\theta^{(t)}$ is the parameter estimate at the final iteration of the MCEM algorithm. These are imputations of the data, but they are "improper" ones. By "improper," we mean that the imputations were generated without correctly accounting for the uncertainty related to the missing data. It is well known in the missing data literature that inference using "improper" imputations can result in bad estimates for the standard errors (Little and Rubin, 2002). Our goal is to obtain draws of $f(D^{(mis)}|D^{(obs)})$, which would produce proper imputations to which we can apply Rubin's combining rules and obtain good standard error estimates.

We first note that

$$f(D^{(mis)}|D^{(obs)}) = \int f(D^{(mis)}|D^{(obs)};\theta)f(\theta|D^{(obs)})d\theta$$

We can obtain an approximate draw from $f(D^{(mis)}|D^{(obs)})$ by first obtaining a draw from $f(\theta|D^{(obs)})$. Then, given that draw, we draw from the conditional predictive distribution, $f(D^{(mis)}|D^{(obs)};\theta)$. We also note that

$$f(\theta|D^{(obs)}) = \int f(\theta|D^{(mis)}, D^{(obs)})f(D^{(mis)}|D^{(obs)})dD^{(mis)}$$

This integral decomposition suggests that we can obtain a draw from $f(\theta|D^{(obs)})$ by drawing from $f(\theta|D^{(mis)}, D^{(obs)})$ using the previous draw of $D^{(mis)}$ from $f(D^{(mis)}|D^{(obs)})$. This strategy for obtaining a draw from $f(D^{(mis)}|D^{(obs)})$ is described in detail in Little and Rubin (2002). Our proposed method for estimating the standard errors takes advantage of these existing approaches.

The goal of the post-processing step proposed is to obtain M independent draws from $f(D^{(mis)}|D^{(obs)})$ using the most recent imputations from the MCEM algorithm. We do the following to each imputed dataset. First, we fit the multistate cure model to a bootstrap sample of the imputed dataset and the complete data. This provides an approximate draw from $f(\theta|D^{(mis)}, D^{(obs)})$. This is a common approach to obtain an approximate draw from $f(\theta|D^{(mis)}, D^{(obs)})$ under flat priors. Given that draw, we re-impute the missing data from $f(D^{(mis)}|D^{(obs)};\theta)$. This provides a draw from $f(D^{(mis)}|D^{(obs)})$. After doing this for each imputed dataset separately and repeating for several iterations, we can obtain M independent approximate draws from $f(D^{(mis)}|D^{(obs)})$. We can then directly apply Rubin's rules.

Appendix I

Identifiability Issues Related to Multistate Cure Model

In the simulations in **Chapter IV**, we see that we can run into some numerical issues when we assume less restrictive assumptions for the $1 \rightarrow 4$ and $2 \rightarrow 4$ transition hazards. We believe the numerical issues are tied to identifiability issues inherent in the multistate cure model.

Just to review, the $1 \rightarrow 4$ and $2 \rightarrow 4$ transitions represent the transition to death from other causes from the non-cured and cured baseline states respectively. Suppose that we assume that subjects still at risk for recurrence after a certain threshold time τ are cured. For subjects known to be cured, we can attribute all of their events and time at risk to the $2 \rightarrow 4$ transition rather than the $1 \rightarrow 4$ transition. However, consider the non-cured subjects. All subjects known to be non-cured experienced the $1 \rightarrow 3$ transition, so we do not have any subjects with known events for the $1 \rightarrow 4$ transition. For subjects with missing cure status, it is unclear whether their time at risk for death from other causes should be attributed to the $2 \rightarrow 4$ or the $1 \rightarrow 4$ transition. Our inference about the $1 \rightarrow 4$ transition comes entirely from 1) the time subjects experiencing the $1 \rightarrow 3$ transition were at risk for the $1 \rightarrow 4$ transition, 2) the model for the probability of being non-cured, and 3) the assumptions we make linking the $2 \rightarrow 4$ and $1 \rightarrow 4$ transitions. It is, perhaps, unsurprising that we would then run into identifiability-related numerical problems when we do not make any additional assumptions about the $1 \rightarrow 4$ transition.

In this section, we will focus our attention to the situation in which we assume $\Lambda_{14}(t) \propto \Lambda_{24}(t)$ assumption. We note that there is only one additional parameter in the model that assumes $\Lambda_{14}(t) \propto \Lambda_{24}(t)$ compared to a model that assumes $\Lambda_{14}(t) = \Lambda_{24}(t)$, and yet in our simulations we see some evidence of undercoverage of the logistic model intercept when we allow the hazards to be proportional. Here, we will briefly explore identifiability issues related to the proportionality assumption and explore the form of the observed data log-likelihood for the simulated data under the $\Lambda_{14}(t) \propto \Lambda_{24}(t)$ assumption.

We assume that we have fully-observed covariates and equal follow-up of the two outcomes. The complete data likelihood (treating cure status as known) can be expressed as

$$L(\theta|\mathbb{D}) = \prod_{i=1}^{n} \left(P(G_i = 0)\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id})\right)^{1-G_i}$$
$$\times \Big(P(G_i = 1) \left[\lambda_{14}(Y_{id})^{\delta_{id}} S_1(Y_{id})\right]^{1-\delta_{ir}} \left[\lambda_{13}(Y_{ir})S_1(Y_{ir})\lambda_{34}(Y_{id} - Y_{ir})^{\delta_{id}} S_3(Y_{id} - Y_{ir})\right]^{\delta_{ir}}\Big)^{G_i}$$

Let R_i be an indicator for whether the cure status for subject i is known. Let $\mathbb{D}^{obs}$ represent the observed information in $\mathbb{D}$. Here, subjects assumed to be cured have $R_i = 1$. The observed data likelihood is

$$\begin{aligned} L^{obs}(\theta|\mathbb{D}^{obs}) =& \prod_{i=1}^{n} \left[P(G_i = 1)\lambda_{13}(Y_{ir})S_1(Y_{ir})\lambda_{34}(Y_{id} - Y_{ir})^{\delta_{id}} S_3(Y_{id} - Y_{ir})\right]^{\delta_{ir}} \\ & \times \left[P(G_i = 0)\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id}) + P(G_i = 1)\lambda_{14}(Y_{id})^{\delta_{id}} S_1(Y_{id})\right]^{(1-\delta_{ir})(1-R_i)} \\ & \times \left[P(G_i = 0)\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id})\right]^{(1-\delta_{ir})R_i} \end{aligned}$$

Suppose we assume that $\Lambda_{14}(t) = \Lambda_{24}(t)\exp\{\beta_0\}$. For notational convenience, we will also assume we use the same covariate set for all components of the multistate cure model. We have that

$$\begin{aligned} L^{obs}(\theta|\mathbb{D}^{obs}) = \prod_{i=1}^{n} \frac{1}{1 + e^{\alpha_0+\alpha_1 X_i}} & \left[e^{\alpha_0+\alpha_1 X_i} S_2(Y_{ir})^{\exp(\beta_0)} \lambda_{13}(Y_{ir}) e^{-\Lambda_{13}(Y_{id})}\right. \\ & \left. \times\lambda_{34}(Y_{id} - Y_{ir})^{\delta_{id}} S_3(Y_{id} - Y_{ir})\right]^{\delta_{ir}} \\ \times \left[\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id}) + e^{\alpha_0+\alpha_1 X_i} e^{\beta_0 \delta_{id}} S_2(Y_{ir})^{\exp(\beta_0)} \lambda_{24}(Y_{id})^{\delta_{id}} e^{-\Lambda_{13}(Y_{id})}\right]^{(1-\delta_{ir})(1-R_i)} & \\ \times \left[\lambda_{24}(Y_{id})^{\delta_{id}} S_2(Y_{id})\right]^{(1-\delta_{ir})R_i} & \end{aligned}$$

We notice that β_0 only ever appears alongside α, but both parameters appear to be identifiable.

We consider two different datasets **simulated under the same model** as in simulation scenario 1 from **Chapter IV** with no missingness in covariates and no unequal follow-up, assuming equal hazards for the 2 → 4 and 1 → 4 transitions. Both simulated datasets are generated using the same parameter values. We will use these two simulated datasets to illustrate the numerical issues tied to estimating β_0 under assumptions that $\Lambda_{14}(t) = \Lambda_{24}(t)\exp\{\beta_0\}$ and demonstrate that the extent of identifiability-related numerical issues may vary across different datasets generated from the same model. Dataset 1 is an example of a dataset producing an observed data log-likelihood that is nearly flat, resulting in difficulty in estimating β_0 and subsequent difficult with α_0. For Dataset 2, the observed data log-likelihood is easier to maximize, and we can estimate β_0 reasonably well. We note that since the data were simulated assuming equal hazards for the 2 → 4 and 1 → 4 transition, the true value of β_0 is 0.

For each dataset, we fit a multistate cure model using the EM algorithm under Weibull baseline hazards and assuming either 1) $\Lambda_{14}(t) = \Lambda_{24}(t)\exp\{\beta_0\}$ or 2) $\Lambda_{14}(t) = \Lambda_{24}(t)$. Evaluation of the corresponding observed data log-likelihoods indicates that the EM algorithm reached convergence for each model fit. **Table I.1** shows the resulting multistate cure model fits. For Dataset 1, we see that the estimated β_0 is far from the true value of zero in Fit 1, but the confidence interval covers zero. In both model fits for Dataset 1, we see that the confidence intervals for $\beta_{24,14}$ do not cover the true values. We notice that the estimated intercept value is lower when we assume $\Lambda_{14}(t) = \Lambda_{24}(t)\exp\{\beta_0\}$ for Dataset 1, and the confidence interval under equality assumptions does not cover the estimate under proportionality assumptions. For Dataset 2, the estimated β_0 is close to the true value of zero, and the confidence intervals all cover the true values. The model fits are very similar under the two sets of assumptions for Dataset 2.

Table I.1: Multistate Cure Model Fits to Two Simulated Datasets

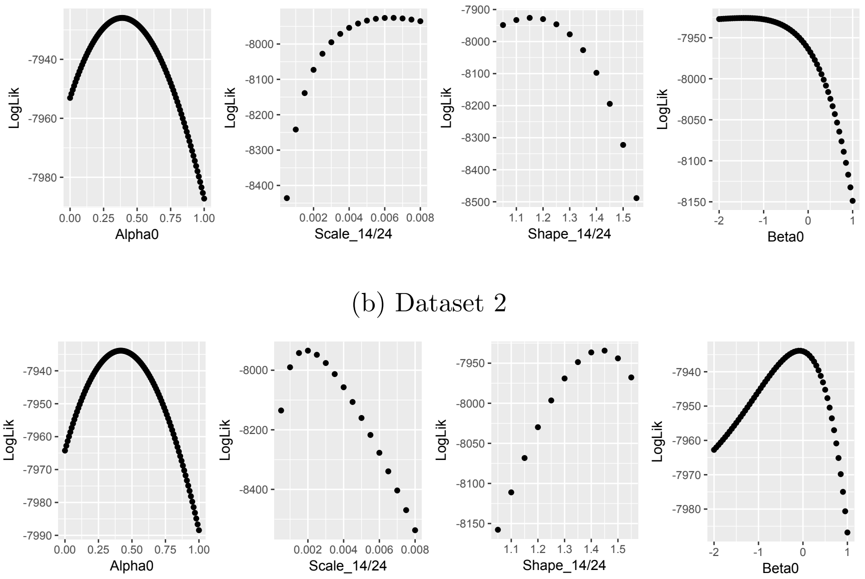
		Dataset 1		Dataset 2	
		Fit 1	Fit 2	Fit 3	Fit 4
Assumption*		Proportional	Equal	Proportional	Equal
Parameter	Truth	log-HR (95% CI)	log-HR (95% CI)	log-HR (95% CI)	log-HR (95% CI)
β_{13} X_1	0.5	0.50 (0.42, 0.58)	0.51 (0.43, 0.59)	0.47 (0.39, 0.56)	0.47 (0.39, 0.56)
β_{13} X_2	0.5	0.54 (0.46, 0.62)	0.54 (0.44, 0.63)	0.50 (0.41, 0.59)	0.50 (0.42, 0.57)
$\beta_{24,14}$ X_1	0.5	0.30 (0.17, 0.43)	0.26 (0.14, 0.40)	0.49 (0.32, 0.66)	0.49 (0.35, 0.62)
$\beta_{24,14}$ X_2	0.5	0.66 (0.52, 0.80)	0.63 (0.51, 0.75)	0.46 (0.31, 0.60)	0.46 (0.32, 0.59)
β_0	0	-1.46 (-3.66, 0.80)	-	-0.07 (-1.16, 1.01)	-
β_{34} X_1	0.5	0.48 (0.41, 0.55)	0.48 (0.40, 0.56)	0.47 (0.40, 0.55)	0.47 (0.41, 0.54)
β_{34} X_2	0.5	0.43 (0.35, 0.50)	0.43 (0.35, 0.50)	0.50 (0.44, 0.55)	0.50 (0.44, 0.57)
α Intercept	0.5	0.38 (0.18, 0.59)	0.53 (0.40, 0.66)	0.41 (0.28, 0.55)	0.42 (0.31, 0.53)
α X_1	0.5	0.48 (0.34, 0.61)	0.49 (0.36, 0.62)	0.56 (0.44, 0.68)	0.56 (0.42, 0.71)
α X_2	0.5	0.51 (0.39, 0.63)	0.58 (0.44, 0.73)	0.40 (0.28, 0.53)	0.40 (0.30, 0.51)

*Proportional: Assume $\Lambda_{14}(t) = \Lambda_{24}(t)\exp\{\beta_0\}$. Equal: Assume $\Lambda_{14}(t) = \Lambda_{24}(t)$

We now want to explore the shape of the observed data log-likelihood for Fits 1 and 3 (proportional 1 → 4 and 2 → 4 hazards for Datasets 1 and 2) at their respective EM-maximized values. We recall that α_0 is the intercept from the logistic part of the model, β_0 is the proportionality parameter, and the shape, scale, and beta parameters are assumed to be equal for the 1 → 4 and 2 → 4 transitions. We first look at the shape of the observed data log-likelihood varying parameters one at a time and *keeping all other parameters fixed at the EM-maximized values.* **Figure I.1** shows the profile log-likelihood values for several different parameters. We notice that the profile likelihood curve for β_0 appears nearly flat for Dataset 1 and has a clear peak for Dataset 2.

