Approaches to Handling Time-varying Covariates in Survival Models

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Nicholas J. Salkowski

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Melanie M. Wall

May, 2011

Acknowledgements

I am grateful to Drs. Michelle Rheault, Clifford Kashtan and Yoav Segal at the University of Minnesota for making their experimental data available for analysis. I also wish to thank Melanie M. Wall for her significant contributions to this work.

Dedication

Dedicated to my wife Erin, whose support and patience made this work possible.

Abstract

Time-varying covariates present special problems in survival analyses. Their measurements are often missing, and their missing status may be related to the survival outcome of interest. This dissertation discusses three approaches to handling time-varying covariates in survival models. First, predictions of event probabilities from a joint model for longitudinal and event time data are are compared to predictions from simpler models. Second, a Bayesian joint modeling approach is used to resolve difficulties relating to inference when measurements of a potentially mediating process are partially missing. Third, many time-varying covariates can be converted into alternative time scales. This dissertation presents an approach to handle vector-valued time scales in semiparametric proportional hazards regression.

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Tables vi

List of Figures vii

1 Introduction 1

2 Evaluating predictions of event probabilities from a joint model for longitudinal and event time data 4

2.1 Introduction 4

2.2 The Models 7

2.2.1 The Two-Step Model 7

2.2.2 The Joint Model 9

2.2.3 Interpreting parameters 10

2.3 Predictions and Prediction Errors 10

2.4 Results 12

2.4.1 Parameter Estimates in the Two-Step Model 12

2.4.2 Parameter Estimates for the Joint Model 13

2.4.3 Fitted Trajectories and Hazards from the Joint Model 14

2.4.4 Prediction Error 15

2.5 Simulation 16
2.6 Discussion 18

3 Inference on a Partially Missing Mediator **31**
3.1 Introduction 31
3.2 Models 33
3.2.1 Modeling X-inactivation and Gene Expression 34
3.2.2 Modeling Disease Severity 35
3.2.3 Joint Model and Prior Specification 36
3.3 Results 38
3.4 Discussion 41

4 Semiparametric Proportional Hazards Regression with Vector-Valued Time Scales **50**
4.1 Introduction 50
4.2 Vector-valued Time Scales and Associated Risk Sets 52
4.3 Example 56
4.4 Simulations 57
4.5 Discussion 59

5 Conclusions **67**

References **69**

List of Tables

2.1 Parameter Estimates: Two-Step Model 21
2.2 Parameter Estimates: Joint Model 22
2.3 Prediction Error Estimates for the Heart Failure Clinical Trial data. . . 23
2.4 Mean Prediction Error in 100 Simulations. 24
2.5 Mean Prediction Error in 100 High Variance Simulations. 25
3.1 Posterior means and 95% credible intervals for Model 1 and Model 2. . . 44
3.2 Posterior means and 95% credible intervals for Model 3 and Model 3–Constrained. 45
4.1 Primary Biliary Cirrhosis Data Results. 61
4.2 Simulation Results: Baseline hazard function $\lambda_A(t_1, t_2)$. 62
4.3 Simulation Results: Baseline hazard function $\lambda_B(t_1, t_2)$. 63
4.4 Simulation Results: Baseline hazard function $\lambda_1(t_1, t_2)$. 64
4.5 Simulation Results: Baseline hazard function $\lambda_2(t_1, t_2)$. 65

List of Figures

2.1 Example Trajectories 26
2.2 $\hat{\mu}_i$ for participants whose death was observed and censored participants. 27
2.3 Two participants with similar longitudinal observations but different survival data. 28
2.4 Predicted survival probabilities for the clinical trial data set under the joint model and under the two-step model. 29
2.5 Simulation Results. 30
3.1 Diagram of the general modeling framework. 46
3.2 Posterior distributions of X-inactivations from Model 2. 47
3.3 Posterior distributions for α_0, α_G, α_X, and η_G for the four models. . . . 48
3.4 Posterior distributions for ϕ_G, ϕ_X, θ_{1G}, and θ_{1X} for the four models. . . 49
4.1 Risk Sets. 66

Chapter 1

Introduction

Time-varying covariates present special problems in survival analyses. Their measurements are often missing, and their missing status may be related to the survival outcome of interest. Time-varying covariates present a history of measurements. This history, or trajectory, can be modeled in multiple ways. Typically, the most recent measurement is used, but characteristics of the trajectory may sometimes be of greater value as predictors.

Joint models are one approach to handling longitudinal and survival data. In one common class of joint models, the longitudinal model and the survival model are assumed to be conditionally independent given a set of shared latent variables.[1] An advantage of employing these models is the hope that the survival data model can help resolve the problems of informative missingness in the longitudinal model, and the longitudinal model can account for measurement error in the time-varying covariates. In practice, these models present unusual, and often overlooked, difficulties. Both longitudinal and survival data provide information about the latent variables. Individuals who experience events carry more information about the latent variables in their survival data than censored individuals. Individuals with long longitudinal data series carry more information about the latent variables in their longitudinal data than individuals with short longitudinal data series. This can be troublesome, since when events truncate the longitudinal data series, data sets can be filled with censored individuals with long longitudinal data series and individuals with little longitudinal data, but specific event times.

Chapter 2 presents an investigation of the predictive performance of a joint longitudinal and event time model for participants in a heart failure clinical trial. This trial included a time-varying, self-reported, quality of life measurement. Since the measure is self-reported, and somewhat subjective, it may be more reasonable to use aspects of a patient's quality of life trajectory, rather than simply the last available measurement, to predict survival. This chapter focuses on evaluating a particular joint model for the purpose of predicting survival using a partial longitudinal data series. The joint model is compared to a computationally simpler, but potentially biased, two-step model for making predictions of survival probability for individuals who have some longitudinal measurements available, but who have not yet died.

Chapter 3 presents an analysis of experimental data. Two groups of genetically different mice were studied. The researchers wished to see whether X-inactivation, as well as measures of disease severity, were affected by the genetic differences. In particular, they wanted to see whether X-inactivation mediated the effect of the genes on disease severity. Several measurements of mice, including one longitudinal measure, were made, but most measures, including the measures of X-inactivation, were not available for mice who died prior to the end of the experiment. While the study in Chapter 2 focuses on the use of joint models when the survival outcome is of primary interest, the study in Chapter 3 focuses on the use of joint models when the survival outcome is a nuisance. By using a Bayesian joint modeling approach, inference on whether X-inactivation mediates survival can be made. This inference is not possible in conventional survival models, since measures of X-inactivation are missing for all the dead mice.

Another approach to dealing with time-varying covariates in survival models is presented in Chapter 4. Many time-varying covariates can be considered as cumulative measures instead of instantaneous measures, typically by integrating the instantaneous measures over time. Indeed, many cumulative measures are timescales in their own right, that is, they are non-decreasing non-negative functions of time. Thus, it may be useful to consider whether such a time-varying covariate should be treated as a timescale instead. Chapter 4 describes a novel extension of Cox proportional hazards regression to include a vector-valued time scale. By combining multiple time scales into a time scale vector, time-varying covariates can be eliminated from the parametric portion of the

semi-parametric regression model, and handled in the nonparametric baseline hazard instead.

The appeal of this approach is flexibility. First, the vector-valued time scale allows the baseline hazard to be a complicated function of all the included time scales, eliminating the need to specify a functional form for the time-varying covariates. This also eliminates the need for the time-varying covariates to satisfy proportionality assumptions. Second, this approach can eliminate the choice of the appropriate time scale for the analysis. There are often multiple plausible time scales for a survival analysis. Time since randomization is commonly used in clinical studies, but age, time since disease onset, time since diagnosis, or even some measure of disease progression could also be time scales for analysis. The vector-valued timescale approach allows all the potentially important timescales to affect the baseline hazard.

Chapter 2

Evaluating predictions of event probabilities from a joint model for longitudinal and event time data

2.1 Introduction

Many studies involve the simultaneous collection of measurements over time (longitudinal data) and measurements of the amount of time until some event of interest (event time data). Medical studies are especially likely to generate both longitudinal and event time data since participants are often monitored and measured regularly while they are at risk of experiencing some event, such as illness or death. Often, the longitudinal process is measured because it is believed to be a potential predictor or surrogate for the event time process. One approach to incorporating measurements of a longitudinal process into an event time model is to treat the longitudinal measurements as a time-varying covariate. Indeed, Cox proportional hazards models with time-varying covariates are commonly used.

The longitudinal measurements, however, are typically made with some error. Also,

the longitudinal measurements are often made at infrequent intervals. Time-varying covariate models treat the longitudinal process as if it were continuously recorded without error. These models assume that the hazard at a particular time is a function of the current covariate value. This may be too limiting when there is some reason to believe that the trajectory, not the absolute measure, of the longitudinal process is related to hazard. In this chapter, the example longitudinal data are scores from a patient survey measuring the impact of their coronary heart disease on their daily life. It is reasonable to think that otherwise similar individuals may have different perceptions of the impact of disease on their lives, and different disease impact scores. How an individual's disease impact scores change over time may better reflect his or her state of health than the most recent score.

Two-step models, such as the model described by Tsiatis et al.[2], model the longitudinal process, then use the fitted longitudinal model as a predictor for the event time model. Typically, a linear mixed model is used for the longitudinal process, and the smoothed trajectory for each individual replaces the actual longitudinal measurements as a time-varying covariate in the event time model. This approach provides a continuous predictor that does not ignore longitudinal measurement error. Any function of the fitted longitudinal model, however, could be used as a predictor in the event time model. For example, each individual's smoothed trajectory parameters (such as the slope) could be used as predictors. When the trajectory parameters are used as predictors, the entire longitudinal process is summarized into a set of non-time-varying covariates for the event time model.

Joint modeling of the longitudinal and event time data have been proposed to account for association between longitudinal measures and event times (e.g. Wulfsohn and Tsiatis[3], Henderson et al.[1], Guo and Carlin[4], Vonesh et al.[5]). Besides the potential to gain statistical efficiency compared to a two-step approach, the joint model incorporates event times to explain missing longitudinal observations. Hence, an additional benefit of joint modeling can be a reduction in bias for the longitudinal parameter estimates when the longitudinal data are not missing at random.[1] There are many possible joint models for longitudinal and event time data. Tsiatis and Davidian[6] provide an excellent overview of joint models.

While there are theoretical reasons for preferring a joint model to a two-step model,

the joint model may not be universally superior in practice. Joint models are usually more computationally difficult to fit. Joint models are more complex, and it is more difficult to infer the influence of particular data on the fitted model. One strength of joint models is that the event time data is used to predict the longitudinal trajectory for subjects with missing longitudinal data. This chapter will demonstrate that this strength can actually be disadvantageous in application.

This chapter will focus on a particular class of joint models described by Henderson et al.[1] These models use latent variables to model both the longitudinal trajectory function and the survival function. Guo and Carlin[4] describe how to fit these models using PROC NLMIXED in SAS. Since the range of possible applications for these joint models is broad, this chapter will focus on prediction for a small, but clearly useful, subset of applications where: (i) events can occur at most once, (ii) events stop the observation of the longitudinal process, and (iii) prediction of the probability of an event during a given interval of time is of primary interest.