Figure I.1: Profile of Observed Data Log-Likelihood under Proportional Baselines (at the EM-Maximized Values for Two Simulated Datasets)

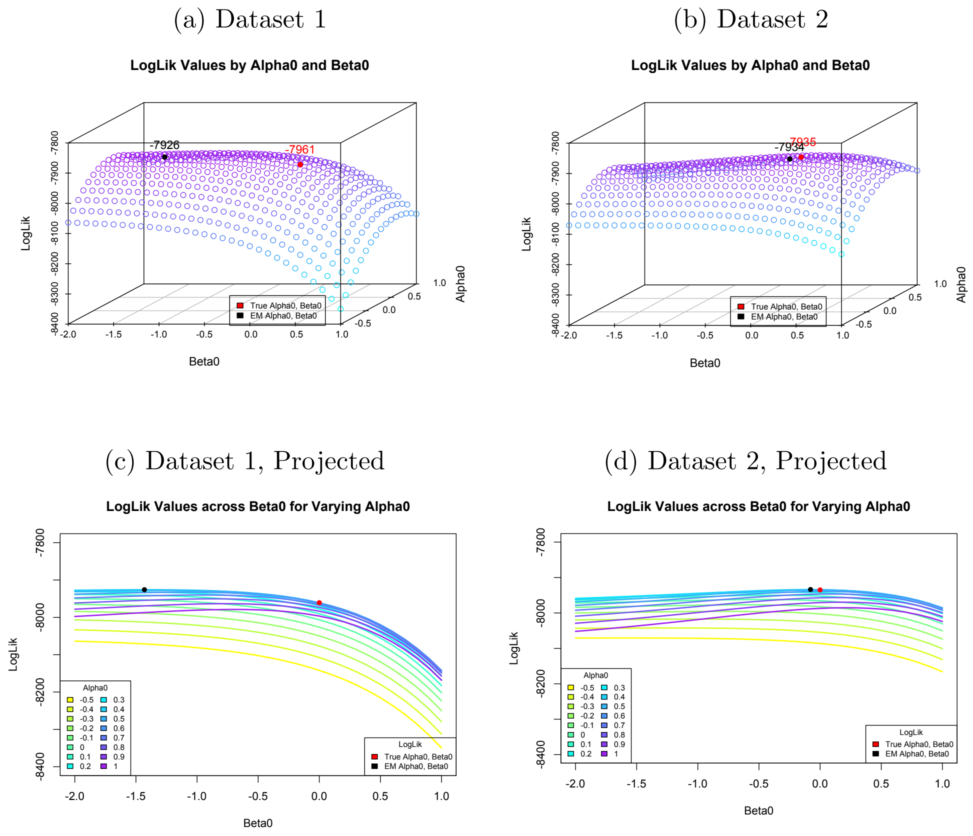
(a) Dataset 1



(b) Dataset 2

We also explore the shape of the log-likelihood surface varying both β_0 and α_0 from their EM-maximized values in **Figure I.2**. **Figures I.2(a) and I.2(b)** show a 3-dimensional surface for Datasets 1 and 2, and **Figures I.2(c) and I.2(d)** show the same curve projected onto the β_0 x log-likelihood plane. The log-likelihood appears particularly flat for Dataset 1, and the EM-maximized value for β_0 differs from the true value substantially. For Dataset 2, the EM-maximized value is near the true value. Based on these plots, it is unsurprising that we have difficulty estimating β_0 for Dataset 1, but we are able to estimate β_0 well for Dataset 2.

Figure I.2: Log-Likelihood Surface under Proportional Baseline Hazards (at the EM-Maximized Values varying α_0 and β_0 for Two Simulated Datasets)



(a) Dataset 1

(b) Dataset 2

(c) Dataset 1, Projected

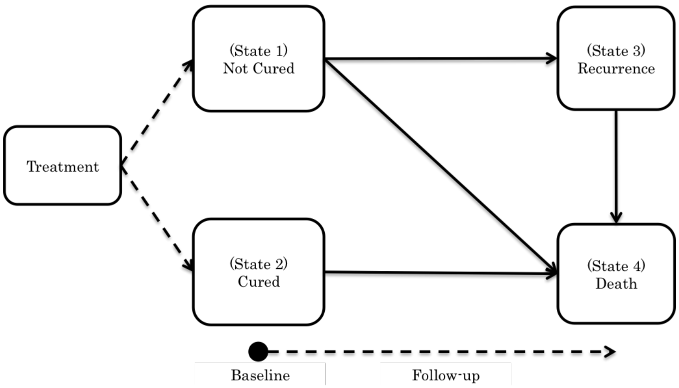
(d) Dataset 2, Projected

Appendix J

Relaxing Hazard Restrictions through Shrinkage

In **Chapter IV** of this dissertation, we develop an EM algorithm for fitting the multistate cure model in **Figure J.1**. We explore different restrictions we can make on the transitions to death from other causes (transitions 1 → 4 and 2 → 4 in diagram below), and in the course of our simulations, we demonstrate at relaxing the restrictions on these two transitions can result in some numerical trouble.

Figure J.1: Diagram of the Multistate Cure Model



In this appendix, we briefly explore a shrinkage-based approach that may allow us to relax some restrictions on the 1 → 4 and 2 → 4 transition hazards while still avoiding some numerical issues. Recall that we explore four different sets of restrictions on the baseline hazards: no restrictions, $\Lambda_{14}(t) \propto \Lambda_{24}(t)$, $\Lambda_{14}^0(t) = \Lambda_{24}^0(t)$, and $\Lambda_{14}(t) = \Lambda_{24}(t)$. The best numerical properties were obtained under the most restrictive assumption, $\Lambda_{14}(t) = \Lambda_{24}(t)$.

Rather than fully restricting the hazards to be equal, we propose shrinking the hazards toward each other. This allows the hazards to be unequal if the data support it and will shrink the hazards toward each other if the data do not support a difference. Suppose, for example, we are unsure if we can assume that $\beta_{24} = \beta_{14}$. We can write

$$\Lambda_{24}(t) = \Lambda_{24}^{0}(t)e^{\beta_{24}X}$$
$$\Lambda_{14}(t) = \Lambda_{14}^{0}(t)e^{\beta_{24}X+(\beta_{14}-\beta_{24})X}$$

In the model likelihood, we can impose a ridge penalty or other shrinkage penalty on the parameter $\beta_{14} - \beta_{24}$. Under this modified likelihood, the E-Step of the EM algorithm and the imputation step of the MCEM algorithm are both unchanged. The modified likelihood impacts the M Step in both cases. As before, we can write the terms we want to maximize for β and the baseline hazard parameters in the form of a single Cox regression model fit. However, we will perform the maximization incorporating the ridge penalty on $\beta_{14} - \beta_{24}$. Along with the ridge penalty comes a tuning parameter related to the amount of penalization. We propose using the software standard approach for determining the tuning parameter. In the *survival* package in R, an estimate for the approximate degrees of freedom is used for the tuning parameter. More sophisticated methods can be used to determine a reasonable tuning parameter, but by using the software standard, we can make the modification to the proposed EM algorithms extremely easy to implement.

We can use a similar shrinkage trick for the restriction $\Lambda_{14}(t) \propto \Lambda_{24}(t)$. We have that

$$\Lambda_{14}(t) = \Lambda_{24}(t)e^{\beta_0} = \Lambda_{24}^{0}(t)e^{\beta_{24}X+\beta_0}$$

As before, we can apply a ridge penalty on β_0, and this will shrink $\Lambda_{24}(t)$ towards $\Lambda_{14}(t)$.

We perform a small simulation study to explore the impact of the shrinkage on the numerical stability of the algorithm (in terms of bias, coverage, and empirical variance). We generate 200 simulated datasets under scenario 1 in **Chapter IV**, where the data are generated under a multistate cure model with no covariate missingness or unequal censoring. For each simulated dataset, we fit a multistate cure model using the proposed EM algorithm. We estimate the bias, empirical variance, and coverage for the resulting parameter estimates and standard errors.

We note that, for these data, the truth is that $\Lambda_{14}(t) = \Lambda_{24}(t)$ (so $\beta_{24} = \beta_{14}$ and $\beta_0 = 0$). In the future, we will explore how the shrinkage approach impacts the numerical properties when the model is misspecified under the stronger set of restrictions.

Tables J.1 and J.2 shows the simulation results. In all the scenarios considered, we don't have too much bias in estimating parameters in β. However, when we only restrict the baseline hazards to be equal, we see increased variances for estimating β_{24} and β_{14}. These variances are substantially reduced when we apply shrinkage to $\beta_{14} - \beta_{24}$. In previous simulations and in **Table J.2**, we see that undercoverage for the intercept parameter of the logistic regression model when we assume the hazards are proportional. When we apply shrinkage to β_0, we see that the undercoverage goes away. We note that when we apply the shrinkage, the estimation time goes up substantially. This is due to the variance estimation. In all of the simulations presented here, we use bootstrap methods to estimate the standard errors (recall, we are using the EM algorithm, not the MCEM algorithm). There is a small increase in time related to fitting the survival model with the ridge penalty, and this time is compounded substantially as the function is called many times within the bootstrap variance estimation procedure. Additionally, in the Weibull case with shrinkage, there is a substantial number of the 200 simulations that have some numerical issues. These arise in the variance estimation, where the survival model fits (using 'survreg' in R) have numerical problems for one or more of the bootstrap samples. The improved numerical properties in terms of bias, empirical variance, and coverage may be explained by failure of the shrinkage-based estimation on the simulated datasets that are more challenging estimation-wise without the shrinkage. Additional explorations are certainly needed in the future, but these quick simulations do suggest that shrinkage may provide some means for improving the statistical properties for the more relaxed model assumptions among the non-failing datasets.

Table J.1: Multistate Cure Model Failure Time Model Estimates (Shrinkage)

Results across 200 simulations are presented using the following notation: Bias (Empirical Variance) Coverage of 95% Confidence Interval, each multiplied by 100. States 1 and 2 are the non-cured and cured baseline states. States 3 and 4 represent recurrence and death.

Baseline	$2 \rightarrow 4$, $1 \rightarrow 4$	Failure Time Models							
Hazard	Assumption	$1 \rightarrow 3$		$2 \rightarrow 4$		$1 \rightarrow 4$		$3 \rightarrow 4$	
		X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
				Scenario 1: No Covariate Missingness or Unequal Follow-up					
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.19) 92	0 (0.14) 97	0 (0.51) 93	0 (0.44) 96	0 (0.51) 93	0 (0.44) 96	0 (0.17) 93	0 (0.14) 97
Weibull	$\Lambda_{14}^0(t) = \Lambda_{24}^0(t)$	0 (0.20) 93	0 (0.15) 97	0 (0.87) 96	0 (0.94) 96	-1 (2.86) 96	1 (2.86) 94	0 (0.17) 94	0 (0.17) 96
Weibull	plus SHRINK β	0 (0.20) 94	0 (0.15) 97	0 (0.50) 96	0 (0.50) 95	0 (0.70) 95	0 (0.54) 98	0 (0.17) 95	0 (0.14) 96
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.19) 93	0 (0.15) 97	0 (0.50) 95	0 (0.46) 95	0 (0.50) 95	0 (0.46) 95	0 (0.17) 93	0 (0.14) 97
Weibull	plus SHRINK β_0	0 (0.20) 93	0 (0.14) 97	-2 (0.42) 96	0 (0.43) 95	-2 (0.42) 96	0 (0.43) 95	0 (0.19) 92	0 (0.14) 95
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.20) 93	0 (0.15) 97	0 (0.51) 95	0 (0.45) 96	0 (0.51) 95	0 (0.45) 96	0 (0.17) 94	0 (0.17) 96
Cox	$\Lambda_{14}^0(t) = \Lambda_{24}^0(t)$	0 (0.20) 94	0 (0.16) 95	0 (0.92) 96	0 (0.93) 98	-2 (3.02) 97	1 (2.85) 96	0 (0.17) 94	0 (0.15) 96
Cox	plus SHRINK β	0 (0.19) 94	0 (0.15) 97	0 (0.54) 94	0 (0.50) 97	-1 (0.71) 93	0 (0.61) 97	0 (0.17) 94	0 (0.15) 96
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.20) 91	0 (0.15) 98	0 (0.50) 95	0 (0.46) 95	0 (0.50) 95	0 (0.46) 96	0 (0.17) 95	0 (0.15) 98
Cox	plus SHRINK β_0	0 (0.19) 93	0 (0.15) 98	0 (0.51) 96	0 (0.45) 95	0 (0.51) 96	0 (0.45) 95	0 (0.17) 94	0 (0.15) 97

Table J.2: Multistate Cure Model Logistic Model Estimates (Shrinkage)

Results across 200 simulations are presented using the following notation: Bias (Empirical Variance) Coverage of 95% Confidence Interval, each multiplied by 100. The number of simulations (out of 200) with numerical issues and the median run time per simulation are also shown.

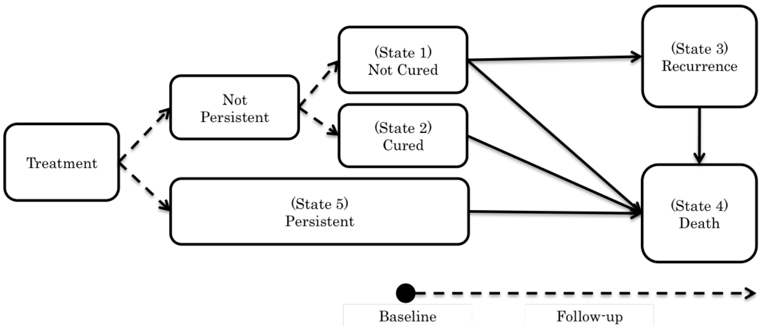
Baseline	$2 \to 4$, $1 \to 4$		Logistic Model		# Failed	Run Time
Hazard	Assumption	Intercept	X_1	X_2	(out of 200)	(mins/sim)
		Scenario 1: No Covariate Missingness or Unequal Follow-up				
Weibull	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.32) 94	0 (0.43) 94	0 (0.35) 98	0	1.94
Weibull	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.31) 94	0 (0.53) 94	0 (0.44) 95	0	1.72
Weibull	plus SHRINK β	0 (0.31) 95	0 (0.43) 93	0 (0.35) 97	12	4.75
Weibull	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	0 (0.55) 93	0 (0.42) 95	0 (0.36) 99	0	3.07
Weibull	plus SHRINK β_0	0 (0.32) 96	0 (0.41) 96	0 (0.34) 97	47	4.54
Cox	$\Lambda_{14}(t) = \Lambda_{24}(t)$	0 (0.32) 97	0 (0.43) 95	0 (0.35) 97	0	9.38
Cox	$\Lambda^0_{14}(t) = \Lambda^0_{24}(t)$	0 (0.31) 94	0 (0.54) 94	0 (0.45) 96	0	9.65
Cox	plus SHRINK β	-1 (0.31) 92	0 (0.44) 93	0 (0.36) 97	0	29.3
Cox	$\Lambda_{14}(t) \propto \Lambda_{24}(t)$	1 (0.88) 84	0 (0.43) 94	0 (0.41) 97	0	9.62
Cox	plus SHRINK β_0	-2 (0.31) 94	0 (0.42) 95	0 (0.35) 98	0	28.6

Appendix K

Additional HNSCC Results for Multistate Model with Persistence

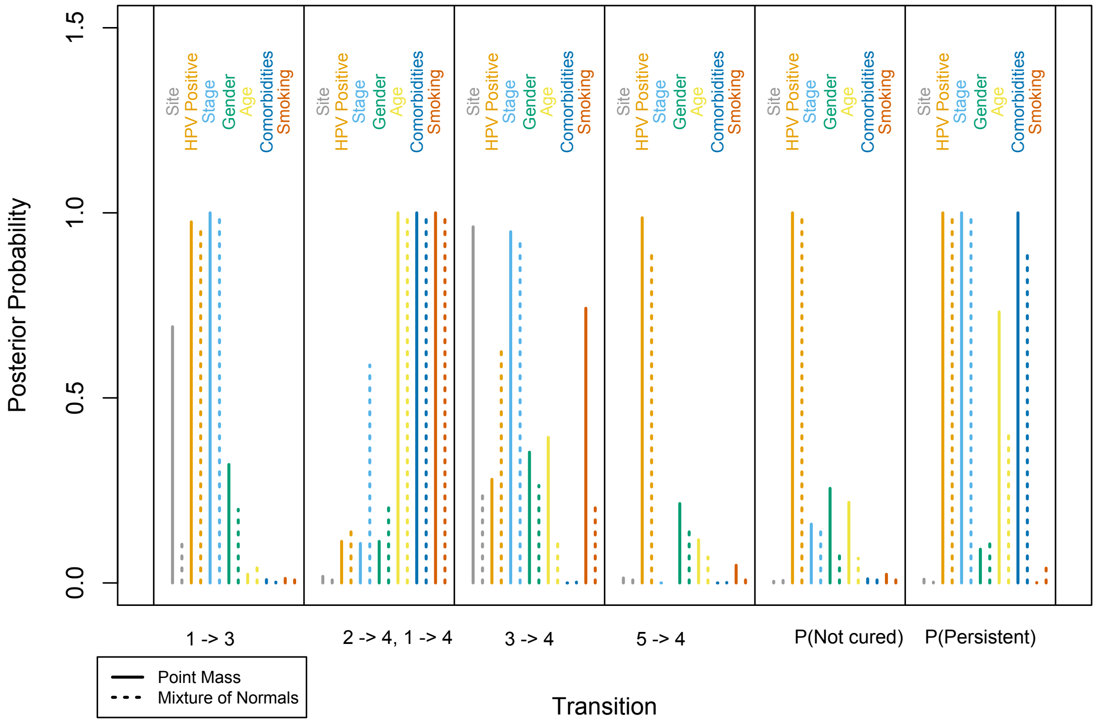
In this appendix, we present some additional exploration into the multistate cure model with persistence fits to the head and neck data. We recall the model structure in **Figure K.1**:

Figure K.1: Diagram of Multistate Cure Model with Persistence



First, we focus on the Bayesian estimation-based fits. **Figure K.2** shows the posterior inclusion probabilities for groups of variables under the two spike and slab priors considered. These probabilities are calculated as the proportion of iterations of the MCMC algorithm in which the group of variables was assigned $\gamma_g = 1$. The two priors tend to give similar results, but this is not always the case. The fit with the point mass prior has a tendency to have larger posterior inclusion rates. This may be a result of our choice of hyperparameters for the two priors. We see large differences in the posterior inclusion probabilities for site for the $1 \to 3$ transition (recurrence given non-cured); cancer stage in the $1 \to 4, 2 \to 4$ transitions (death from other causes given not persistent); cancer site, age, and smoking status for the $3 \to 4$ transition (death after recurrence), and age for the model for persistence.

Figure K.2: Posterior Inclusion Probabilities



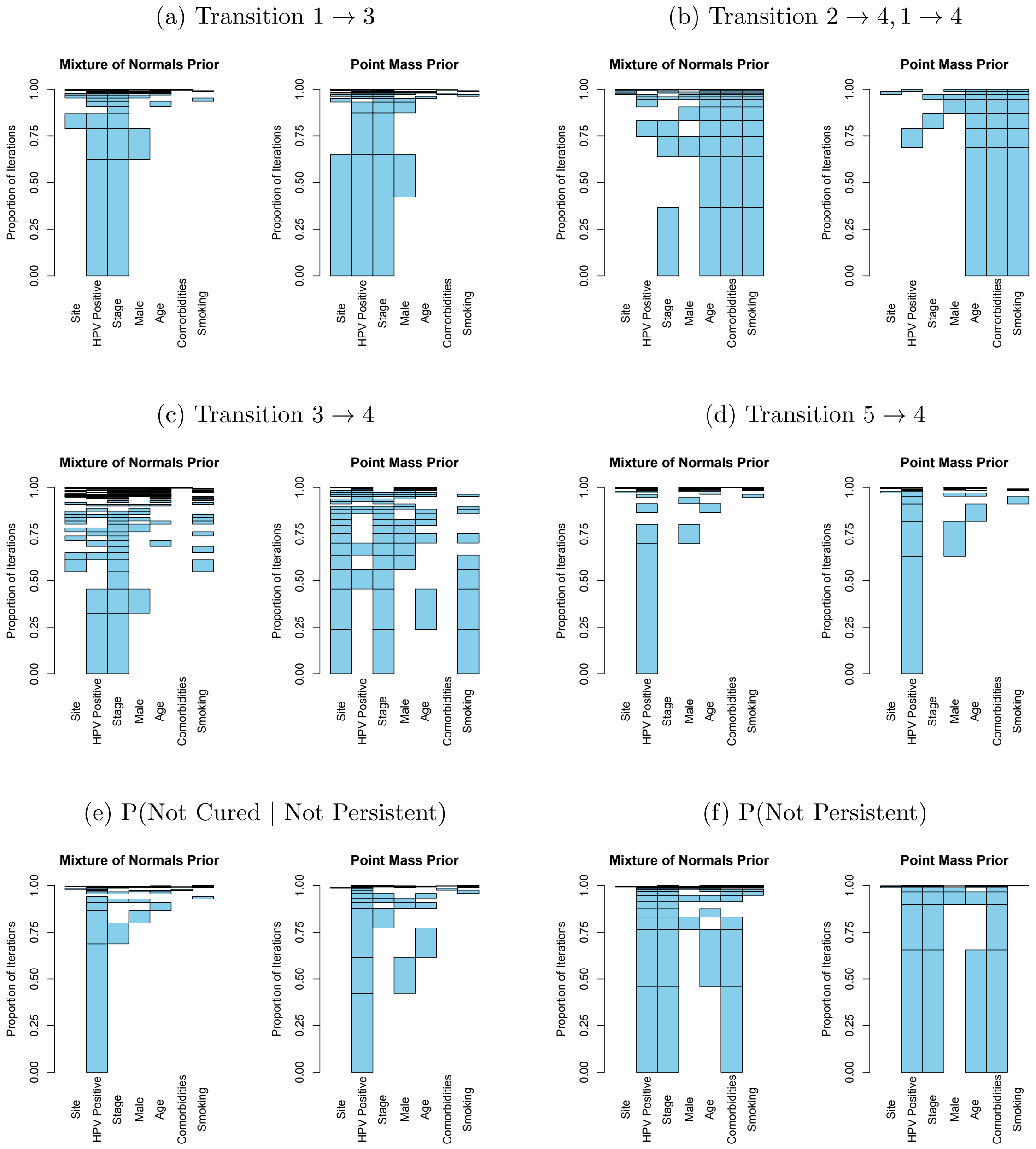
The line height for a group indicates the proportion of MCMC iterations in which that group was included in the model.

Figure K.3 visualizes the (cumulative) proportion of MCMC iterations each combination of predictors was chosen, where the filled-in area corresponds to the chosen covariates and the height of the filled-in area corresponds to the proportion of iterations in which that combination of covariates was chosen. The various model formulations chosen by the MCMC algorithm for each submodel were sorted from most to least often chosen. We note that the proportion of colored area above a particular variable in these plots corresponds to the posterior inclusion probability for that variable in the corresponding submodel. This type of plot was used in Chipman (1996) as a way to visualize the highest posterior models. For example, consider the first plot under Transition 1 $\rightarrow$ 3. The results suggest that the combination of HPV and cancer stage was the model chosen for over 60% of the MCMC iterations. The cancer subsite variables (hypopharynx, larynx, oropharynx) were included in only a few of the iterations; they were included with HPV and stage for about 5% of iterations and in some other covariate combinations for smaller fractions of the iterations. They were excluded from the model for the majority of the iterations. In contrast, cancer stage was included in all iterations. Overall, there are many instances in which the posterior weight put on different combinations of variables differs between the two priors.

We use the posterior means of θ (using Bayesian model averaging) for each of the four Bayesian model fits to predict the 5-year overall survival probability. Given θ, the estimated overall survival probability can be calculated using the state occupancy probabilities derived in **Section 5.5** using a time of 5 years. **Figure K.4** compares the predictions across the Bayesian fits. The four Bayesian fits give nearly identical 5-year OS predictions. We obtain similar results when we compare the predicted 5-year event-free survival rates.

Figure K.5 compares the 5-year OS predictions for the Bayesian fit without shrinkage and the two maximum likelihood estimation-based fits (with and without ridge shrinkage). The two fits without any selection/shrinkage give very similar predictions, but we obtain different predictions for the ridge-penalized model fit. For this fit, the predicted 5-year survivals appear to be more clustered around the population average.

Figure K.3: Posterior Probabilities for Different Covariate Combinations



(a) Transition 1 → 3

(b) Transition 2 → 4, 1 → 4

(c) Transition 3 → 4

(d) Transition 5 → 4

(e) P(Not Cured | Not Persistent)

(f) P(Not Persistent)

These plots show the proportion of MCMC iterations in which each combination of covariates was chosen for inclusion for each submodel. A filled-in rectangle indicates that the covariate group was included in the model. The height of the rectangle indicates the proportion of iterations.