The motivating dataset comes from an international, randomized, placebo-controlled, double-blind, parallel-group trial. Its goal was to evaluate the application of a drug in chronic heart failure patients, but treatment assignments, as well the other available baseline covariates, have been ignored in the current analysis. A primary endpoint of the study was all-cause mortality, and time to death is the event time of interest in the joint model considered in this chapter. A secondary outcome of the study was the self-reported impact of the disease on daily living, longitudinally measured using the Minnesota Living with Heart Failure (MLHF) questionnaire of Rector et al.[7] MLHF scores range from 0, for no impact of heart failure on daily living, to 105 for strong impact of heart failure on daily living. Disease impact on daily living, measured by MLHF scores, will serve as the longitudinal process in the joint model.

Data for 2,102 participants were analyzed. Each participant had MLHF measured during up to 10 scheduled visits over the 24 month study period, for a total of 15,630 MHLF observations. 438 participants died during observation, and 1,664 were censored. It is reasonable to hypothesize that the impact of the disease on daily living is related to survival. Participants with poor or diminishing health could experience greater impact of heart failure on their daily lives and could be at higher risk of death than participants in better or improving health. Also, since daily living impact is a self-perception

measure, how MLHF scores change over time may be more important than the MLHF measurement at any particular time. This also supports using the trajectory of the impact of the disease on daily living to predict survival. The joint model of Henderson et al.[1] provides a structure to model longitudinal trajectory and survival jointly. A quadratic model was considered for daily living impact trajectory over time and the coefficients of the longitudinal trajectory were taken as predictors of survival. Inclusion of other predictors could improve the model, but would unnecessarily complicate the focus of the current demonstration of the joint model and its comparison with other simple modeling strategies.

Section 2.2 will describe the models used to analyze the heart failure data set. Section 2.3 will discuss calculating prediction error and describe the procedures used to make predictions. Section 2.4 will summarize the model parameter and prediction error estimates. Section 2.5 will describe simulation studies to demonstrate circumstances when the joint model predictions are inferior. This section will conclude with an evaluation of the joint model's value in survival prediction.

2.2 The Models

2.2.1 The Two-Step Model

Consider a two-step model for comparison to the joint longitudinal-survival model. The first step of a two-step model is to fit a linear mixed model to the longitudinal data. Let $\boldsymbol{y}_i^T = [y_{i1}\ y_{i2}\ \cdots\ y_{in_i}]$ be the vector of n_i longitudinal observations for participant i. Let $\boldsymbol{x}_i$ be a $(n_i \times r)$ matrix of covariate data for participant i. In particular, $\boldsymbol{x}_i$ includes the measurement times. Also, let $\boldsymbol{u}_i$ be a $(p \times 1)$ vector of unobserved latent variables for the i^{th} participants, and let $\boldsymbol{\theta}_L$ be a vector of longitudinal model parameters. Suppose $\boldsymbol{y}_i$ is a realization from some distribution F_L from a family of distributions indexed by $\boldsymbol{x}_i$, and $\boldsymbol{\theta}_L$:

$$\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\theta}_L \overset{ind}{\sim} F_L(\boldsymbol{x}_i, \boldsymbol{\theta}_L)$$

$$\boldsymbol{u}_i \mid \boldsymbol{\theta}_U \overset{iid}{\sim} G(\boldsymbol{\theta}_U)$$

where $\boldsymbol{\theta}_U$ is a vector of parameters that control the distribution of the latent variables. Maximum likelihood estimation can be performed to obtain $\hat{\boldsymbol{\theta}}_L$ and $\hat{\boldsymbol{\theta}}_U$.

The second step is to model the event time data. Let s_i and c_i be the event time and the censoring time for participant i, respectively. Only $t_i = \min(s_i, c_i)$ and $\delta_i = I(t_i = s_i)$ are observed. Suppose that t_i and δ_i are realizations from some distribution F_S which is in a family of distributions indexed by $\boldsymbol{x}_i$, $\hat{\boldsymbol{u}}_i$, and $\boldsymbol{\theta}_S$:

$$t_i, \delta_i \mid \boldsymbol{x}_i, \hat{\boldsymbol{u}}_i, \boldsymbol{\theta}_S \overset{ind}{\sim} F_S(\boldsymbol{x}_i, \hat{\boldsymbol{u}}_i, \boldsymbol{\theta}_S)$$

where $\boldsymbol{\theta}_S$ is a vector of parameters for the event time model, and $\hat{\boldsymbol{u}}_i$ is the predictor of $\boldsymbol{u}_i$ based on the longitudinal model. Specifically $\hat{\boldsymbol{u}}_i = \mathrm{E}[u_i \mid \boldsymbol{y}_i, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_L, \hat{\boldsymbol{\theta}}_U]$, that is, $\hat{\boldsymbol{u}}_i$ is the empirical Bayes estimate for $\boldsymbol{u}_i$. Note that the distribution of the event time data depends only on the longitudinal data through its functional relationship with $\hat{\boldsymbol{u}}_i$.

For the current study, linear and quadratic longitudinal trajectories were considered, and the quadratic trajectory model was selected based on AIC:

$$y_{ij} \mid x_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}, \sigma^2 \overset{ind}{\sim} N\left(\begin{bmatrix} 1 & x_{ij} & x_{ij}^2 \end{bmatrix} (\boldsymbol{\beta} + \boldsymbol{u}_i), \sigma^2\right)$$

$$\boldsymbol{u}_i = \begin{bmatrix} u_{i0} & u_{i1} & u_{i2} \end{bmatrix}^T \overset{iid}{\sim} N(\boldsymbol{0}, \Sigma)$$

where x_{ij} is the time of the j^{th} measurement of MLHF score for participant i, $\boldsymbol{\beta}^T = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \end{bmatrix}$, and Σ is the covariance matrix for the latent variables. The maximum likelihood estimates $\hat{\boldsymbol{\theta}}_L = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\Sigma})$ are obtained along with the Empirical Bayes estimate $\hat{\boldsymbol{u}}_i$.

A Weibull survival model is used for the second step:

$$\Pr(s_i \geq r | \mu_i, \gamma) = \exp\left\{-\exp\left\{-\mu_i \gamma\right\} r^{\gamma}\right\}$$

$$f(s_i | \mu_i, \gamma) = \gamma \exp\left\{-\mu_i \gamma\right\} s_i^{\gamma - 1} \exp\left\{-\exp\left\{-\mu_i \gamma\right\} s_i^{\gamma}\right\}$$

$$\text{where } \mu_i = \beta_S + \boldsymbol{\alpha}^T (\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_i)$$

such that γ, β_S and $\boldsymbol{\alpha}$ determine the relationship between the fitted summary of the longitudinal data, $\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_i$, and survival. Note that if $\hat{\boldsymbol{u}}_i$ instead of $\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_i$ is used, the value of β_S changes, but not the values of $\boldsymbol{\alpha}$ and γ. Further note that $\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_i$ is a person specific summary and does not change over time. It is important to note that a person's longitudinal trajectory acts as a baseline covariate in the survival component of the joint model. That is, as a person's MLHF scores change over time, it is assumed to be due to an unchanging personal trajectory parameter $\hat{\boldsymbol{u}}_i$, which is not a time-varying covariate for survival.

2.2.2 The Joint Model

Consider a joint model where the longitudinal and the event time components are conditionally independent given the latent variables. The assumption of conditional independence follows the example of previous joint modeling papers, including Henderson et al.[1]. That is, the complete data joint density of the longitudinal, the survival data, and the latent variables, is assumed to be the product of the conditional densities:

$$f(\boldsymbol{y}_i, t_i, \delta_i, \boldsymbol{u}_i \mid \boldsymbol{x}_i, \theta_L, \theta_S, \theta_U) =$$

$$f_L(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{u}_i, \theta_L, \theta_S, \theta_U) \times f_S(t_i, \delta_i \mid \boldsymbol{u}_i, \boldsymbol{x}_i, \theta_L, \theta_S, \theta_U) \times f_U(\boldsymbol{u}_i \mid \theta_L, \theta_S, \theta_U).$$

This conditional independence formulation implies the longitudinal and survival data are not related to one another except through their shared relation with u_i.

The natural extension of the two-step model described in subsection 2.1 to the joint model case is:

$$y_{ij} \mid x_{ij}, \boldsymbol{u}_i, \boldsymbol{\beta}, \sigma^2 \overset{ind}{\sim} N\left(\begin{bmatrix} 1 & x_{ij} & x_{ij}^2 \end{bmatrix} (\boldsymbol{\beta} + \boldsymbol{u}_i), \sigma^2\right) \tag{2.1}$$

$$\Pr(s_i \geq r | \mu_i, \gamma) = \exp\left\{-\exp\left\{-\mu_i \gamma\right\} r^{\gamma}\right\} \tag{2.2}$$

$$f(s_i | \mu_i, \gamma) = \gamma \exp\left\{-\mu_i \gamma\right\} s_i^{\gamma - 1} \exp\left\{-\exp\left\{-\mu_i \gamma\right\} s_i^{\gamma}\right\} \tag{2.3}$$

$$\mu_i = \beta_S + \boldsymbol{\alpha}^T (\boldsymbol{\beta} + \boldsymbol{u}_i) \tag{2.4}$$

$$\boldsymbol{u}_i \overset{iid}{\sim} N(\boldsymbol{0}, \Sigma) \tag{2.5}$$

where the key difference between the two-step and the joint model is in the specification for μ_i. In the two-step model, μ_i is based on the fitted values, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}_i$ from the longitudinal model, while in the joint model, μ_i is a function of $\boldsymbol{\beta}$ and $\boldsymbol{u}_i$. The joint approach uses all the data, longitudinal and event time, to simultaneously estimate all the parameters. Note that in the two-step model, the latent variables are predicted using the longitudinal data, and those predicted latent variable values are used to model

survival, but, in the joint model, latent variables are predicted using both the longitudinal and event time data. In particular, the predictor of the latent variable $\tilde{\boldsymbol{u}}_i$ for the joint model is $\mathrm{E}(\boldsymbol{u}_i \mid \boldsymbol{y}_i, \boldsymbol{x}_i, t_i, \delta_i, \tilde{\theta}_L, \tilde{\theta}_S, \tilde{\theta}_u)$, where $\tilde{\theta}_L$, $\tilde{\theta}_S$, and $\tilde{\theta}_u$ are the maximum likelihood estimators from the joint model.

2.2.3 Interpreting parameters

An individual's true longitudinal trajectory is governed by both $\boldsymbol{\beta}$ and his or her $\boldsymbol{u}_i$. The mean trajectory is determined by $\boldsymbol{\beta}$, and $\boldsymbol{u}_i$ describe how an individual's trajectory differs from the mean trajectory. A positive u_{i0} (corresponding to the intercept) indicates that the i^{th} individual's trajectory starts higher than the average trajectory. Similarly, positive values of u_{i1} (slope) and u_{i2} (quadratic term) indicate that the i^{th} individual's trajectory increases as a function of time and time squared, respectively, faster than the mean trajectory.

The effect of the $\boldsymbol{u}_i$ on the probability of the i^{th} individual experiencing an event is clear when the log hazard function is examined. Assuming a Weibull distribution for the event times, as in (2) and (3), the log hazard function is: $\log h_i(t) = \log\gamma + (\gamma - 1)\log t - \gamma\mu_i$, where μ_i is as in (4). Since γ is always positive, large values of μ_i produce small log hazards and small values of μ_i produce large log hazards. Thus, individuals with good survival probabilities have large μ_i and individuals with poor survival probabilities have small μ_i.