Figure K.4: 5-Year OS Predictions across Bayes Fits

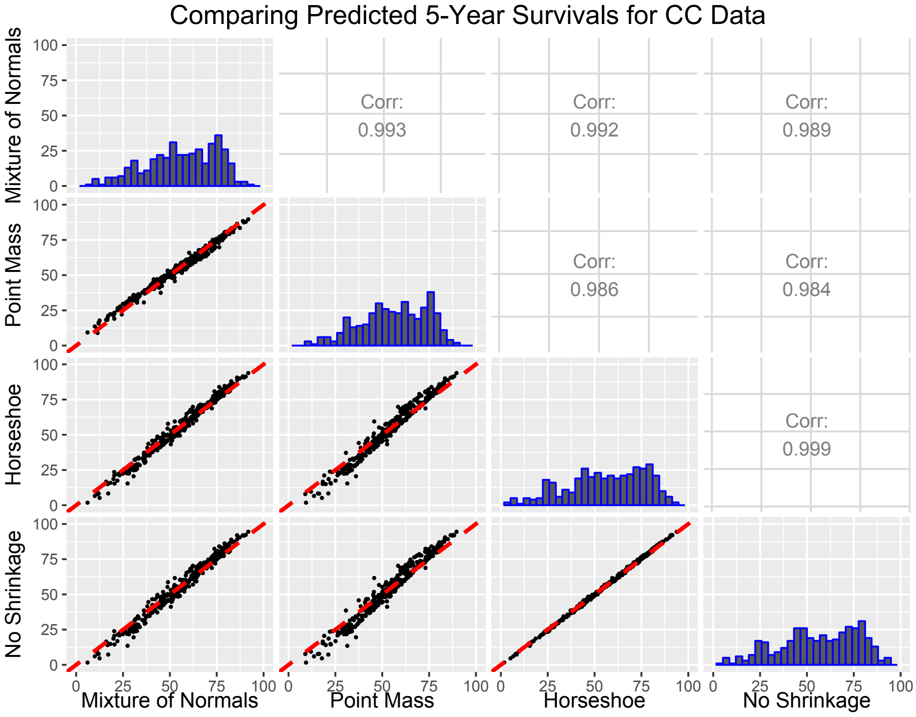


Figure K.5: 5-Year OS Predictions across MLE and Bayes Fits

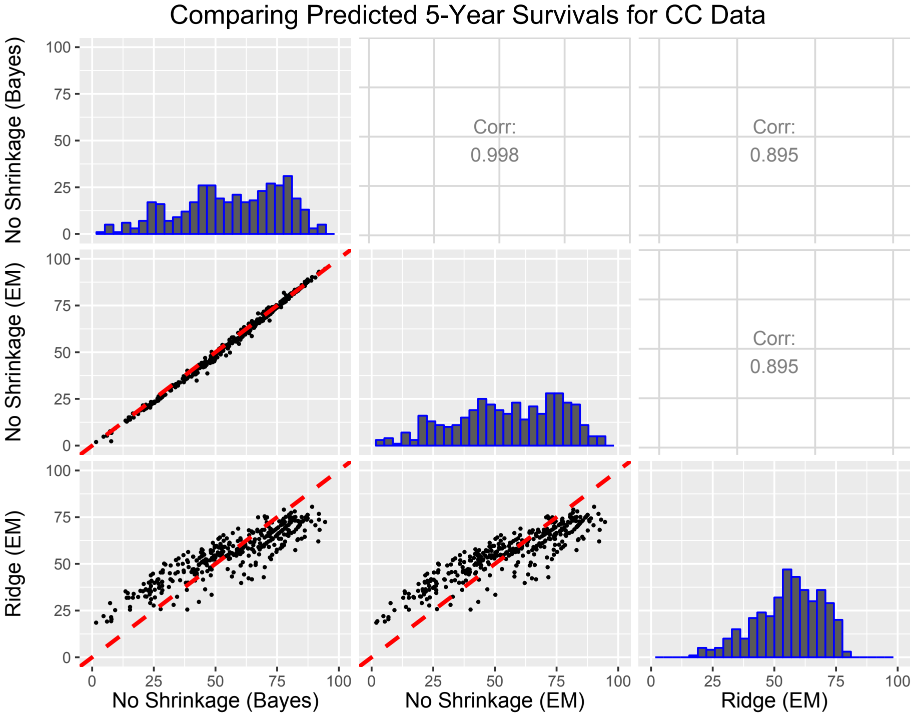


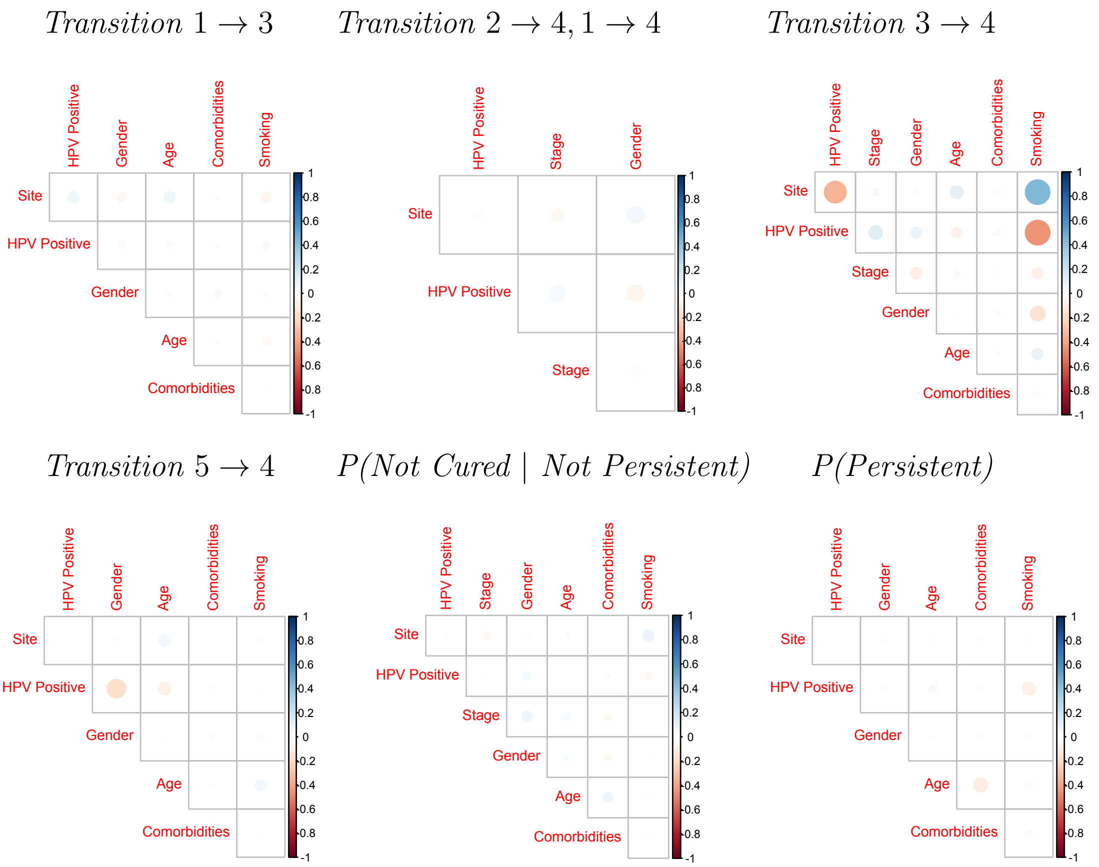
Figure K.6 shows the correlations of the posterior draws of γ across iterations for each transition under the mixture of normals prior. If γ always equals zero or one (for all iterations) for a particular group, that group is not plotted. We can see that the correlations are small for all but the $3 \rightarrow 4$ transition. For this transition, we have some mild correlations in the inclusion/exclusion across groups. This may be due to a smaller amount of available data for this transition. In the main paper, we expressed a concern about correlation of the inclusion/exclusion indicators for a particular group across submodels. **Figure K.7** shows the correlations across submodels of interest for the mixture of normals prior. We do not see evidence of correlation in γ across submodels.

Figure K.8 shows the correlations of the posterior draws of γ across iterations for each transition under the point mass at zero prior. We see greater evidence for γ correlation issues within the $3 \rightarrow 4$ transition than we did with the mixture of normals prior. **Figure K.9** shows the correlations across submodels of interest for the point mass at zero prior. Again, we see greater evidence of cross-submodel correlation in γ than we did for the mixture of normals prior.

This last set of figures looking at the correlations of γ indicates that we may be at a higher risk of correlation issues using the point mass at zero prior. This may be due to the reversible jump algorithm, which only makes small modifications in the covariate set at each iteration. We do see that the correlations across submodels are generally very small. It is important to note that these diagnostics are looking at the correlation for γ, not the resulting values of θ. The correlation structure for θ may be different.

Figure K.10 shows the confidence and credible intervals for each one of the model parameters under each of the 6 methods (four Bayesian methods, 2 maximum likelihood methods). We tend to see similar results between the methods with no shrinkage (Bayesian and MLE), but the ridge regression fit produces very different estimates from the other methods.

Figure K.6: Correlations of Gamma within Submodels (Mixture of Normals)



Transition 1 → 3

Transition 2 → 4, 1 → 4

Transition 3 → 4

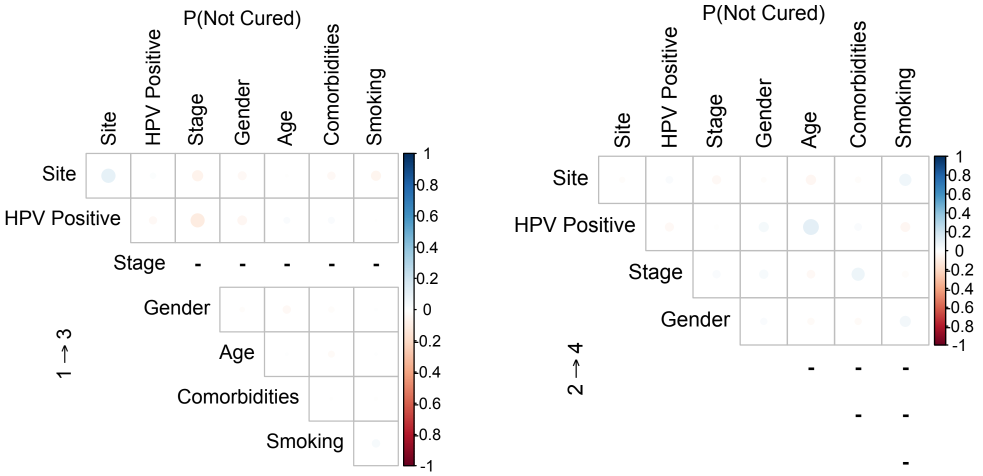
Transition 5 → 4

P(Not Cured | *Not Persistent)*

P(Persistent)

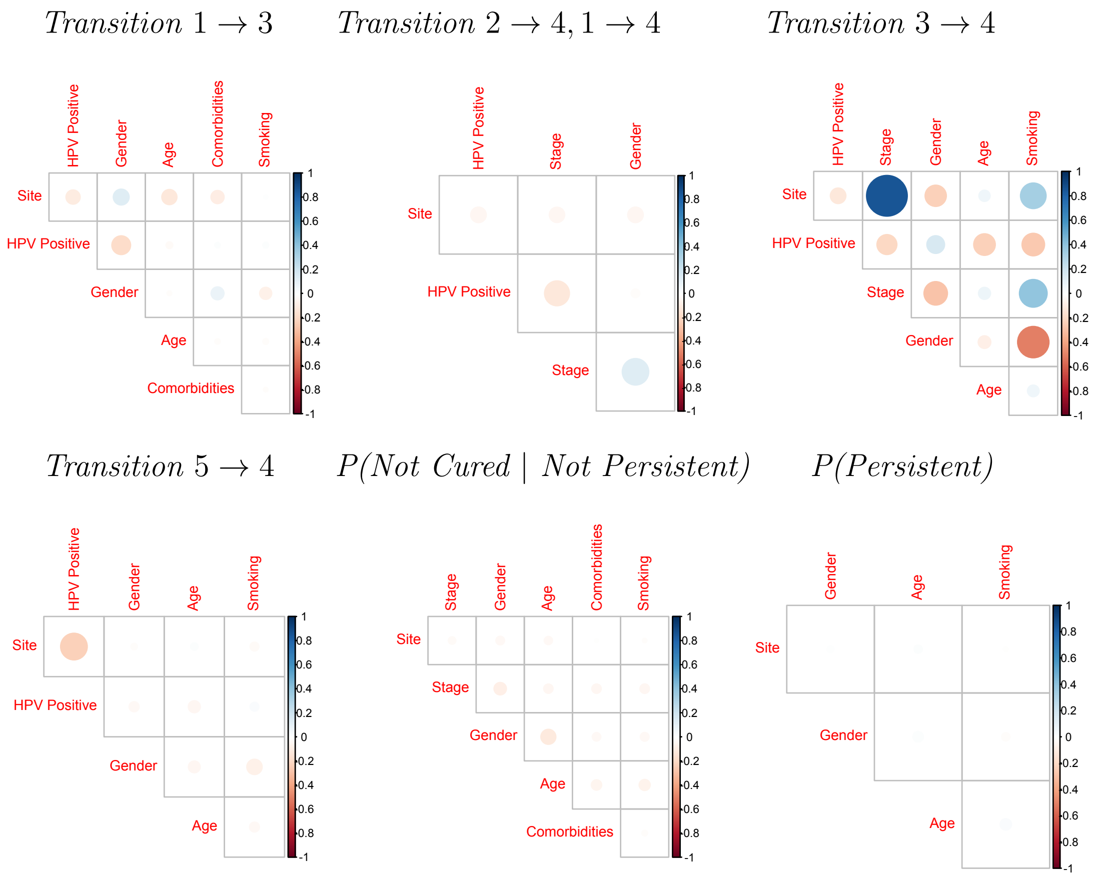
This figure presents heatmaps for the correlation of the γ indicators across MCMC iterations *within* submodels. If γ was always 0 or 1, that group was not included in the plot.

Figure K.7: Correlations of Gamma across Submodels (Mixture of Normals)



This figure presents heatmaps for the correlation of the γ indicators across MCMC iterations *across* submodels. If γ was always 0 or 1 for one of the variables, the correlation is listed as '-'

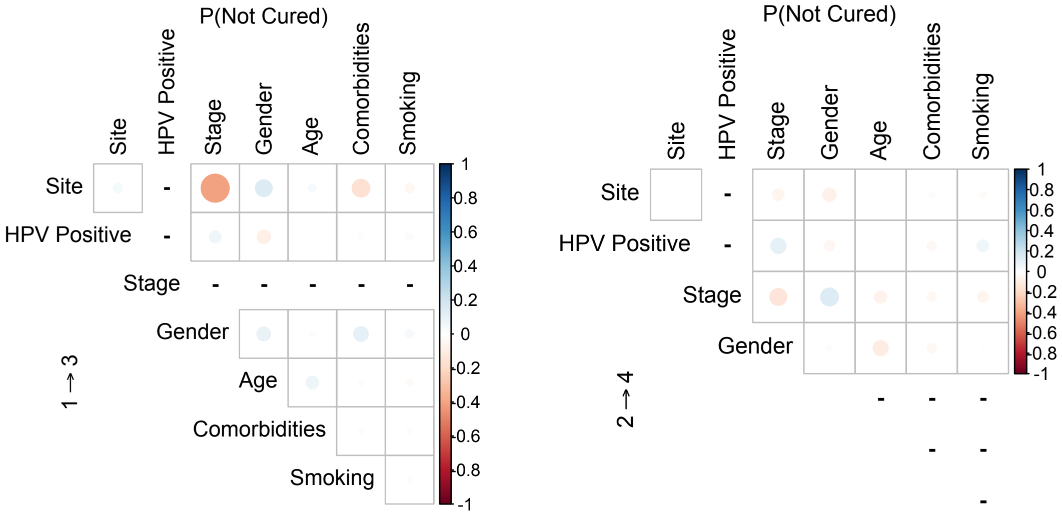
Figure K.8: Correlations of Gamma within Submodels (Point Mass at Zero)


Transition 1 → 3 *Transition* 2 → 4, 1 → 4 *Transition* 3 → 4

Transition 5 → 4 *P(Not Cured* | *Not Persistent)* *P(Persistent)*

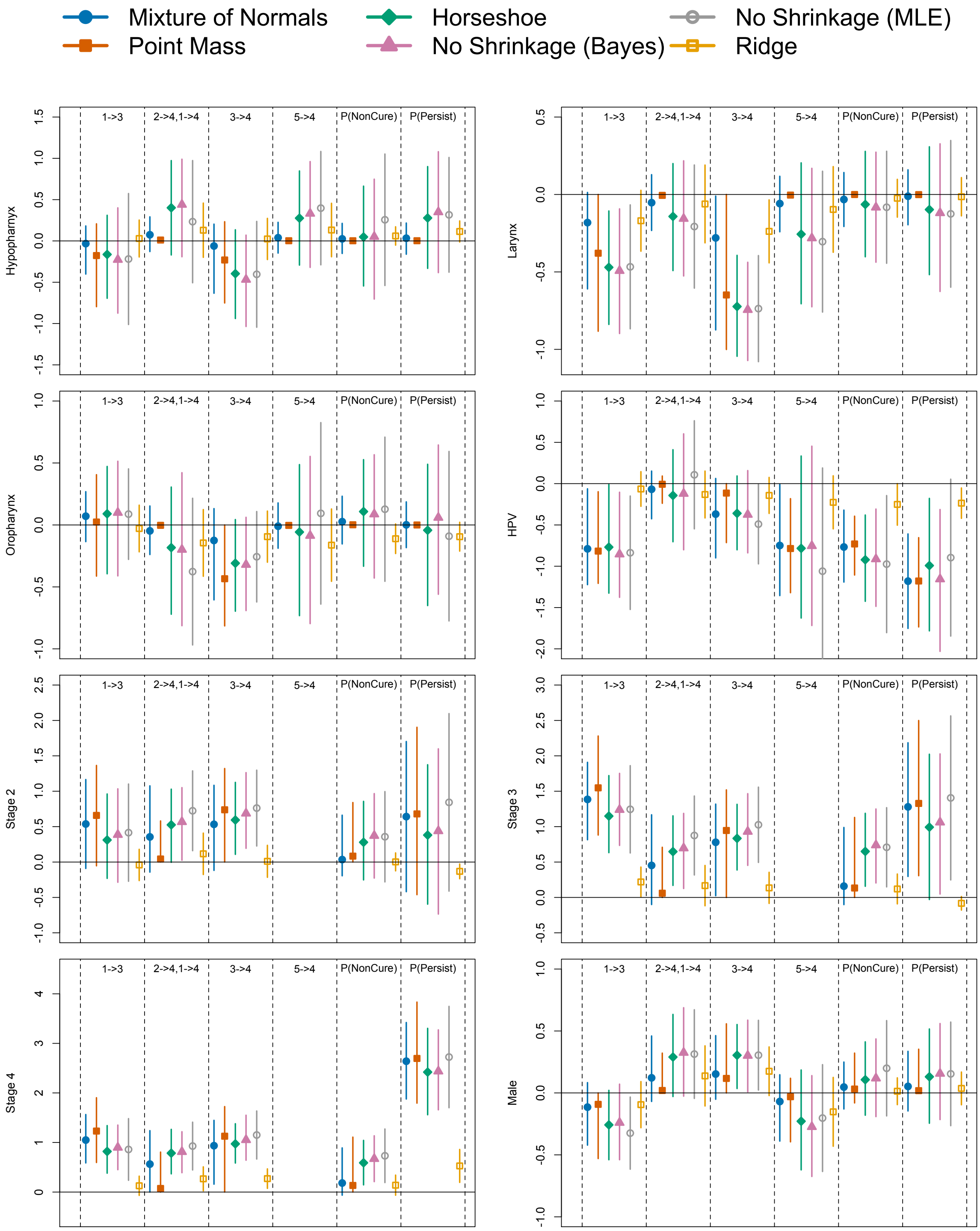
This figure presents heatmaps for the correlation of the γ indicators across MCMC iterations *within* submodels. If γ was always 0 or 1, that group was not included in the plot.

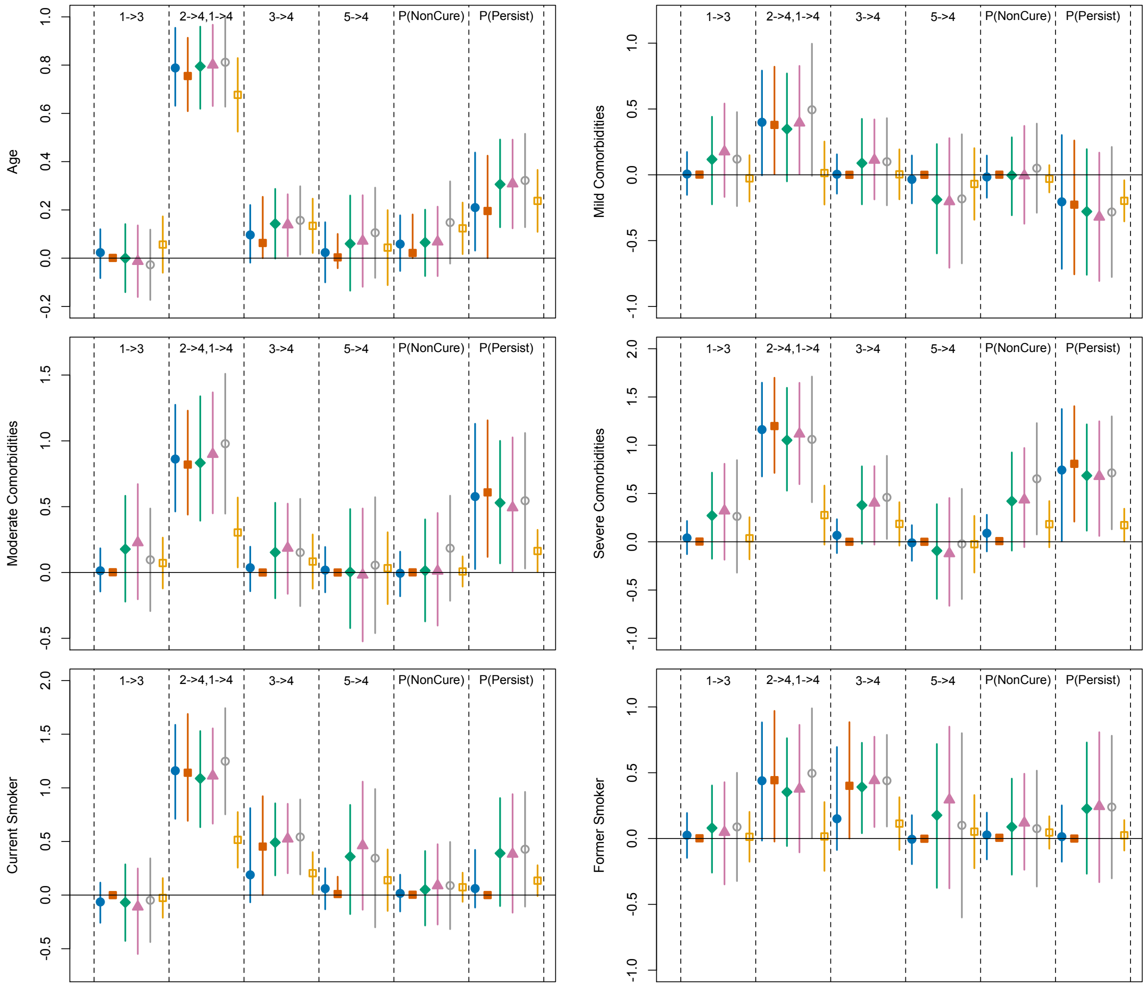
Figure K.9: Correlations of Gamma across Submodels (Point Mass at Zero)



This figure presents heatmaps for the correlation of the γ indicators across MCMC iterations *across* submodels. If γ was always 0 or 1 for one of the variables, the correlation is listed as '-'

Figure K.10: Credible and Confidence Intervals Across All Methods




Age
Mild Comorbidities
Moderate Comorbidities
Severe Comorbidities
Current Smoker
Former Smoker
1->3
2->4,1->4
3->4
5->4
P(NonCure)
P(Persist)
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

Appendix L

Additional Simulation Results for Order-Restricted Selection with Interactions

In this appendix, we present some additional results for the simulations in **Chapter VI**. We perform simulations under both linear and logistic regression.

L.1 Linear Regression

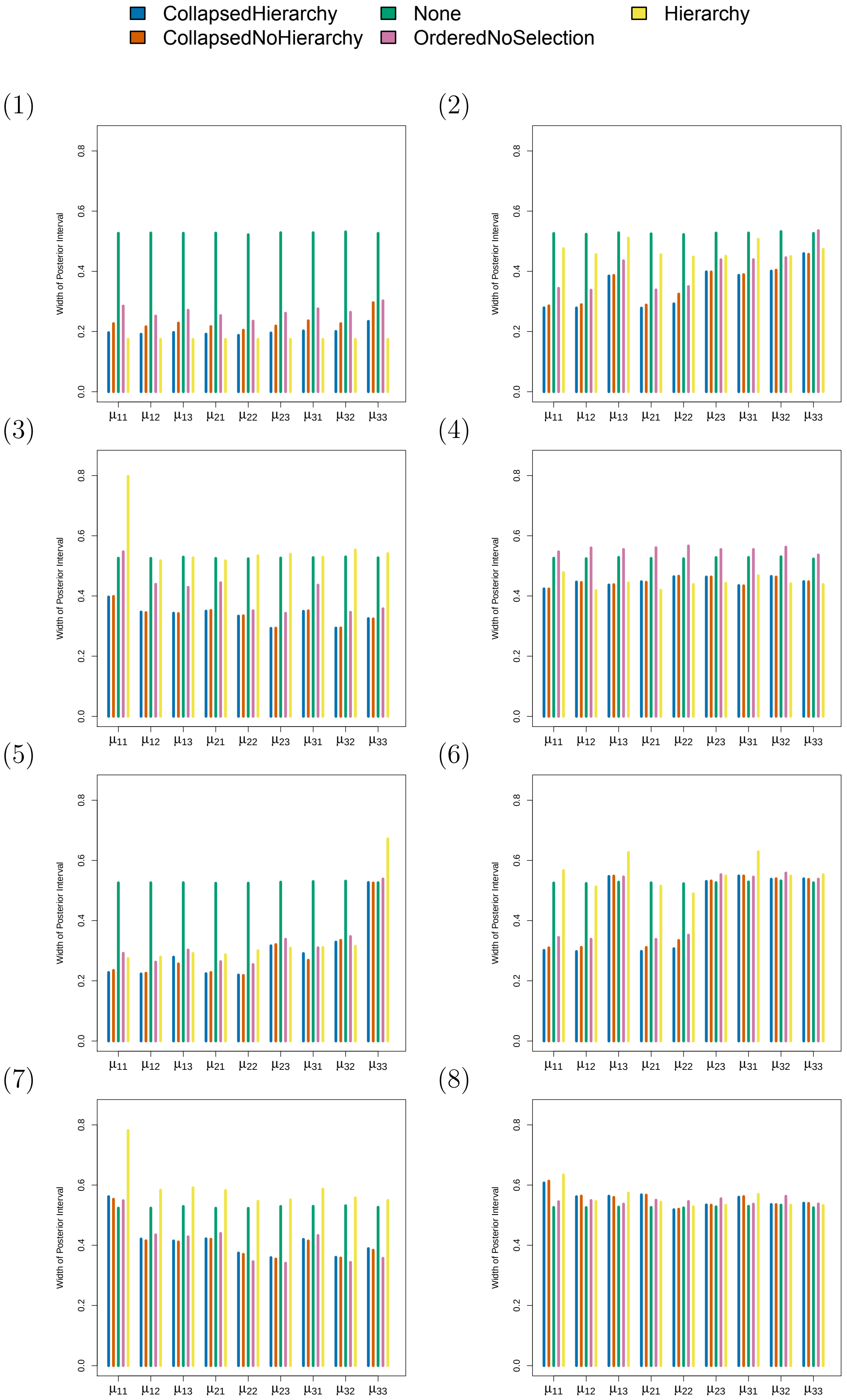
Figure L.1 shows the average 95% credible interval widths for μ. We can see that the collapsed methods generally result in the smallest posterior intervals. When we have strict ordering and all interactions included, the *None* prior (which has no selection or constraints) has the narrowest posterior intervals. The *Hierarchy* prior outperforms the other methods in terms of credible interval width for Simulation 1, where there are no main effects or interactions.

Figure L.2 shows the coverage rates for the 95% credible intervals for μ. The *Ordered-NoSelection* method results in good coverage properties when strict ordering is present (Simulations 4 and 8), but it results in undercoverage in other simulation settings. In Simulation 5 (which violates weak hierarchy constraints), the *Hierarchy* prior produces undercoverage. In contrast, the *CollapsedHierarchy* is able to obtain good coverage properties in spite of the violation to the hierarchy constraint. The collapsed methods generally produce good coverage properties except in Simulations 7 and 8. In these simulations, we have all interaction terms present and some or all main effects present. Undercoverage seen for the collapsed methods is a result of bias due to the shrinkage.

Figure L.3 shows the average posterior credible intervals for β. The horizontal black bars correspond to the true values. We notice that for many of the simulation settings, the *OrderedNoSelection* prior results in biased parameter estimates. In contrast, the collapsed methods generally produce low bias and narrow intervals. Some hints of bias can be seen in Simulations 5, 7, and 8 for the collapsed methods, but these biases appear very small. The intervals for the *None*, *OrderedNoSelection* and *Hierarchy* priors tend to be large in comparison to the collapsed methods. **Figure L.4** shows the average posterior credible intervals for μ. The results are similar.

Figure L.5 shows the posterior mean of μ for each one of the methods along with the true values through heatmaps. This figure allows us to more clearly see the impact of the various priors on the resulting parameter orderings and magnitudes. Each three-by-three grouping corresponds to the posterior means of μ for one of the methods. The color in a particular cell corresponds to the magnitude of the posterior mean. The *OrderedNoSelection* prior tends to result in more spread out μ estimates compared to the other priors. The two collapsed methods produce very similar results. The *None* method also performs similarly. The *Hierarchy* prior performs similarly except for Simulation 5.

Figure L.1: Average Width of 95% Credible Intervals for μ (Linear Regression)



This figure shows the width of the 95% credible intervals, averaged across 200 simulations.

Figure L.2: Coverage of Credible Intervals for μ (Linear Regression)