2.3 Predictions and Prediction Errors

Graf et al.[8] suggest a method for calculating the prediction error (PE) when the status of an individual is known at the end of the prediction interval, and Schoop et al.[9] applied this approach to event time models with time-varying covariates. This approach can also be applied to two-step and joint models for longitudinal and event time data. The first step in prediction is to determine the prediction scenarios of interest. Each scenario involves both a time that the prediction is made and an interval of interest. The prediction time is the point at which the prediction is made. It is a time at which an individual is at risk for experiencing the event. The prediction itself is an estimate of the probability that an individual does not experience the event during

some interval after the prediction time.

Consider two prediction times (6 months and 12 months), and two prediction intervals (6 months and 12 months) to produce four scenarios. Let t_p be the prediction time and w be the prediction interval. Then, if $t_p = 6$ and $w = 6$, a prediction of the probability that each individual alive at t_p will be event free during the next 6 months will be made based on the basis of data up to 6 months. That is, $\Pr(t_p+w \leq s_i \mid t_p \leq s_i)$ is estimated.

Let $\hat{p}_i = \hat{\Pr}(t_p+w \leq s_i \mid t_p \leq s_i)$ for a particular model for participant i. This is only meaningful for individuals who are known to be at risk at t_p. Let $q_i = I(t_p + w \leq s_i)$, where q_i is only known if the participant was not censored during the interval. If participant i was at risk at t_p and was not censored before t_p+w, then the squared error for participant i is $(\hat{p}_i - q_i)^2$. A weighted average of the squared errors is an estimate of the prediction error for the model. Schoop et al.[9] give the following formula for calculating prediction error, when censoring is independent:

$$\hat{PE}(t_p, w) = \frac{1}{n_t} \sum_{i=1}^{n} \left[\frac{I(t_p < t_i \leq t_p + w)\hat{p}_i^2 \delta_i}{\Pr(c_i \geq t_i \mid c_i \geq t_p)} + \frac{I(t_i > t_p + w)(1 - \hat{p}_i)^2}{\Pr(c_i \geq t_p + w \mid c_i \geq t_p)} \right] \tag{2.6}$$

where n_t is the number at risk at t_p. Since nearly all the censoring in the example heart failure data set was due to the end of the study, it is reasonable to assume that the censoring was independent of the longitudinal and event time processes. In equation (6), the probability of censoring is used to weight each individual's prediction error. A Kaplan-Meier model was used to estimate the probability of censoring when calculating the prediction errors.

For the two step model, the empirical Bayes estimates $\hat{\boldsymbol{u}}_i^*$ are found, where $\hat{\boldsymbol{u}}_i^* = \mathrm{E}(u_i \mid \boldsymbol{x}_i^*, \boldsymbol{y}_i^*, \hat{\beta}, \hat{\sigma}^2, \hat{\Sigma})$, and $\boldsymbol{x}_i^*$ and $\boldsymbol{y}_i^*$ are the predictors and responses for individual i up to time t_p and $\hat{\beta}$, $\hat{\sigma}^2$, and $\hat{\Sigma}$ are the maximum likelihood estimators from the fit of the two step model to the whole data set. Setting $\hat{\mu}_i^* = \hat{\beta}_S + \hat{\boldsymbol{\alpha}}^T(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_i^*)$, where $\hat{\beta}_S$ and $\hat{\boldsymbol{\alpha}}$ are the maximum likelihood estimators from the fit of the two step model to the whole data set, then leads to:

$$\hat{p}_i = \frac{\hat{\Pr}(t_p + w \leq s_i)}{\hat{\Pr}(t_p \leq s_i)} = \frac{\exp\left(-\exp\left[-\hat{\mu}_i^* \hat{\gamma}\right](t_p + w)^{\hat{\gamma}}\right)}{\exp\left(-\exp\left[-\hat{\mu}_i^* \hat{\gamma}\right](t_p)^{\hat{\gamma}}\right)}$$

For the joint model, the empirical Bayes estimates for the latent variables is found:

$\tilde{\boldsymbol{u}}_i^* = \mathrm{E}(u_i \mid \boldsymbol{x}_i^*, \boldsymbol{y}_i^*, t_i^*, \delta_i^*, \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \tilde{\Sigma}, \tilde{\beta}_S, \tilde{\gamma}, \tilde{\boldsymbol{\alpha}})$, where $\boldsymbol{x}_i^*$, $\boldsymbol{y}_i^*$, t_i^*, and δ_i^* are the predictors and responses for individual i up to time t_p, and $\tilde{\boldsymbol{\beta}}$, $\tilde{\sigma}^2$, $\tilde{\Sigma}$, $\tilde{\beta}_S$, $\tilde{\gamma}$, and $\tilde{\boldsymbol{\alpha}}$ are the maximum likelihood estimators from the joint model fit to the whole data set. If an individual is at risk at the time of prediction, then the event time must be greater than the time of prediction, that is $s_i > t_p$. Since, for the purpose of prediction, the exact value of s_i at the prediction time is unknown, $t_i^* = t_p$ and $\delta_i^* = 0$. In effect, when making predictions, each individual at risk is treated as censored at the prediction time for the purpose of obtaining $\tilde{\boldsymbol{u}}_i^*$. Calculation for $\hat{\mu}_i$ and $\hat{p}_i$ proceed in the same manner as the two-step model, but using the parameter estimates from the joint model.

Schoop et al.[9] suggest using the Kaplan-Meier probability estimates as a baseline for comparison. The Kaplan-Meier estimates do not incorporate covariate data, such as longitudinal measurements. Predicted survival probabilities based on the Kaplan-Meier survival curve represent the best prediction based only the available event time data. Predictions from models with covariates are expected to outperform the Kaplan-Meier predictions, since some of the variability in survival is expected to be explained by the covariates. Schoop et al.[9] also suggest using the decrease in prediction error relative to the Kaplan-Meier prediction error as a measure of the variation explained by a model. Since parametric event time models were used, the predictive performance of the joint and two-step models were compared to both a Kaplan-Meier model and a parametric Weibull survival model without covariates. The predictive performance of the Kaplan-Meier model and the Weibull model were nearly identical, so the choice of baseline model did not affect the results.

2.4 Results

2.4.1 Parameter Estimates in the Two-Step Model

The SAS procedures MIXED and LIFEREG were used to obtain parameter estimates for the two-step model. First, consider the parameters of the longitudinal part of the two-step model. Table 2.1 contains the parameter estimates from the two-step model. Examining $\hat{\boldsymbol{\beta}}^T = [\hat{\beta}_0 \ \hat{\beta}_1 \ \hat{\beta}_2]$, i.e. estimates for the parameters governing the average quadratic MLHF trajectory, the population mean MLHF score at time zero for the

sample is found to be $\hat{\beta}_0 = 34.85$, and the quadratic term $\hat{\beta}_2$ is positive and statistically significant. Examination of the covariance of $\boldsymbol{u_i}$ finds significant person to person trajectory differences. The variance of the u_{0i} shows that MLHF scores at time zero covered the full range of possible scores. The correlation between the u_{1i} and u_{2i} is large and negative, approximately -0.89.

Since $\hat{\gamma}$ is greater than one, the hazard increases over time. The $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are all negative and statistically significant, indicating that larger values of the latent variables tend to increase the hazard and decrease survival. Since the latent variables are the individual deviations from the mean longitudinal MLHF trajectory, this implies that individuals whose apparent trajectories are higher than average, or increase faster than average, are at increased risk of death.

2.4.2 Parameter Estimates for the Joint Model

The SAS procedure NLMIXED was used to generate the joint model parameter estimates using the clinical trial data. Parameter estimates from separate longitudinal and survival models were used as starting values for the joint model for all parameters except for $\boldsymbol{\alpha}$, which was started as a zero vector. Table 2.2 contains the parameter estimates from the joint model.

The parameter estimates for the longitudinal part of the joint model are generally similar to their counterparts in the two-step model. The estimate for β_1 is positive in the fitted two-step model, and negative in the joint model, but both are near zero and not statistically significant. The mean fitted trajectory is shown in the plots of Figure 2.1 as a thick black line. The covariance matrix for the latent variables in the fitted two-step model is also similar to its counterpart in the fitted joint model, and the correlation between the u_{1i} and u_{2i} is large and negative, approximately -0.92.

The joint model estimate for γ is larger than the two-step model estimate, indicating that the hazard changes over time more in the joint model than in the two-step model. Values for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in particular are larger in the joint model than in the two-step model. So, based on the parameter estimates the joint model indicates there is a stronger relationship between longitudinal trajectory and survival.

2.4.3 Fitted Trajectories and Hazards from the Joint Model

In the joint model, the individual trajectories and survival curves for each individual are both functions of the latent variables. Using empirical Bayes estimates $\tilde{\boldsymbol{u}}_{\boldsymbol{i}}$ from the joint model given all the available data, fitted trajectories and survival curves can be produced for each individual. Since the survival curve parameter γ is fixed for all participants, each participant's fitted survival curve is a function of $\hat{\mu}_i = \tilde{\beta}_S + \tilde{\boldsymbol{\alpha}}^T(\tilde{\beta} + \tilde{\boldsymbol{u}}_i)$. Thus, $\hat{\mu}_i$ provides a convenient summary of the i^{th} individual's fitted survival curve and hazard. The upper left panel of Figure 2.1 presents a histogram of the $\hat{\mu}_i$, ranging from 5.42 to 11.06. The range of the $\hat{\mu}_i$ describes meaningful differences in fitted survival. The probability that an individual with $\mu_i = 6.23$ survives beyond 2 years is only 0.176. A person with $\mu_i = 9.43$ will live beyond 2 years with probability 0.987.

Figure 2.1 also displays a sample of fitted trajectories for individuals with a variety of estimated risks. Individuals with relatively high $\hat{\mu}_i$ tend to have low or decreasing MLHF trajectories over the study period. Participants with trajectories near the 0.25 and 0.5 quantiles of $\hat{\mu}_i$ have trajectories that appear to increase at about the rate of the mean trajectory, but have a variety of intercepts. The worst fitted trajectories tend to increase rapidly over the study period, though in many cases the trajectory only increases substantially after longitudinal measurements stopped.

Figure 2.2 provides scatterplots of the estimates of $\hat{\mu}_i$ for the participants whose death was observed and censored participants separately. All but three participants whose death was observed had a fitted hazard higher than the hazard associated with the mean trajectory. The two panels demonstrate that the joint model estimates high log hazards (i.e. low $\hat{\mu}_i$) for participants with observed deaths and log hazards near the average for censored individuals. This also explains the bimodality of the histogram in Figure 2.1. The mode at smaller values of $\hat{\mu}_i$ is largely individuals whose death was observed, and the other mode is largely censored individuals. In fact, 95.7% of the individuals who died had $\hat{\mu}_i$ in the lowest quartile.

Clearly, the distribution of the $\hat{\mu}_i$ for the participants whose death was observed is different from the distribution of $\hat{\mu}_i$ for censored individuals. This difference in fitted hazards for those who died and those who did not may seem to indicate the trajectories for participants who died were truly different from the trajectories of censored individuals. In other words, examining these fitted values may lead to the conclusion

that the joint model is effectively describing the relationship between QOL trajectory and survival. But a more likely explanation is that since the predicted latent variable values $\tilde{\boldsymbol{u}}_{\boldsymbol{i}}$ are based on both the longitudinal and survival data, the fitted trajectories are being more strongly influenced by the survival data than by the longitudinal data itself. Figure 2.2 also shows that there is a strong relationship between survival time and $\hat{\mu}_i$ in the participants whose death was observed. The relationship between censoring time and $\hat{\mu}_i$ does not appear to be nearly as strong, indicating that censoring time has little influence on the fitted hazard.