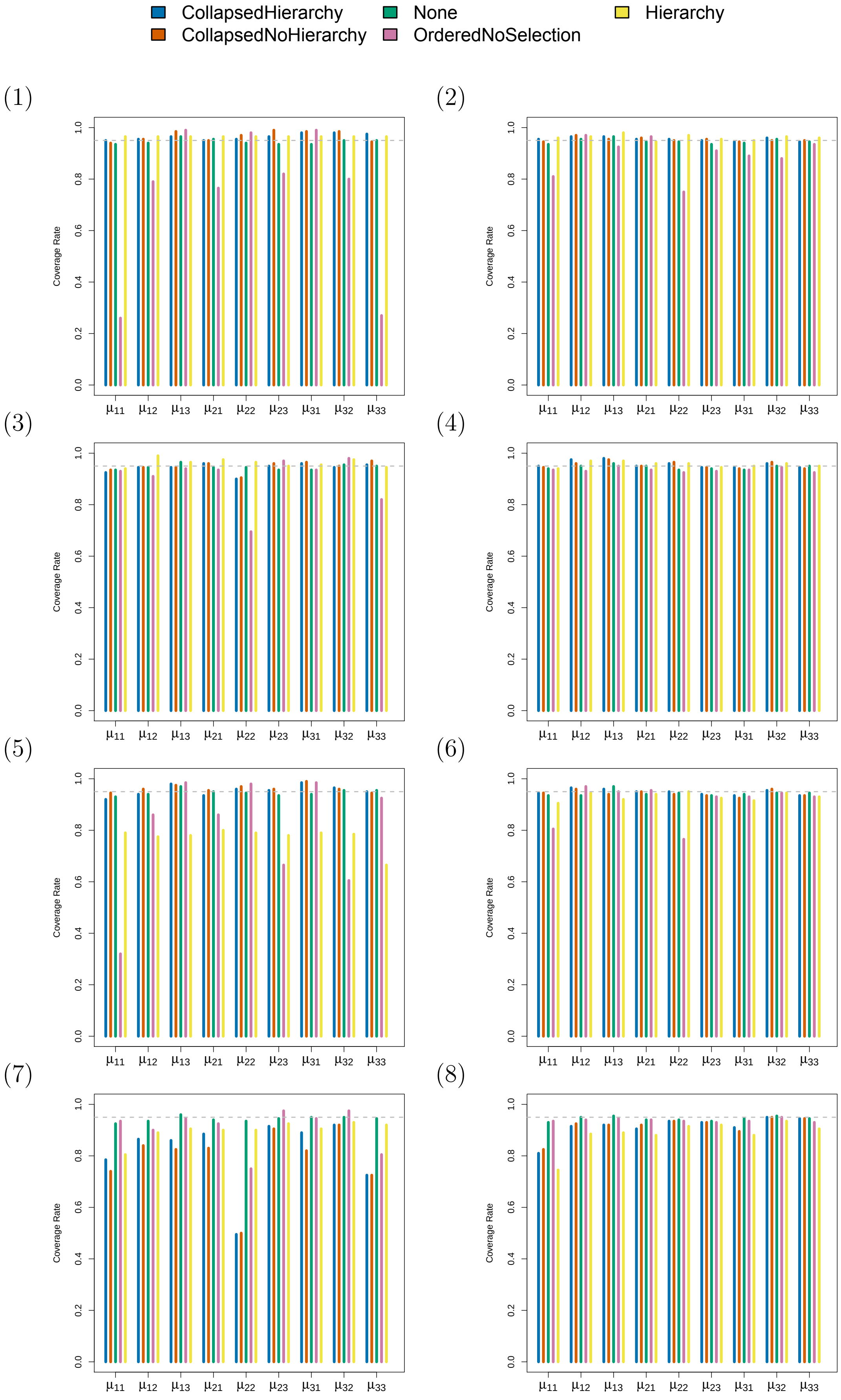
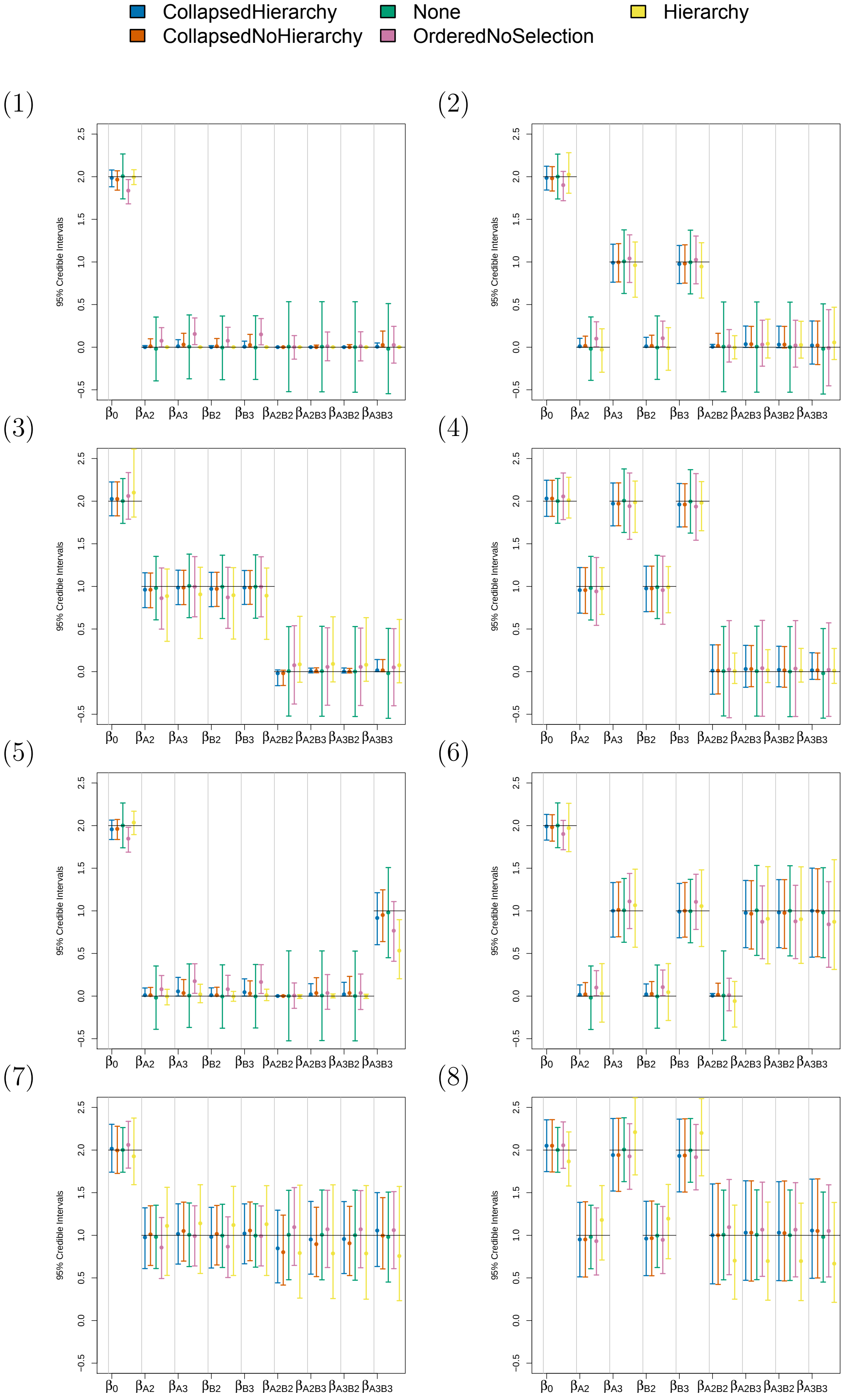
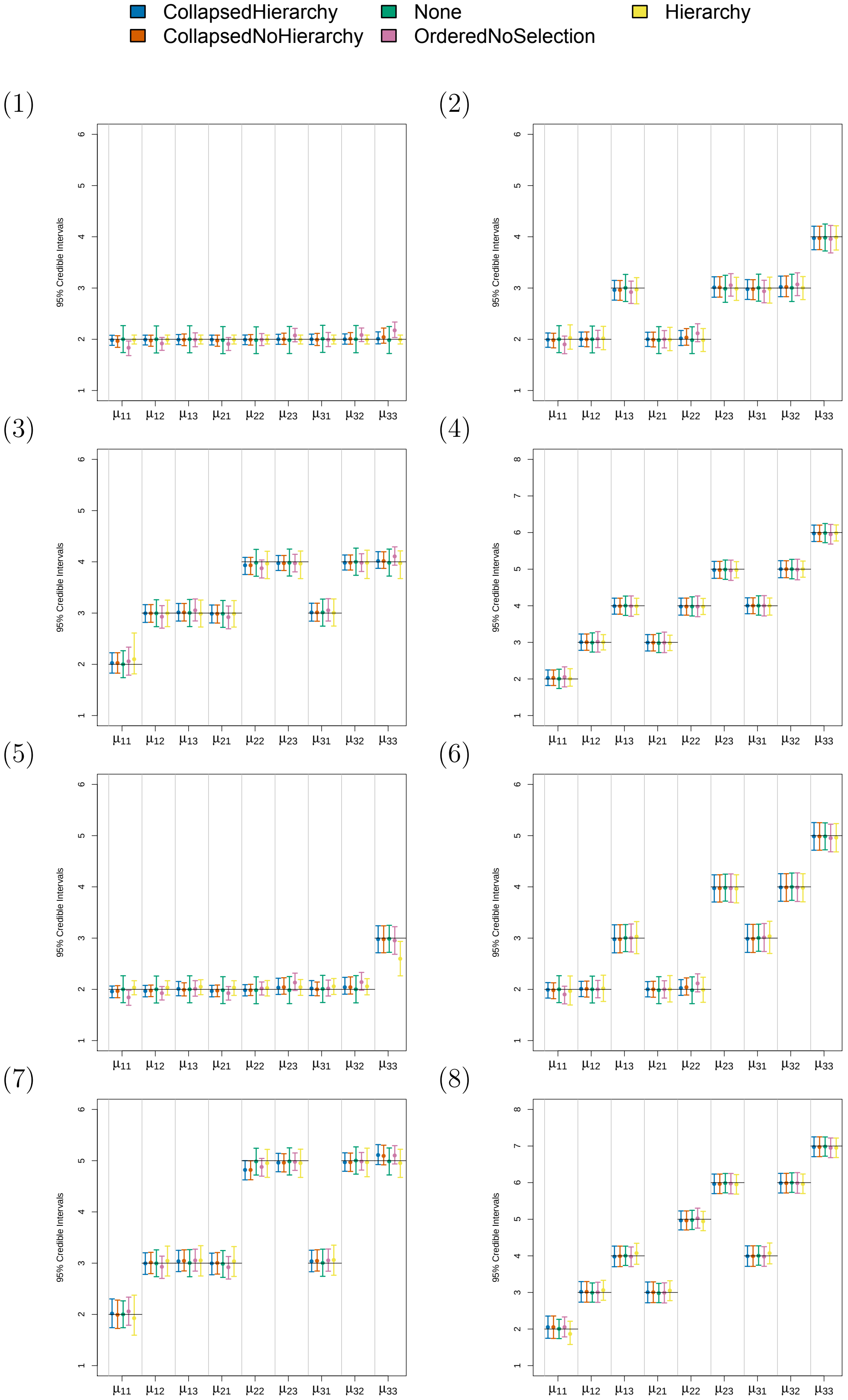


Figure L.3: Average Credible Intervals for β (Linear Regression)



This figure shows the posterior mean, 97.5% quantile, and 2.5% quantile of the posterior draws of β, averaged across the 200 simulations. The horizontal black bars correspond to the true values.

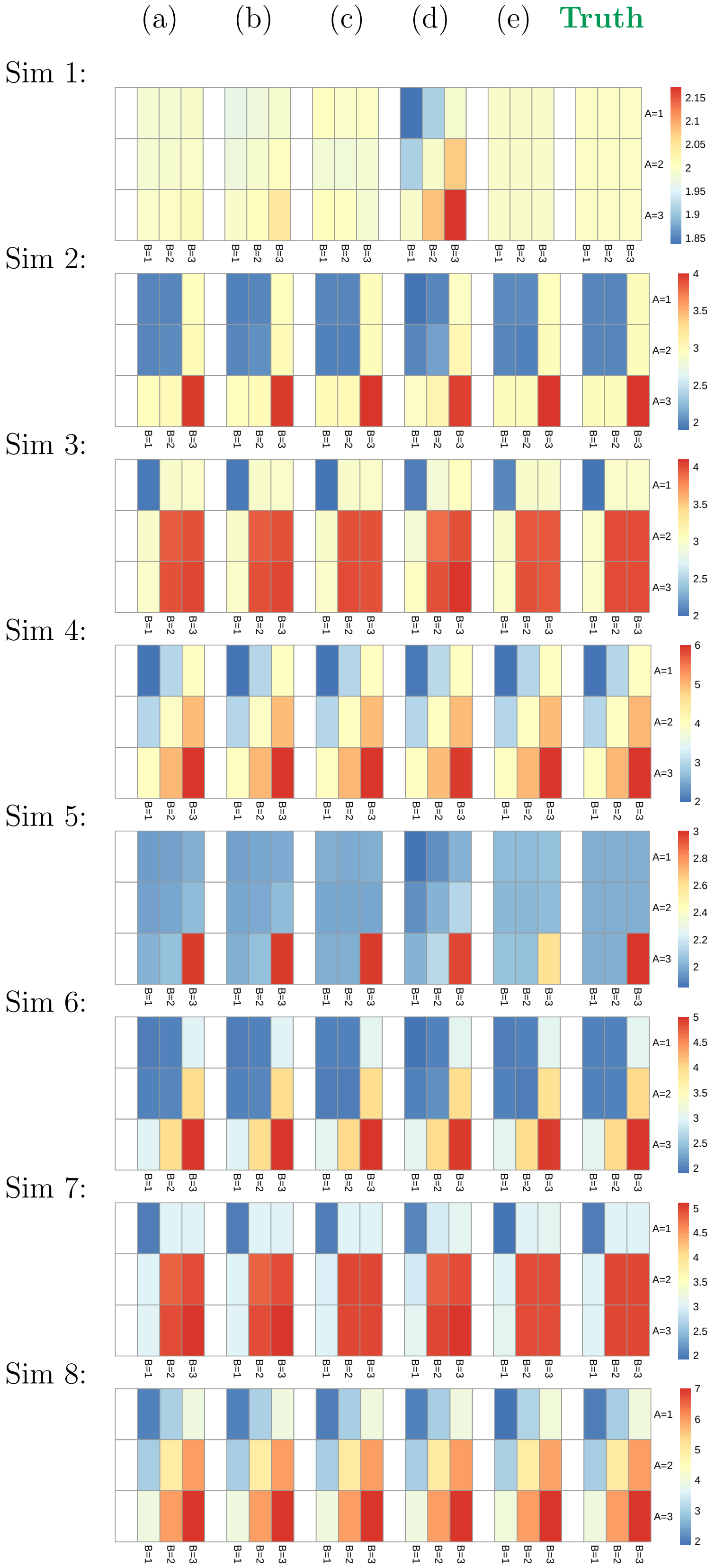
Figure L.4: Average Credible Intervals for μ (Linear Regression)



This figure shows the posterior mean, 97.5% quantile, and 2.5% quantile of the posterior draws of μ, averaged across the 200 simulations. The horizontal black bars correspond to the true values.

Figure L.5: Heat Maps for Posterior Mean of μ (Linear Regression)

(a) CollapsedHierarchy (b) CollapsedNoHierarchy (c) None (d) OrderedNoSelection (e) Hierarchy



This figure shows the posterior mean values of μ for different combinations of A and B, averaged across 200 simulations. The final column shows the true values.

L.2 Logistic Regression

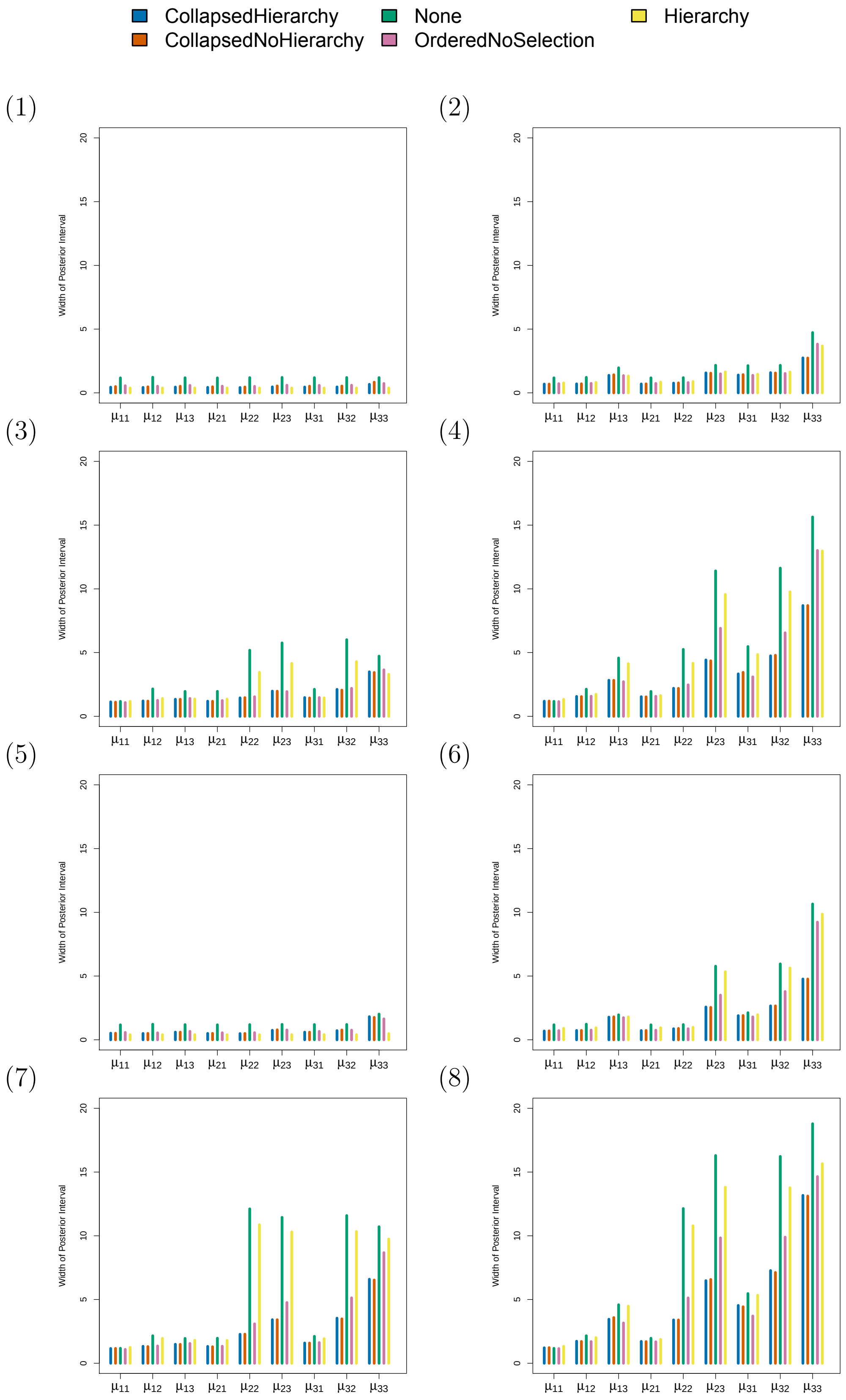
Figure L.6 shows the average 95% credible interval widths for μ. As in the linear setting, the collapsed methods generally result in the smallest posterior intervals. The *None* prior generally produces the largest intervals. The *Hierarchy* prior often produces inflated intervals for parameters corresponding to interaction terms.