To demonstrate the strong influence the event time has on fitted trajectory, Figure 2.3 shows the fitted trajectories for two individuals with similar longitudinal QOL observations. Participant 18 and participant 1931 both have four QOL observations, and they were taken at the same time points. Each observed score is less than 20, which indicates little impact of heart failure on daily life. Participant 18 died at 232 days, but participant 1931 was censored at 757 days. The fitted QOL trajectory for participant 18 clearly curves upward, and increases faster than the mean trajectory. The fitted QOL trajectory for participant 1931 has little curvature, and does not increase substantially faster than the mean trajectory. This demonstrates that differences in survival dramatically affect fitted trajectories. The fitted hazards for the two participants are also clearly different, with $\hat{\mu}_{18} = 6.44$ and $\hat{\mu}_{1931} = 8.58$. Although the fitted trajectories for the two participants are close to the their observed MLHF scores, the two fitted trajectories are clearly different after the longitudinal observations stop.

2.4.4 Prediction Error

Using the heart failure clinical trial data, prediction errors were estimated for four models, a Kaplan-Meier nonparametric survival model, a parametric Weibull survival model without covariates, the two-step model, and the joint model. Each model was applied to four prediction scenarios described earlier. Figure 2.4 shows histograms of predicted survival probabilities for all individuals at risk from the two-step and the joint models under the four prediction scenarios. In each scenario, the joint model predicted higher survival probabilities than the two step model predicted. In general, the joint model predicted survival probabilities higher than the predicted survival probability from the fitted Kaplan-Meier model, while the two step model's predicted survival

probabilities were near the survival probability predicted by the fitted Kaplan-Meier model. Table 2.3 shows the prediction error estimates based on the method described in Section 2.3. The two-step model had the lowest prediction error and the joint model had the highest prediction error in all four prediction scenarios. Kaplan-Meier and Weibull models produced similar prediction errors in each of the four scenarios. Prediction errors were smaller when the prediction interval was smaller.

2.5 Simulation

To further examine the prediction ability of the joint model as compared to the two step model, a simulation study was performed. One hundred data sets, each with one thousand individuals, were generated under the assumptions of a joint longitudinal-survival model with independent censoring. Parameters were chosen to produce similar results to the motivating data set, but a linear longitudinal model instead of a quadratic longitudinal model was selected to achieve a high model convergence rate. For each simulation, predictions were made according to the fitted joint model, the fitted two-step model, a Kaplan-Meier survival model, and a Weibull survival model without covariates for each of the four scenarios described in Section 2.3. Fitting the joint model is more computationally challenging than fitting the two step model. Even with very good starting values, each joint model takes about three minutes to converge, while the two-step models converge in less than one minute. Prediction errors were calculated for each scenario and model. Since the data were simulated, the true status of each individual was known at all times, so the weighting of the individual prediction errors by the probability of censoring was not necessary. The mean prediction error for the four models and the four scenarios is shown in Table 2.4. In these simulations, the prediction errors for the two-step model and the joint model were similar. This suggests that the additional effort required to fit a joint model over a two-step model may not yield substantially better predictions, even when the joint model assumptions hold. Both the joint model and the two-step model performed slightly better than the Kaplan-Meier model and the Weibull model without covariates.

Because these simulations were unable to reproduce the prediction error difficulties

shown in the heart failure data set, another set of simulations was produced to investigate the effect of changes to the relative information content of the longitudinal and event time data. In this second set of simulated data sets, the parameters were identical to the previous simulations, except the measurement error standard deviation in the longitudinal model was increased by 50%, from $\sigma^2 = 100$ to $\sigma^2 = 225$. This weakens the longitudinal data relative to the event time data. Since this change also reduces the overall information about each subject, the number of subjects in each data set was doubled to 2000. Table 2.5 shows the mean prediction errors for the four models and four model scenarios using the high variance simulated data.

As in the heart failure data set, the joint model's prediction error is higher than the prediction errors for the Kaplan-Meier and Weibull models without covariates, while the two-step model had the lowest prediction error. Additional investigations reveal why the joint model is performing poorly. Because each simulated data set had a somewhat different difficulty of prediction, it is useful to look at the difference between the prediction error of the joint model or two-step model and the prediction error of the Kaplan-Meier model. The top two panels of Figure 2.5 show histograms of these prediction error differences for the joint model and the two-step model. The histogram of prediction error differences for the joint model is distinctly bimodal, with one mode centered near zero and another model centered near 0.001. Some, but not all, of the fitted joint models are performing substantially worse than the Kaplan-Meier models. This bimodality is not seen for the two-step models, whose histogram has a single mode centered near -0.0001.

The left middle panel of Figure 2.5 shows the histogram of the mean prediction for the fitted joint models, where the mean prediction is the average probability of survival of all of the at-risk subjects for a particular data set. Again, this distribution is also bimodal. One mode is near the mean of the 100 Kaplan-Meier predictions, and the other mode is clearly higher. This second mode echoes the results shown in Figure 2.4 which showed joint model predictions that were higher than the Kaplan-Meier prediction in the heart failure data set. The right middle panel of Figure 2.5 shows the bimodal distribution of the fitted $\tilde{\alpha}_1$ for the 100 fitted joint models. One mode is near the true value of α_1, with the other mode representing a stronger association between the slope and the hazard.

The bottom two panels of Figure 2.5 show the scatterplots of the prediction error differences and the mean predictions of fitted $\tilde{\alpha}_1$. It is clear that the fitted joint models with the most optimistic mean predictions had worse prediction error. The fitted models with the strongest association between slope and hazard also had worse prediction error.

2.6 Discussion

The prediction error results are counterintuitive. Henderson et al.[1] suggest that the parameter estimates from the joint model should be less biased than the two-step model. Intuition suggests that the joint model should predict better, but the prediction error for the two-step model using the clinical trial data was lower than the prediction error for the joint model. Furthermore, the joint model's prediction error was even higher than the Kaplan-Meier model.

The strong influence of the survival data on the predicted μ_is in the joint model may provide some insight into the results. The joint model manages to sort the participants into a higher risk group and a lower risk group–this can be seen in the bimodal histogram for the predicted $\hat{\mu}_i$s in Figure 2.1. This sorting, though, is strongly influenced by the survival data rather than the longitudinal data. When predictions are made, each individual is treated as censored at the prediction time, and hence, for the purpose of prediction, the joint model tends to place every individual in the lower risk group. Figure 2.4 shows that nearly all joint model survival probability predictions are higher than the Kaplan-Meier prediction, indicating overoptimistic predictions by the joint model.

The joint model appears to achieve this effect through shrinkage. In the joint model, like a linear mixed model, trajectories tend to shrink toward the mean trajectory. The event time model component, however, allows event time data, as well as the longitudinal data, to work against the shrinkage tendency. Most of the participants survived to be censored at the end of the study, so the average participant had low apparent risk of death.

The joint model appears to be more prone to this undesirable behavior when the longitudinal data is weaker. Asymptotic calculations very likely cannot capture this effect because a arbitrarily large sample of longitudinal data should generally overcome

a problem of large error variation that could be decisive in a finite sample. When the longitudinal data is weak, the event time data can have a greater influence on the fitted trajectory for individuals who experience events. The best-fitting models are more likely to have an overly strong association between trajectory and hazard. During prediction, when the event time data is minimal for all subjects, the joint model cannot distinguish between the low and high hazard groups well because the joint model depended on the observed events to make that distinction.

In contrast, the two-step model has no mechanism to allow survival data to influence predicted trajectories. The survival component of the two-step model uses the predicted trajectories from the longitudinal modeling. As a result, the two-step model does a better job of capturing the predictive value of an individual's apparent trajectory on survival. And, although the longitudinal trajectory appears to have limited value in predicting survival in this clinical example, the two-step model does outperform the Kaplan-Meier predictions. Figure 2.4 shows that the two-step model predictions tend to be distributed around the Kaplan-Meier prediction, unlike the joint model predictions.

The high statistical significance of the $\boldsymbol{\alpha}$ parameter estimates in the joint model suggests that there is a relationship between survival and trajectory. Since the joint model is fit using a single step, it is impossible to infer from the fitted model parameters alone whether the longitudinal trajectory is a good predictor of survival or whether survival is a good predictor of longitudinal trajectory. It is apparent that the joint model allows the survival data to strongly influence the predicted trajectory. In contrast, the highly significant $\boldsymbol{\alpha}$ parameters in a two-step model can indicate whether the longitudinal trajectory is a useful predictor of survival. Even highly significant $\boldsymbol{\alpha}$ can lead to only modest improvements in prediction performance, though. The two-step model, both in the heart failure clinical trial data and in simulations, performed only marginally better than the predictions from a Kaplan-Meier model.

Hanson et al.[10] demonstrated that joint model predictions may not be substantially better than predictions from two-step models or event-time models with time-varying covariates in a Bayesian framework. Our example and simulation confirms that the two-step model can provide comparable, or even superior, prediction performance, at least in some situations. In addition, our motivating data set demonstrates that joint model predictions may not even outperform predictions based on a Kaplan-Meier model,

despite the joint model's inclusion of additional information. Simulation results show that the prediction error of the joint model can be similar to the prediction error of the two-step model, even when the joint model's assumptions are correct. The high variance simulations also demonstrate that the joint model is prone to overestimate the association between trajectory and hazard when the contribution of the longitudinal component is relatively weak, which can lead to poor prediction performance.

These simulations do not present a compelling case for choosing the joint model over the two-step model. It is substantially more difficult computationally to fit the joint model, and model fitting algorithms may have difficulty converging. In contrast, the two-step model can be easily fit using standard software using algorithms that rarely experience convergence problems. If prediction is important, then the extra effort to fit a joint model may not be justified.

Parameter	Estimate	Standard Error	t Statistic	p-value
β_0	34.8539	0.4906	71.04	$< .0001$
β_1	0.0141	0.05148	0.36	0.7206
β_2	0.006104	0.002140	2085	0.0044
σ^2	90.1994	1.2724	70.89	$< .0001$
β_S	8.3336	0.1241	67.13	$< .0001$
α_0	-0.00814	0.002113	-3.85	0.0001
α_1	-0.4735	0.1141	-4.15	$< .0001$
α_2	-10.0491	3.5489	-2.83	0.0046
γ	1.1232	0.05051	22.24	$< .0001$
$\text{Var}[u_{i0}]$	458.65	15.6212	29.36	$< .0001$
$\text{Var}[u_{i1}]$	2.7496	0.1764	14.06	$< .0001$
$\text{Var}[u_{i2}]$	0.003171	0.000305	10.40	$< .0001$
$\text{Cov}[u_{i0}, u_{i1}]$	-4.4599	1.1995	-3.72	0.0002
$\text{Cov}[u_{i0}, u_{i2}]$	0.09469	0.04916	1.93	0.0542
$\text{Cov}[u_{i1}, u_{i2}]$	-0.08287	0.007134	-11.62	$< .0001$

Table 2.1: Table of Parameter Estimates: Two-Step Model

Parameter	Estimate	Standard Error	t Statistic	p-value
β_0	34.9033	0.4905	71.16	$< .0001$
β_1	-0.02168	0.05078	-0.43	0.6695
β_2	0.01177	0.002067	5.7	$< .0001$
σ^2	92.5744	1.7625	52.52	$< .0001$
β_S	9.5359	0.3374	28.3	$< .0001$
α_0	-0.0132	0.003537	-3.73	0.0002
α_1	-2.1129	0.4294	-4.92	$< .0001$
α_2	-70.7266	19.1581	-3.69	0.0002
γ	1.5221	0.2265	6.72	$< .0001$
$\text{Var}[u_{i0}]$	457.09	15.6481	29.21	$< .0001$
$\text{Var}[u_{i1}]$	2.1059	0.2528	8.33	$< .0001$
$\text{Var}[u_{i2}]$	0.002235	0.000556	4.02	$< .0001$
$\text{Cov}[u_{i0}, u_{i1}]$	-3.917	1.2256	-3.2	0.0014
$\text{Cov}[u_{i0}, u_{i2}]$	0.07891	0.05031	1.57	0.117
$\text{Cov}[u_{i1}, u_{i2}]$	-0.06303	0.01153	-5.47	$< .0001$

Table 2.2: Table of Parameter Estimates: Joint Model

Model	$t_p = 6$, $w = 6$	$t_p = 6$, $w = 12$	$t_p = 12$, $w = 6$	$t_p = 12$, $w = 12$
Kaplan-Meier	0.0455	0.0909	0.0503	0.0948
Weibull	0.0456	0.0909	0.0503	0.0948
Two-Step	0.0453	0.0902	0.0498	0.0934
Joint	0.0464	0.0945	0.0510	0.0978

Table 2.3: Table of Prediction Error Estimates for the Heart Failure Clinical Trial data. The prediction time, t_p, and the prediction interval, w, are both in months. The two-step model had the lowest prediction error, and the joint model had the highest prediction error.

Model	$t_p = 6$, $w = 6$	$t_p = 6$, $w = 12$	$t_p = 12$, $w = 6$	$t_p = 12$, $w = 12$
Kaplan-Meier	0.0493	0.0956	0.0545	0.1037
Weibull	0.0494	0.0956	0.0546	0.1037
Two-Step	0.0493	0.0952	0.0542	0.1022
Joint	0.0493	0.0953	0.0541	0.1021

Table 2.4: Table of Mean Prediction Error in 100 Simulations. On average, the prediction errors for the two-step model and the joint model are similar and smaller than the mean prediction errors for the Kaplan-Meier model and the Weibull model without covariates.

Model	$t_p = 6$, $w = 6$	$t_p = 6$, $w = 12$	$t_p = 12$, $w = 6$	$t_p = 12$, $w = 12$
Kaplan-Meier	0.0496	0.0957	0.0545	0.1027
Weibull	0.0496	0.0957	0.0556	0.1028
Two-Step	0.0495	0.0954	0.0543	0.1018
Joint	0.0499	0.0969	0.0546	0.1029

Table 2.5: Table of Mean Prediction Error in 100 High Variance Simulations. The joint model has the highest average prediction error. The two-step model prediction errors are lower than the prediction errors of the Kaplan-Meier and Weibull models without covariates.

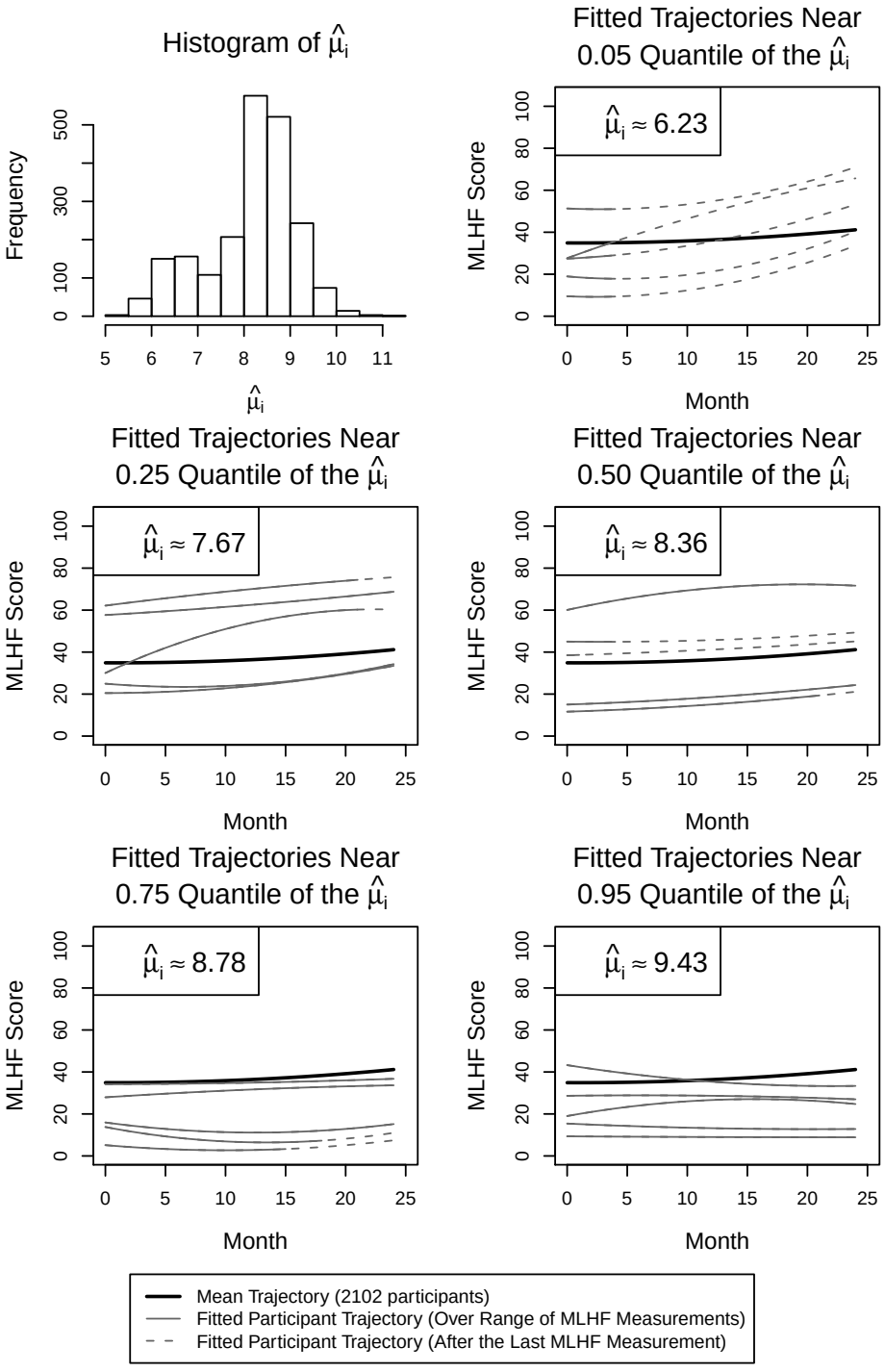

Figure 2.1: Example Trajectories

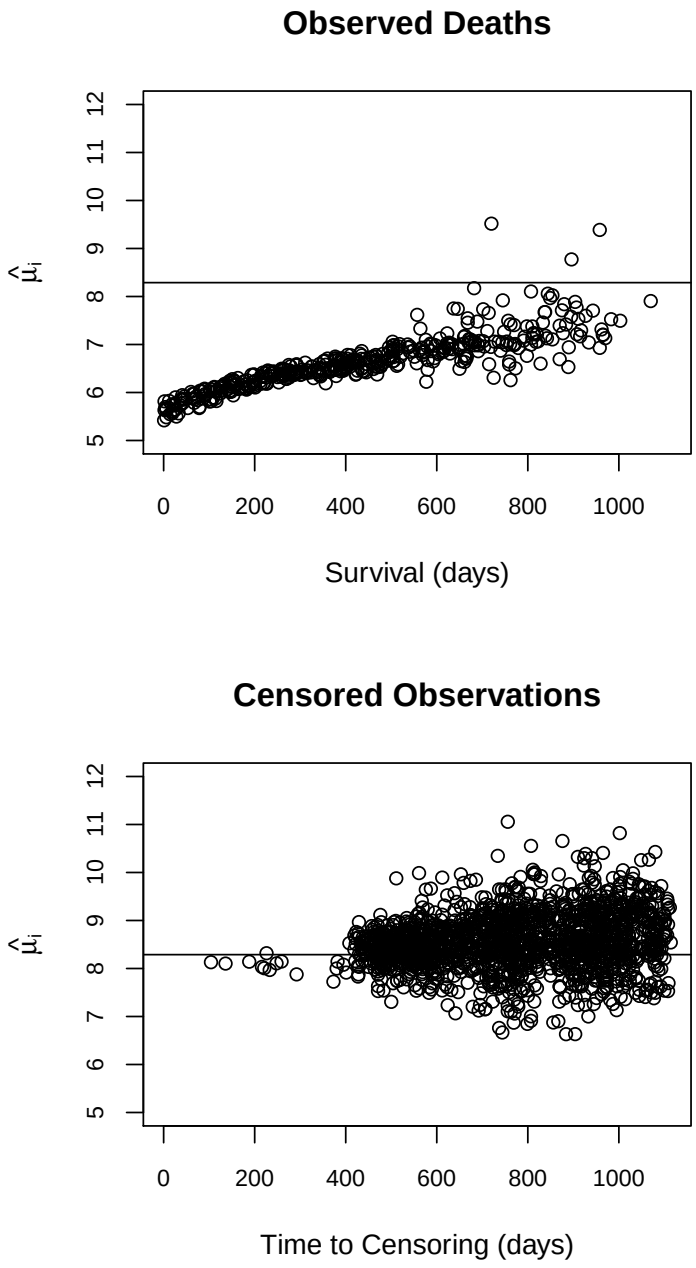

Figure 2.2: Plot of $\hat{\mu}_i$ for participants whose death was observed and censored participants. The horizontal line indicates the value of $\hat{\mu}_i$ associated with the mean longitudinal trajectory.