Figure L.7 shows the coverage rates for the 95% credible intervals for μ. We see more evidence of some undercoverage for the collapsed methods than in the linear case. As before, this may be due to bias induced by the shrinkage. We note that the *Hierarchy* prior tends to produce some undercoverage in all the simulation settings. The *None* prior produces good coverage in all simulation settings.

Figures L.8 and L.9 show the average posterior credible intervals for β and μ respectively. One striking difference between the logistic regression and linear settings is the width of the credible intervals, which are much larger in the logistic regression case. However, this is to be expected. All methods generally produce reasonable estimates for the main effect parameters. The credible intervals for the interaction parameters tend to be much narrower for the collapsed methods, which helps to control some of the large variability and stabilize the parameter estimates.

Figure L.10 shows the posterior mean of μ for each one of the methods along with the true values through heatmaps. We generally see greater differences across the methods than we did in the linear regression case. As before, the *OrderedNoSelection* prior tends to result in estimates of μ that are more "spread out." The collapsed methods tend to perform similarly, although there is more spread for the *CollapsedNoHierarchy* method in Simulation 1 than there should be. We can see greater differences between the collapsed methods and the *Hierarchy* prior than in the linear regression case.

Figure L.6: Average Width of Credible Intervals for μ (Logistic Regression)



This figure shows the width of the 95% credible intervals, averaged across 200 simulations.

Figure L.7: Coverage of Credible Intervals for μ (Logistic Regression)

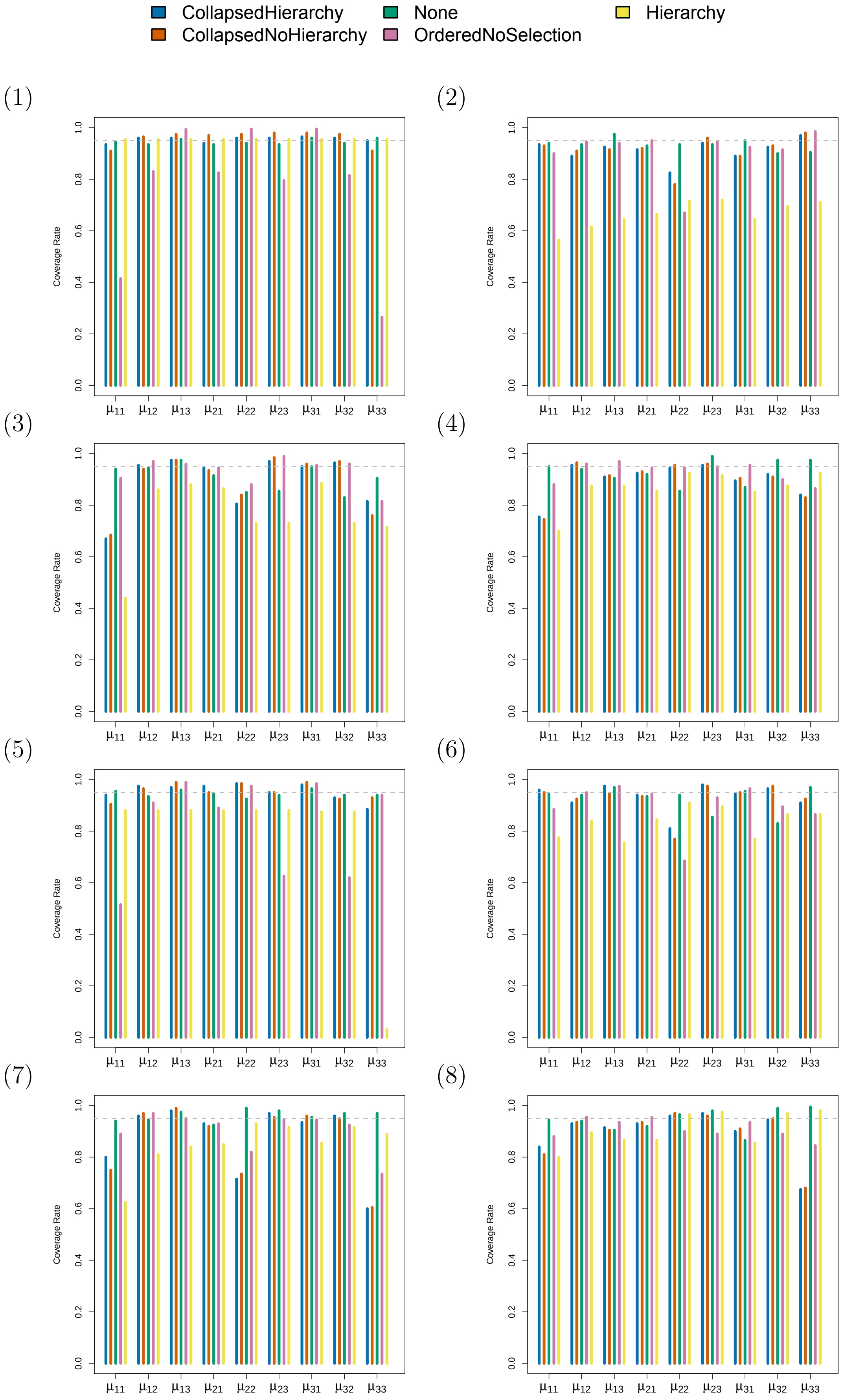
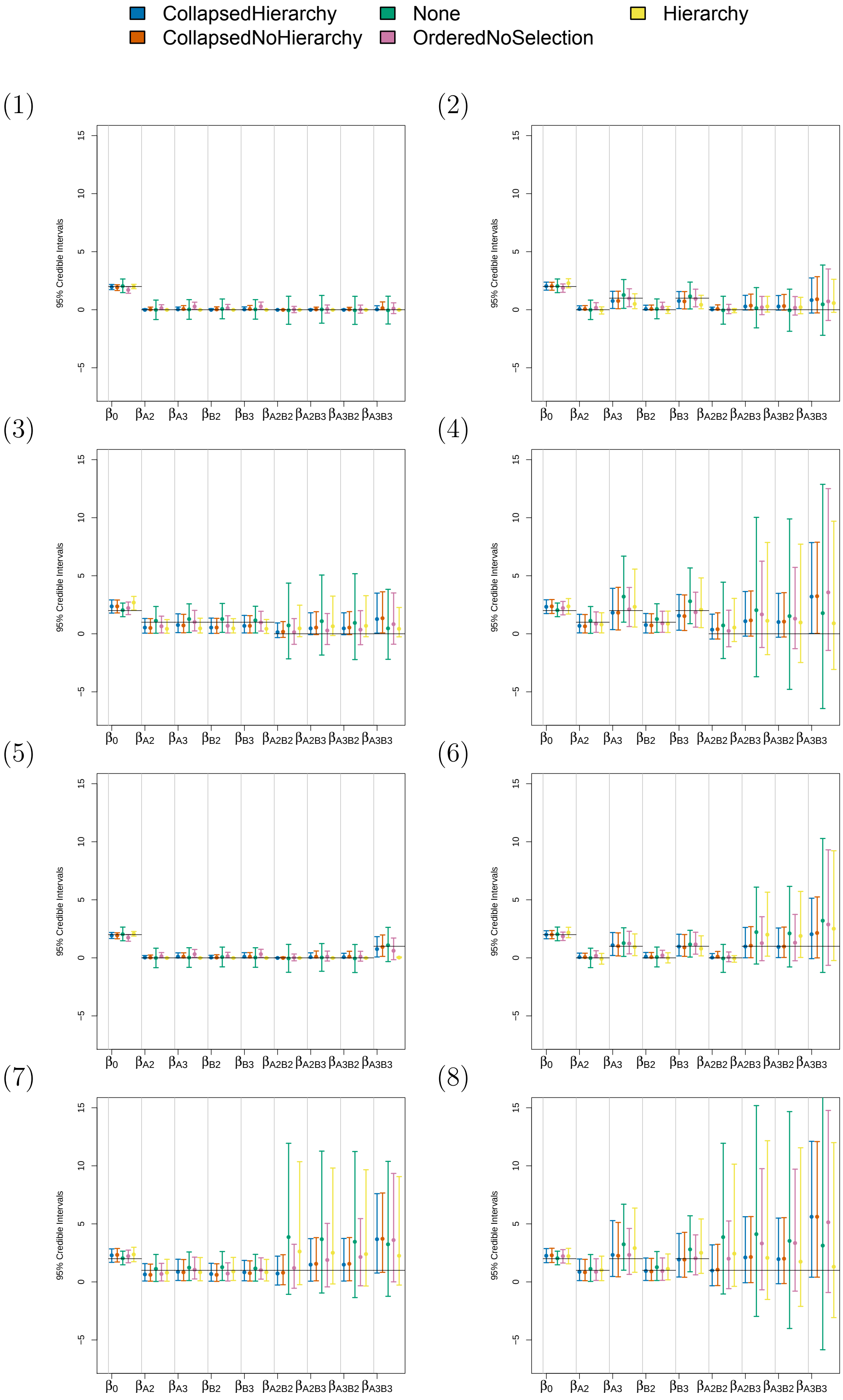
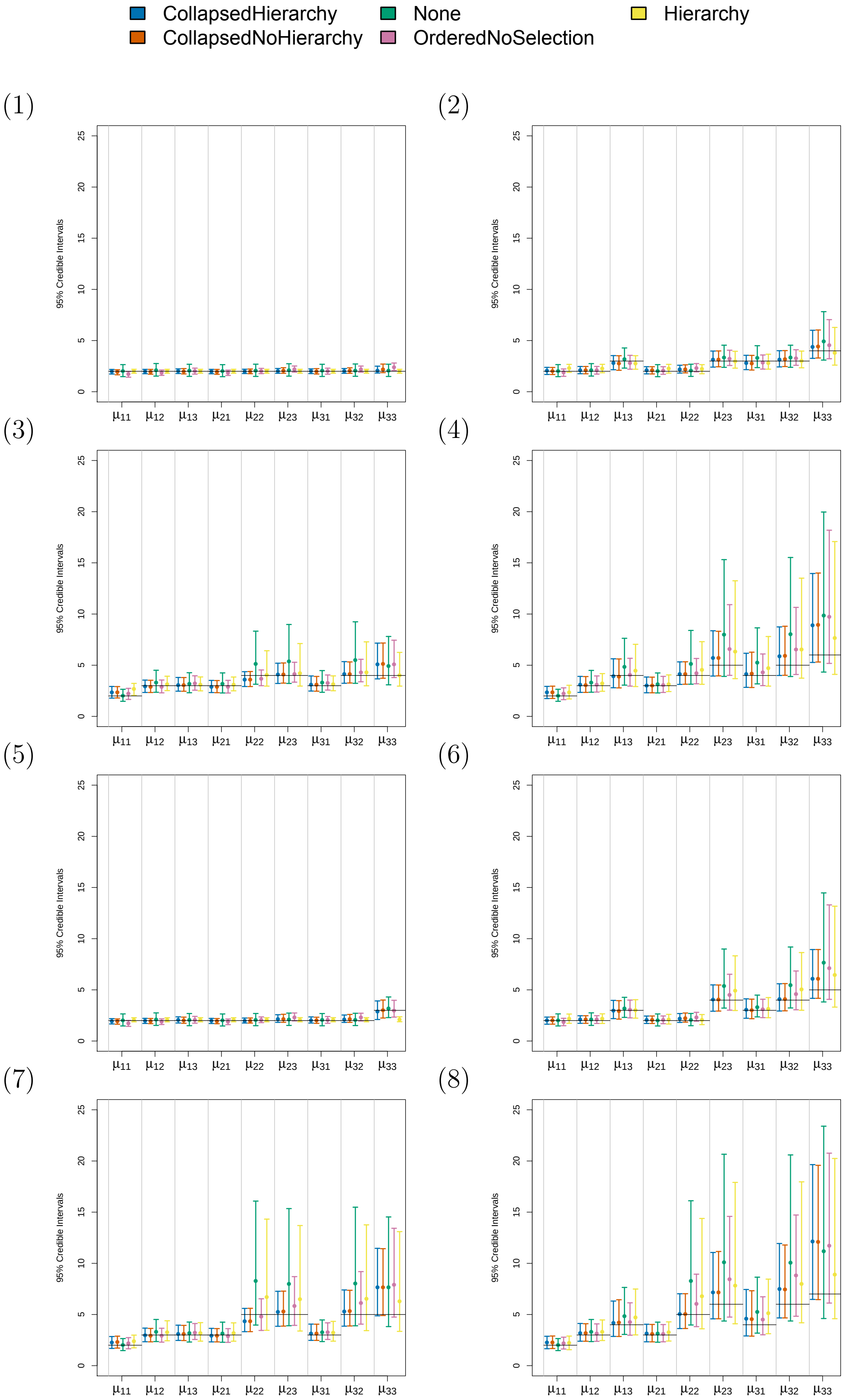


Figure L.8: Average Credible Intervals for β (Logistic Regression)



This figure shows the posterior mean, 97.5% quantile, and 2.5% quantile of the posterior draws of β, averaged across the 200 simulations. The horizontal black bars correspond to the true values.