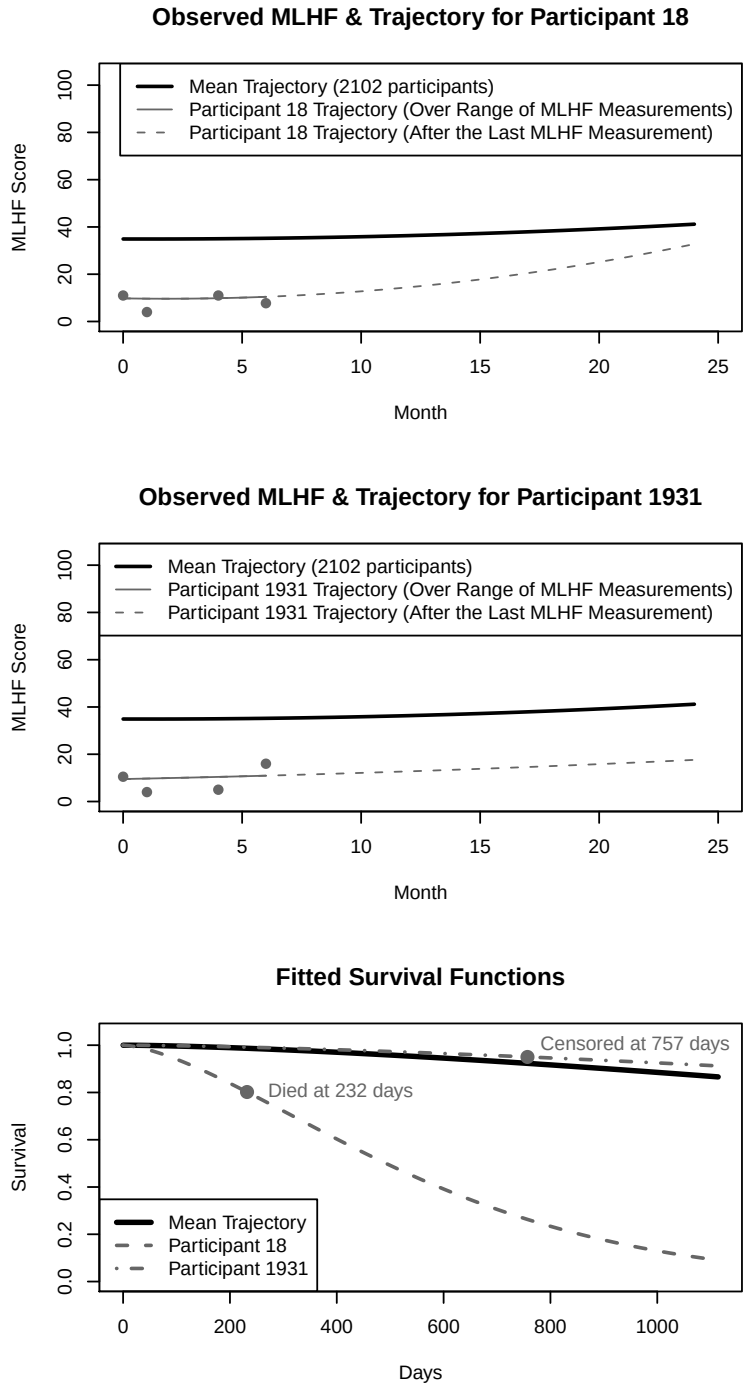

Figure 2.3: Two participants with similar longitudinal observations but different survival data.

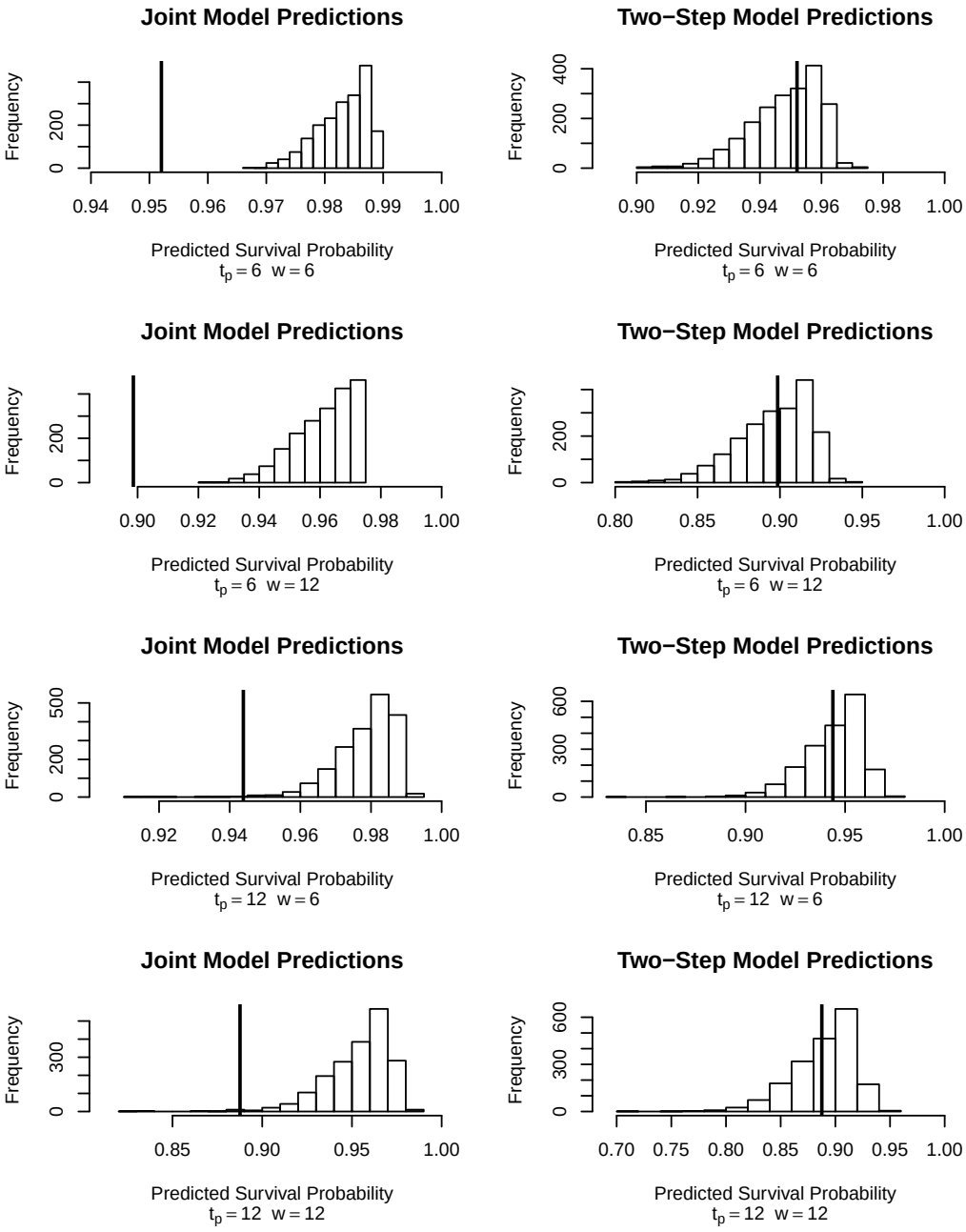

Figure 2.4: Histograms of predicted survival probabilities for the clinical trial data set under the joint model and under the two-step model. The Kaplan-Meier predicted survival probability is indicated by the thick vertical line.

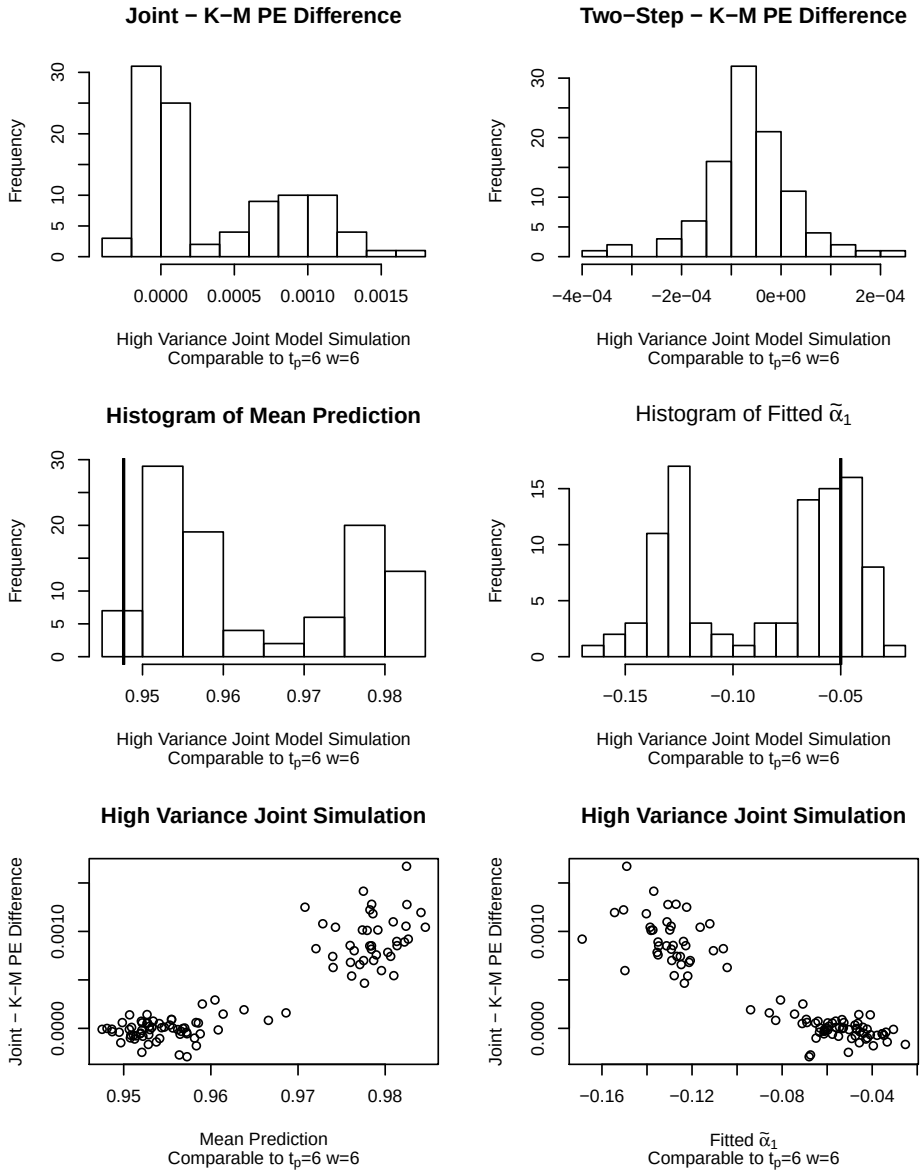

Figure 2.5: The top two panels show histograms of the difference in the prediction errors between the joint model or two-step model and the Kaplan-Meier model for the 100 simulated high variance data sets for the first prediction scenario. The middle two panels show histograms for the mean prediction for the joint model and the fitted value of $\tilde{\alpha}_1$ in the 100 joint model fits. The vertical lines are the mean Kaplan-Meier prediction and the true value of α_1, respectively. The bottom two panels show scatterplots of the prediction difference and the mean prediction or $\tilde{\alpha}_1$.

Chapter 3

Inference on a Partially Missing Mediator

3.1 Introduction

The mechanism by which an experimental treatment achieves its effect is often the subject of scientific interest. In particular, treatments in medical studies almost always affect the human body in multiple ways. Measuring a process that is believed to mediate the effect of the treatment is a useful way to infer whether the treatment works through the hypothesized mechanism or another mechanism. Unfortunately, measurements of the mediating process may be missing.

When the missing data mechanism is nonignorable, some model for that mechanism is needed. Joint models can provide a useful framework for inference when there is nonignorable missing data. When longitudinal data is missing because an event occurs, joint models for the longitudinal and event processes have been proposed that model the association between the two processes (e.g. Wulfsohn and Tsiatis[3], Henderson et al.[1], Guo and Carlin[4], Vonesh et al.[5]). These joint models for longitudinal and event time data have been shown to decrease bias in longitudinal model parameter estimation when longitudinal data is not missing at random (Henderson et al.[1]). The joint model framework is quite flexible, and the overview of joint models by Tsiatis and Davidian[6] is particularly useful. This chapter will use a specific class of joint models that includes shared latent variables to link a model for the longitudinal data and a model for the

missing data mechanism (Henderson et al.[1]). Guo and Carlin[4] demonstrated how to fit Bayesian versions of these models using WinBUGS software.

The motivating data set for this chapter comes from an investigation into the effect of X-inactivation on the severity of disease in a mouse model of X-linked Alport syndrome (XLAS). X-inactivation is the process by which one of the two X chromosomes in female cells is rendered inactive. X-linked Alport syndrome is a hereditary disorder of the basement membranes that can progress to end-stage kidney disease. Mutations of the *COL4A5* gene on the X chromosome cause the disorder. Men with XLAS, who only have one X chromosome, generally have more severe symptoms than women with XLAS, who have two X chromosomes and thus are carriers. Differences in X-inactivation have been hypothesized to explain variability in disease severity.[11]

Rheault et al.[12] developed a mouse model for XLAS by introducing a mutation into the mouse *Col4a5* gene, corresponding to the human *COL4A5* gene. Female mice, like women, can be carriers of the mutation. That is, they can carry both the mutant and wild-type versions of the gene. The mouse X-inactivation process is regulated by the X-controlling element *Xce* gene on the X chromosome. By selectively breeding 88 mice, [11] were able to produce two groups of carrier female mice with different X-inactivation probabilities for the mutant mouse *Col4a5* gene. Thirty-eight Group 1 mice preferentially inactivated the X chromosome that carried the wild-type *Col4a5*, while the fifty Group 2 mice preferentially inactivated the mutant-*Col4a5*-carrying X chromosome. They hypothesized that the Group 1 mice should exhibit more severe disease than the Group 2 mice and that this effect would be due to (i.e., mediated by) increased X-inactivation in the Group 1 mice.

Two measurements of mouse disease severity were made. Urine protein excretion was measured at 2, 4, and 6 months and at 6 months plasma urea nitrogen (PUN) was measured. Measurements of X-inactivation were indirectly made by measuring the expression of the *Srpx*, *Rpgr* and *Aff2* genes. These measurements of gene expression could only be made at the time of sacrifice (i.e., just after the 6 month measurements). Unexpectedly, 14 Group 1 mice and 3 Group 2 mice died before the planned sacrifice at 6 months, which resulted in their PUN, gene expressions, and final protein excretion levels not being measured. Although the early deaths of these mice was not anticipated, the survival status of the mice effectively serves as a third measure of disease severity.

Based on the mice that did not die prematurely, Group 1 mice appear to have higher mean PUN measurements, higher X-inactivation measurements (i.e., *Srpx*, *Rpgr* and *Aff2* gene expressions), and more rapidly increasing protein excretion rates. These measurements suggest that the surviving Group 1 mice had more severe disease than the surviving Group 2 mice. The gene expressions, log PUN, and 6-month urine protein excretions are unknown for the mice that died prematurely, but their deaths imply that their disease was more severe than the surviving mice. As a result, excluding these mice from the analysis may introduce bias because of informative censoring. The goal of this chapter is to make inference on X-inactivation as a mediator between experimental group and disease severity including premature death using all the mice, not just those that survived to the end of the experiment.

This chapter reanalyzes the Rheault et al.[11] data, using a Bayesian joint model to allow inference on the partially missing X-inactivation mediator. The Bayesian joint model will treat X-inactivation as a latent variable underlying the 3 gene expression measures and will examine X-inactivation as a full and partial mediator between experimental group and disease severity. WinBUGS software is used to fit all the models. Section 3.2 describes the structure of the fitted models. Section 3.3 presents the analysis results, and Section 3.4 provides discussion of the results.

3.2 Models

Figure 3.1 shows the different models which were considered. Model 1 depicts the assumed effects of the experimentally manipulated group variable, i.e., it was expected that group membership would cause differences in X-inactivation which in turn would manifest as differences in gene expression levels (i.e., *Srpx*, *Rpgr*, *Aff2*), and it was expected that group membership would cause differences in disease severity (i.e., PUN, protein excretion, and premature death). However, by not relating X-inactivation to disease severity, Model 1 implicitly assumes that X-inactivation does not cause disease severity and hence is not a mediator. Models 2, on the other hand, depicts that X-inactivation has a causal effect on disease severity and that the causal effect of group membership on disease severity is completely mediated by X-inactivation. Model 3 also depicts X-inactivation as a causal mediator between group membership and disease

severity but allows for the possibility that group membership may have some relationship with disease severity through some other mechanism than X-inactivation. Models 1 and 2 are nested within Model 3 such that certain parameters of Model 3 are simply fixed to zero. The Bayesian model specifications (parametric relationships, distributional assumptions and priors) are described below.

3.2.1 Modeling X-inactivation and Gene Expression

Let X_i be the latent X-inactivation for mouse i, and let G_i be an indicator that mouse i is in Group 1. The relationship between latent X-inactivation and group membership is modeled as follows,

$$X_i = \eta_0 + \eta_G G_i + u_i \ , \tag{3.1}$$

where random error u_i is assumed normally distributed with mean zero and standard deviation σ_U. The intercept, η_0 is the Group 2 mean X-inactivation, and η_G is the difference between the Group 1 and Group 2 mean X-inactivations. The mice were bred specifically to generate higher X-inactivations in Group 1, so η_G is expected to be positive. Indeed a significant positive effect for η_G is of particular interest to the scientist since it verifies that the lab measurement of X-inactivation (through the 3 gene expressions) is able to detect the experimental manipulation of X-inactivation.

Let Y_{iS}, Y_{iR}, and Y_{iA} be mouse i's gene expression measures for *Srpx*, *Rpgr* and *Aff2*, respectively. The three gene expression measures were moderately correlated (Pearson correlations: 0.48 (*Srpx*/*Rpgr*), 0.68 (*Srpx*/*Aff2*), 0.46 (*Rpgr*/*Aff2*)) in the mice that did not die prematurely). As in a common factor analysis model, it is assumed that this correlation is due to a common shared latent factor, here represented by the latent X-inactivation level, X_i. That is,

$$\begin{pmatrix} Y_{iS} \\ Y_{iR} \\ Y_{iA} \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_{0R} \\ \lambda_{0A} \end{pmatrix} + \begin{pmatrix} 1 \\ \lambda_{1R} \\ \lambda_{1A} \end{pmatrix} X_i + \begin{pmatrix} \epsilon_{iS} \\ \epsilon_{iR} \\ \epsilon_{iA} \end{pmatrix} \ . \tag{3.2}$$

The random error vector $(\epsilon_{iS}, \epsilon_{iR}, \epsilon_{iA})'$ is assumed to be multivariate normal with zero mean vector and diagonal covariance matrix with standard deviations represented σ_{XS}, σ_{XR}, and σ_{XA} respectively. Hence, given latent X-inactivation, X_i, the three gene

expression measures are assumed to be conditionally independent. For identifiability in (3.2), the *Srpx* gene expression, Y_{iS}, is assumed to be an unbiased measure of X-inactivation, i.e. its intercept is fixed to zero and its "factor loading" is fixed to 1. The choice of fixing *Srpx* over *Rpgr* and *Aff2* was arbitrary, but it allows for more direct comparison with the results from Rheault et al.[11] where the *Srpx* measure was used alone as a direct measure of X-inactivation.

3.2.2 Modeling Disease Severity

First, consider the model for protein excretion. Let Y_{i2}, Y_{i4}, and Y_{i6} be the 2-month, 4-month, and 6-month protein excretion measurements for mouse i. Let the following linear model relate protein excretion across time with group membership G_i and latent X-inactivation X_i,

$$Y_{it} = \theta_{0i} + \theta_{1i}t + \delta_{it} \tag{3.3}$$

$$\theta_{0i} = \theta_0 + \theta_{0G}G_i + \theta_{0X}(X_i - \eta_0) \tag{3.4}$$

$$\theta_{1i} = \theta_1 + \theta_{1G}G_i + \theta_{1X}(X_i - \eta_0) \tag{3.5}$$

where δ_{it} is assumed to be i.i.d normal with mean zero and standard deviation σ_L. Notice in (3.4-3.5) that the latent X-inactivation is centered at η_0 which is the mean of the Group 2 from (3.1). This parameterization allows θ_0 and θ_1 to be interpreted as the intercept and slope for a Group 2 mouse with X-inactivation at the Group 2 mean (η_0). The two groups can have different intercepts and slopes when θ_{0G} and θ_{1G} are included, and X-inactivation is allowed to affect both the protein excretion intercept and slope when θ_{0X} and θ_{1X} are included.

Based on the original experimental plan, the other measure of disease severity collected was Plasma Urea Nitrogen (PUN) at 6 months. Due to right skew, log PUN was modeled instead of PUN. The data set included up to two PUN measurements for each mouse, one from each of two lab technicians. The correlation between the two technicians' PUN measurements was very high, so the average of the two log PUN measurements was used whenever two measurements were available. Let Y_{iPUN} be the average log PUN measurement for mouse i. The following model relating latent X-inactivation and group membership to PUN was used,

$$Y_{iPUN} = \phi_0 + \phi_G G_i + \phi_X(X_i - \eta_0) + \zeta_i \tag{3.6}$$

where ζ_i was assumed to be distributed normally with mean 0 and standard deviation σ_N. Similar to the parameterization for Protein Excretion, in this model for PUN, ϕ_0 is the mean log PUN for a Group 2 mouse with X-inactivation equal to the Group 2 mean, η_0.

Finally, as was mentioned in the introduction, although unplanned, there were mice that died prematurely (prior to 6 months). Hence, premature death is also used as a measure of disease severity and provides an opportunity for incorporating into the model information about mice that had neither gene expression data nor a PUN measurement of disease severity. A simple logistic model was used to relate group membership and X-inactivation with premature death. For mouse i, let D_i be a 0-1 indicator of premature death and assume D_i follows a Bernoulli distribution with probability of premature death π_i such that

$$\log \frac{\pi_i}{1-\pi_i} = \alpha_0 + \alpha_G G_i + 100\alpha_X (X_i - \eta_0) \; . \tag{3.7}$$

The inclusion of the constant 100 in (3.7) serve two purposes: first, it reduces α_X, which otherwise could be quite large, since differences in X-inactivation between mice are less than one and second, it facilitates interpretation of α_X, since it represents the effect of a 0.01 increase in X-inactivation, which is closer to the scale of X-inactivation variability between mice. This model takes α_0 to be the logit of the probability of death for a Group 2 mouse with X-inactivation equal to η_0. This is particularly useful since, as will be discussed in the next section, it will be necessary to place an informative prior on α_0, and thus interpretability of α_0 will help select a reasonable, meaningful prior distribution.

3.2.3 Joint Model and Prior Specification

Equations (3.1 - 3.7) are combined to create the following joint model for mouse i with data $\mathbf{Z}_i$ =(Y_{iS}, Y_{iR}, Y_{iA}, Y_{i2}, Y_{i4}, Y_{i6}, Y_{iPUN}, D_i) conditioned on observed group membership G_i.

$$\begin{aligned} P(\mathbf{Z}_i|G_i) \;\; = \;\; & \int \prod_{t=2,4,6} P(Y_{it}|X_i, G_i) * P(Y_{iPUN}|X_i, G_i) * P(D_i|X_i, G_i) \\ & * \prod_{j=S,R,A} P(Y_{ij}|X_i) * P(X_i|G_i) dX_i \end{aligned} \tag{3.8}$$

Since X-inactivation is latent and hence not observed, it is marginalized out of the joint model but can also be considered a parameter within the Bayesian framework. The joint model (3.8) represents Model 3 in Figure 1 since it allows both X-inactivation (X_i) as well as Group membership (G_i) to directly effect disease severity. Model 1 and Model 2 in Figure 1 take either X_i or G_i (respectively) out of the conditional statements for the disease severity measures in (3.8).