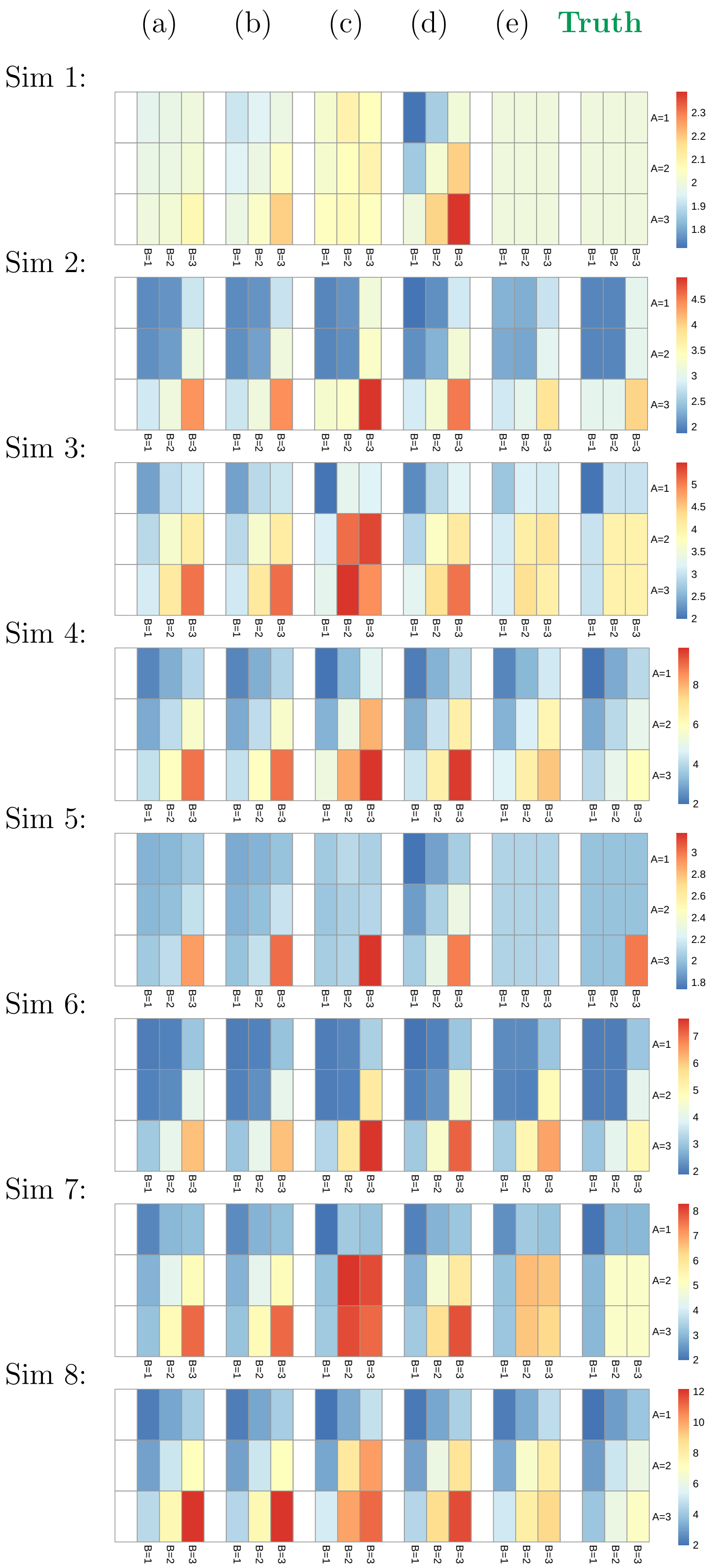
Figure L.9: Average Credible Intervals for μ (Logistic Regression)



This figure shows the posterior mean, 97.5% quantile, and 2.5% quantile of the posterior draws of μ, averaged across the 200 simulations. The horizontal black bars correspond to the true values.

Figure L.10: Heat Maps for Posterior Mean of μ (Logistic Regression)

(a) CollapsedHierarchy (b) CollapsedNoHierarchy (c) None (d) OrderedNoSelection (e) Hierarchy



This figure shows the posterior mean values of μ for different combinations of A and B, averaged across 200 simulations. The final column shows the true values.

Bibliography

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R. (2014). Multiple Imputation of Covariates by Fully Conditional Specification: Accomodating the Substantive Model. *Statistical Methods in Medical Research*, 24(4):462–487.

Beesley, L. J., Bartlett, J. W., Wolf, G. T., and Taylor, J. M. G. (2016). Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Statistics in Medicine*, 35(26):4701–4717.

Bijwaard, G. E. (2014). Multistate event history analysis with frailty. *Demographic Research*, 30(58):1591–1620.

Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? *Structural equation modeling : a multidisciplinary journal*, 15(4):651–675.

Breslow, N. E. (1972). Contribution to the discussion of the paper by DR Cox. *Journal of the Royal Statistical Society*, 24:216–217.

Brooks, S. P. B. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An R-package for Estimating Semiparametric Mixture Cure Models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling Sparsity via the Horseshoe. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chen, M.-H. and Ibrahim, J. G. (2002). Bayesian Methods for Missing Covariates in Cure Rate Models. *Lifetime Data Analysis*, 8(2):117–146.

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, 24(1):17–36.

Chipman, H., George, E. I., and McCulloch, R. E. (2001). The Practical Implementation of Bayesian Model Selection. *Model Selection*, pages 65–116.

Cho, M., Schenker, N., Taylor, J. M. G., and Zhuang, D. (2001). Survival Analysis with

Long-Term Survivors and Partially Observed Covariates. *The Canadian Journal of Statistics*, 29(3):421–436.

Chung, H., Flaherty, B. P., and Schafer, J. L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society*, 169(4):723–743.

Cognetti, D. M., Weber, R. S., and Lai, S. Y. (2008). Head and Neck Cancer: An Evolving Treatment Paradigm. *Cancer*, 113(70):1911–1932.

Conlon, A. S. C., Taylor, J. M. G., and Sargent, D. J. (2013). Multi-state Models for Colon Cancer Recurrence and Death with a Cured Fraction. *Statistics in Medicine*, 33(10):1750–1766.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2):187–220.

de Castro, M., Chen, M.-h., and Zhang, Y. (2015). Bayesian Path Specific Frailty Models for Multi-State Survival Data with Applications. *Biometrics*, 71(3):760–771.

de Wreede, L. C., Fiocco, M., and Putter, H. (2011). mstate : An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, 38(7):1–30.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Duffy, S., Taylor, J. M. G., Terrell, J., Islam, M., Yuan, Z., Fowler, K., Wolf, G., and Teknos, T. (2008). IL-6 predicts recurrence among head and neck cancer patients. *Cancer*, 113(4):750–757.

Dunson, D. B. and Neelon, B. (2003). Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models. *Biometrics*, 59(2):286–295.

Farcomeni, A. (2010). Bayesian Constrained Variable Selection. *Statistica Sinica*, 20(3):1043–1062.

Farewell, V. T. (1982). The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, 38(4):1041–1046.

Follmann, D. and Wu, M. C. (1995). An Approximate Generalized Linear Model with Random Effects for Informative Missing. *Biometrics*, 51(1):151–168.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379.

Frydman, H. and Kadam, A. (2004). Estimation in the continuous time mover – stayer model with an application to bond ratings migration. *Applied Stochastic Models in*

Business and Industry, 20(2):155–170.

Gelfand, A. E., Smith, A. F. M., and Lee, T.-m. (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, 87(418):523–532.

Gelman, A. (2004). Parameterization and Bayesian Modeling. *Journal of the American Statistical Association*, 99(466):537–545.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation using Multiple Sequences. *Statistical Science*, 7(4):457–511.

George, E. I. (2010). Dilution priors: Compensating for model space redundancy. *IMS Collections*, 6(1):158–165.

George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7(2):339–373.

Giusti, C. and Little, R. J. A. (2011). An Analysis of Nonignorable Nonresponse to Income in a Survey with a Rotating Panel Design. *Journal of Official Statistics*, 27(2):211–229.

Grau, J. J., Cuchi, A., Traserra, J., Arias, C., Blanch, J. L., and Estapé, J. (1997). Follow-Up Study in Head and Neck Cancer : Cure Rate according to Tumor Location and Stage. *Oncology*, 54(1):38–42.

Green, P. J. (1995). Reversible Jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Harel, O. (2003). *Strategies for Data Analysis with Two Types of Missing Values*. PhD thesis, Pennsylvania State University.

Harel, O. and Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50.

Hastie, D. I. and Green, P. J. (2012). Model Choice using Reversible Jump Markov Chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, 5(4):475–492.

Hoerl, A. U. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model

Averaging: A Tutorial. *Statistical Science*, 14(4):382–401.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. C. (2014). Joint modeling rationale for chained equations. *BMC Medical Research Methodology*, 14(28):1–10.

Ishwaran, H. and Rao, J. S. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, 33(2):730–773.

Jackson, C. H. (2011). Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, 38(8):1–28.

Jung, H. (2007). *A Latent-Class Selection Model for Nonignorable Missing Data*. PhD thesis, Pennsylvania State University.

Kasim, A., Shkedy, Z., and Kato, B. S. (2012). Estimation and Inference Under Simple Order Restrictions : Hierarchical Bayesian Approach. In *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*, chapter 13, pages 193–214. Springer.

Kuk, A. Y. C. and Chen, C.-H. (1992). A Mixture Model Combining Logistic Regression with Proportional Hazards Regression. *Biometrika*, 79(3):531–541.

Kuo, L. and Mallick, B. K. (1998). Variable Selection for Regression Models. *The Indian Journal of Statistics, Series B*, 60(1):65–81.

Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(8):1–18.

Little, R. J. (2009). Selection and pattern-mixture models. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, chapter 18, pages 409–431. Taylor & Francis Group.

Little, R. J., Ibrahim, J. G., and Molenberghs, G. (2009). Comments on : Missing data methods in longitudinal studies : a review. *Test*, 18(1):47–50.

Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc, Hoboken, NJ, 2nd edition.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. (2013). On the stationary distribution of iterative imputation. *Biometrika*, 101(1):155–173.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*, 44(2):226–233.

Lu, W. and Ying, Z. (2004). Semiparametric Transformation Cure Models. *Biometrika*, 91(2):331–343.

Lu, Z. L., Zhang, Z., and Lubke, G. (2011). Bayesian Inference for Growth Mixture Models with Latent Class Dependent Missing Data. *Multivariate Behavioral Research*, 46(4):567–597.

McCulloch, C. E., Neuhaus, J. M., and Olin, R. L. (2016). Biased and Unbiased Estimation in Longitudinal Studies with Informative Visit Processes. *Biometrics*, 72(4):1315–1324.

Meira-Machado, L., Uña-álvarez, J. D., Cadarso-suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, 18(2):195–222.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538–573.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Miao, W., Ding, P., and Geng, Z. (2016). Identifiability of Normal and Normal Mixture Models with Nonignorable Missing Data. *Journal of the American Statistical Association*, 111(516):1673–1683.

Molenberghs, G., Beunckens, C., and Sotto, C. (2008). Every Missing Not at Random Model Has Got a Missing at Random Counterpart with Equal Fit. *Journal of the Royal Statistical Society (Series B)*, 70(2):371–388.

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., and Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59(10):1092–1101.

Neath, R. C. (2012). On Convergence Properties of the Monte Carlo EM Algorithm. *Institute of Mathematical Statistics*, 10(1):1–21.

Newcombe, P. J., Ali, H. R., Blows, F. M., Provenzano, E., Pharoah, P. D., Caldas, C., and Richardson, S. (2017). Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical Methods in Medical Research*, 26(1):414–436.

Otava, M., Shkedy, Z., Hothorn, L. A., Talloen, W., Kasim, A., Otava, M., Shkedy, Z., Hothorn, L. A., Talloen, W., Gerhard, D., and Kasim, A. (2017). Identification of the minimum effective dose for normally distributed data using a Bayesian variable selection approach. *Journal of Biopharmaceutical Statistics*, 27(6):1–16.

Otava, M., Shkedy, Z., Lin, D., Göhlmann, H. W. H., Bijnens, L., Talloen, W., Kasim, A., Otava, M., Shkedy, Z., Lin, D., Göhlmann, H. W. H., and Bijnens, L. (2014). Dose – Response Modeling Under Simple Order Restrictions Using Bayesian Variable Selection Methods. *Statistics in Biopharmaceutical Research*, 6(3):252–262.

Peng, Y. and Dear, K. B. G. (2000). A Nonparametric Mixture Model for Cure Rate

Estimation. *Biometrics*, 56(1):237–243.

Peterson, L. A., Bellile, E. L., Wolf, G. T., Virani, S., Shuman, A. G., and Taylor, J. M. G. (2016). Cigarette use, comorbidities, and prognosis in a prospective head and neck squamous cell carcinoma population. *Head and Neck*, 38(12):1810–1820.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics : Competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Raghunathan, T. E. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27(1):85–95.

Rathouz, P. J. (2007). Identifiability Assumptions for Missing Covariates in Failure Time Regression Models. *Biostatistics*, 8(2):345–356.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edition.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Inc, New York, NY, 1st edition.

Schafer, J. L. (1997). Imputation of missing covariates under a multivariate linear mixed model. Technical report, Pennsylvania State University.

Schafer, J. L. (1999). Multiple imputation : a primer. *Statistical Methods in Medical Research*, 8(1):3–15.

Schafer, J. L. and Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *Journal of Computational and Graphical Statistics*, 11(2):437–457.

Sherlock, C., Fearnhead, P., and Roberts, G. O. (2010). The random walk Metropolis: linking theory and practice through a case study. *Statistical Science*, 25(2):170–190.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Taylor, J. M. G. (1995). Semiparametric Estimation in Failure Time Mixture Models. *Biometrics*, 51(3):899–907.

Taylor, J. M. G., Murray, S., and Hsu, C.-H. (2002). Survival estimation and testing via multiple imputation. *Statistics and Probability Letters*, 58(3):221–225.

Troughton, P. T. and Godsill, S. J. (1997). A Reversible Jump Sampler for Autoregressive Time Series, Employing Full Conditionals to Achieve Efficient Model Space Moves. Technical report, University of Cambridge.

Van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical Methods in Medical Research*, 16(3):219–242.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, 18(6):681–694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations. *Journal of Statistical Software*, 45(3):1–67.

Vidotto, D., Vermunt, J. K., and Kaptein, M. C. (2015). Multiple Imputation of Missing Categorical Data using Latent Class Models : State of the Art. *Psychological Test and Assessment Modeling*, 57(4):542–576.

Virani, S., Bellile, E., Bradford, C. R., Carey, T. E., Chepeha, D. B., Colacino, J. A., Helman, J. I., McHugh, J. B., Peterson, L. A., Sartor, M. A., Taylor, J. M. G., Walline, H. M., Wolf, G. T., and Rozek, L. S. (2015). NDN and CD1A are novel prognostic methylation markers in patients with nead and neck squamous cell carcinomas. *BMC Cancer*, 15(825):1–13.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Aug mentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

White, I. R. and Royston, P. (2009). Imputing Missing Covariate Values for the Cox Model. *Statistics in Medicine*, 28(15):1982–1998.

White, I. R. and Royston, P. (2011). Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, 30(4):377–399.

Wu, M. C. and Carroll, R. J. (1988). Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics*, 44(1):175–188.

Yamaguchi, K. (1992). Accelerated Failure-Time Regression Models with a regression model of survival fraction: an application to the analysis of “permanent employment” in japan. *Journal of the American Statistical Association*, 87(418):284–292.

Yang, X., Lu, J., and Shoptaw, S. (2008). Imputation-based Strategies for Clinical Trial Longitudinal Data with nonignorable missing values. *Statistics in Medicine*, 27(23):2826–2849.

Zhang, N. and Little, R. J. (2011). A Pseudo-Bayesian Shrinkage Approach to Regression with Missing Covariates. *Biometrics*, 68(3):933–942.

Zhuang, D., Schenker, N., Taylor, J. M. G., Mosseri, V., and Dubray, B. (2000). Analysing the Effects of Anaemia on Local Recurrence of Head and Neck Cancer when Covariate Values are Missing. *Statistics in Medicine*, 19(9):1237–1249.