To complete specification of the Bayesian joint model it is necessary to define the prior distributions for all parameters in (3.1 - 3.7). All prior distributions used were proper. For most parameters, it was feasible to use prior distributions with very little information value. Specifically, normal prior distributions with mean zero and standard deviation 50 were used for the regression coefficient parameters: η_0 and η_G in equation (3.1), θ_0, θ_1, θ_{0G}, θ_{1G}, θ_{0X}, and θ_{1X} in equations (3.4-3.5), ϕ_0, ϕ_G, and ϕ_X in (3.6), and λ_{0A}, λ_{0R}, λ_{1A}, and λ_{1R} in (3.2). Uniform$[0, 1]$ priors were used for σ_{XA}, σ_{XR}, and σ_{XS}, while Uniform$[0, 10]$ priors were used for σ_N and σ_L. A $\Gamma[3, 30]$ distribution was used as a prior distribution for σ_U to show a prior belief that the variability of X-inactivation from mouse to mouse may be small, but not zero. A $\Gamma[3, 30]$ distribution has mean 1/10, variance 1/300, and its 0.025 and 0.0975 quantiles are approximately 0.02 and 0.24, respectively. Without this limitation, posterior distributions can show a nonsensical mode for σ_U very near zero. The prior distribution for parameters in the model for premature death (3.7) required special attention when fitting Model 2 or 3.

Instead of a usual diffuse normal distribution for the logit intercept α_0, a Uniform$[-5.5, 5.5]$ was used corresponding to the probability of death for an average Group 2 mouse not being allowed to drop below below 0.004. Recall that the observed rate of premature death in the Group 2 was 0.06 (3/50) corresponding to a logit of -2.75. The prior distribution for α_0 needed to be somewhat informative in order to anchor the logistic model when X-inactivation was used as a predictor. Without a strong lower bound for the value of α_0, or some other informative prior distribution, the logistic model (3.7) with latent X-inactivation as a predictor resulted in α_0 becoming arbitrarily small, which in turn lead to α_X becoming very large. When this occurs, computational difficulties arise since the probability of death π_i for surviving mice essentially goes to 0 while the π_i for the mice that died prematurely goes to 1. The bounded prior distribution for α_0 allows survival status to inform X-inactivation without becoming a

perfect measure of X-inactivation by requiring nonzero probabilities of death for the surviving mice.

In Model 3 where both X-inactivation and group membership predict premature death, using vaguely specified priors for α_G and α_X (i.e. $N(0, 50)$) lead to distinctly bimodal posterior distributions for those two parameters. This phenomena and its relation to our ability to tease out independent effects of group versus X-inactivation on disease severity will be described in the results and discussion. One attempt to resolve these difficulties involved fitting a constrained version of Model 3, where the prior distributions of α_X and ϕ_X were Uniform$[0, 10]$ and Uniform$[0, 100]$, respectively. These priors restrict α_X and ϕ_X to be positive.

3.3 Results

All the models were fit using WinBUGS. The posterior distributions were approximated using 500,000 posterior samples after a burn-in period of 100,000 iterations using four 125,000 sample chains.

Table 3.1 provides a summary of the results from Models 1 and 2. In Model 1, the posterior for η_G representing the relationship between group membership and X-inactivation is centered near 0.077 and is found to be smaller than the posterior for η_G in Model 2 which is centered near 0.097. This difference comes from the way that the mice that died prematurely are handled by the two models. In Model 1, the mice that died prematurely provide no information about X-inactivation, while in Model 2, the survival status of the mice (as well as their other disease severity information) informed the posterior distribution of the X-inactivations. In Model 2, the dead mice are predicted to have higher X-inactivations, which increased the difference in mean X-inactivation between the two groups, since Group 1 had more dead mice. Figure 3.2 shows posterior distributions for the mice under Model 2, split into separate graphs by group membership and survival status. As expected, Group 1 mice tend to have higher X-inactivations than Group 2 mice. Also, dead mice tend to have higher X-inactivations than surviving mice. Hence Model 2 incorporates the mice that died prematurely. It is worthwhile to note that the posterior distributions of the X-inactivations of the dead mice tend to be broader than the distributions of the surviving mice. This is likely

explained by the fact that the dead mice had more missing measurements than the surviving mice.

In both Models 1 and 2, similar findings are found regarding the relationship between either Group or X-inactivation and the Protein Excretion and PUN measurements. In particular being in Group 1 (in Model 1) or increases in X-inactivation (in Model 2) are both positively related to increases in the PUN measurement (indicating worse disease), but the posterior 95% credible intervals for their effects on Protein Excretion in both Models 1 and 2 contain zero indicating no significant relation with Protein Excretion. There is some indication of a relationship between protein excretion slope and group membership in Model 1, but θ_{1G} is not clearly different from zero (95% Credible Interval: $(-0.544, 0.461)$). In Model 2, there is also some weak evidence of an effect of X-inactivation of protein excretion slope, but θ_{1X} is also not significantly different from zero (95% Credible Interval: $(-3.505, 3.27)$).

The survival submodel for Model 1 is a straightforward Bayesian logistic regression on group membership, hence as expected, the posterior distribution for α_0 is centered near the logit of the rate of death in group 2 which is $logit(5/30) = -2.75$. The credible interval for the log odds ratio α_G in Model 1 is significantly positive indicating an increased odds of death in Group 1. In Model 2, when latent X-inactivation is included as a predictor of premature death it was necessary to specify a somewhat informative prior on α_0. Since latent X-inactivation is not observed, it is not possible to explicitly condition on it in the same way that conditioning on group membership (which is observed) can be done. This means that in Model 2, the disease severity measures, in addition to the gene expression variables, provide information about X-inactivation. For the mice that died prematurely, they have missing data on the gene expression variables and PUN, so X-inactivation is informed only by the Group status, one or two Protein Creatinine measurements and the indicator that they died prematurely. Because of this confounding between the survival status and missing data, the parameters in the logistic survival model are very weakly identified without some prior information. Figure 3.3 shows the posterior distributions for α_0, α_G and α_X for Models 1, 2, and 3 when an informative Uniform$[-5.5, 5.5]$ prior is used for α_0 and a diffuse prior is used for both α_G and α_X. Focusing still on Model 2, it can be sees that the prior clearly has an effect on limiting the posterior of α_0, but that the data have provided strong information for

the posterior of the log odds ratio α_X which has posterior mean 0.383 and 95% credible interval $(0.190, 0.644)$ indicating a significant increased odds of premature death for increased X-inactivation.

Now consider Model 3, whose results are summarized in Table 3.2. Given the same priors used for Model 1 and 2, the posterior in Model 3 of α_G is clearly bimodal (Figure 3.3), with one positive mode and another mode near zero and the posterior for α_X is bimodal, with one positive and one negative mode. There is a strong negative correlation (-0.935) between the posterior samples of α_G and α_X for Model 3, such that the positive mode for α_G is related to the negative mode for α_X and the mode near zero for α_G corresponds to the positive mode for α_X. The results for Model 3 also included bimodal posteriors of X-inactivations for the mice that died prematurely (not shown). In Model 3, the distribution of η_G peaks near 0.08, but there is substantial distributional weight near 0.05. This corresponds to the bimodality in α_X, such that when α_X is large, η_G tends to be large also, but when α_X is negative, η_G tends to be small. Finally, while no clear bimodality is detected for the parameters in the Protein Creatinine and PUN outcome submodels in Model 3, there is a strong negative correlations between ϕ_G and ϕ_X (-0.83), as well as between θ_{1G} and θ_{1X} (-0.63). The posterior distributions of ϕ_G, ϕ_X, θ_{1G}, and θ_{1X} are shown in Figure 3.4 for each of the four models.

In addition to the bimodality and high collinearity of the posterior results from Model 3, the effective degrees of freedom, pD, used in calculating DIC is negative. Clearly there is something problematic with Model 3 such that the parameters are only weakly identified. Carefully inspecting the bimodality, it is clear that the data cannot differentiate between two distinct explanations. In one case, high X-inactivations are harmful to survival, and no additional group effects on survival is present (i.e. full mediation of the group effect by X-inactivation on survival). In this case, the mice that died prematurely tend to have high X-inactivations (high posterior means for their latent X-inactivation). In the second case, membership in Group 1 is very harmful for survival and also low X-inactivations are harmful to survival. In this second case, the posterior distributions of X-inactivation for the mice that died prematurely to be very low. The first case is reasonable, the second case is not biologically plausible. Given that the data cannot discern between these two cases, an informative prior was added to force α_x to be positive. Because X-inactivation appears to be a strong predictor of

PUN in Model 2, ϕ_X was also constrained to be positive in Model 3-Constrained so that the effect of group on PUN could be examined above and beyond the effect of X-inactivation.

Model 3–Constrained results are summarized in Table 3.2. The posterior mean for η_G is 0.093, similar to the result from Model 2. The 95% credible intervals for both α_G and ϕ_G contain zero, suggesting that there is insufficient evidence to conclude that group membership affects PUN or survival other than through its effect on X-inactivation. Model 3–Constrained maintains many of the features of Model 2, but contains more parameters. Despite the fact that all the additional parameters contained in Model 3–Constrained are not significantly different from zero, the DIC for Model 3–Constrained (230.8) is lower than the DIC for Model 2 (249.4). The DIC criterion suggests that Model 3–Constrained should be the preferred model.

3.4 Discussion

The Bayesian joint model framework allows the consideration of the joint posterior distributions of the X-inactivations of the eighty-eight mice and the various model parameters. It is important to consider the posterior distributions jointly, since the distribution of the X-inactivations is needed to infer the model parameters and the model parameters are needed to infer the latent X-inactivations.

Model 1 describes a situation where none of the outcomes including survival can be used to infer the X-inactivation values for the dead mice. As a result, η_G is very close to the mean difference in *Srpx* measurements among the surviving mice. The other models all have substantial posterior support for a larger mean difference in X-inactivation between the two groups, that is, η_G near 0.09, as shown in Figure 3.3. Adding α_X to the models, allows mouse survival to influence X-inactivations, and that influence is notable in both the posterior distribution of η_G and the posterior distributions of the X-inactivations of the dead mice as shown in Figure 3.2.

The evidence for an effect of either group membership or X-inactivation on protein excretion appears weak. Neither θ_{1G} nor θ_{1X} is clearly different from zero. The relationships between X-inactivation or group membership with log PUN are not clear. Model 1 shows that group membership can be a significant predictor of log PUN, and

Model 2 shows that X-inactivation can be a significant log PUN predictor. Model 3, however, suggests that the data are not sufficient to distinguish the group membership effect from the X-inactivation effect on log PUN. Both group membership and X-inactivation can be good predictors of log PUN, but it is unclear which is better. Model 3–Constrained shows that when the model is forced to have a positive relationship between X-inactivation and log PUN (i.e., positive ϕ_X), the marginal effect of group membership on log PUN is small. The posterior distributions of ϕ_G, ϕ_X, θ_{1G}, and θ_{1X} are shown in Figure 3.4 for each of the four models.

Perhaps the role of group and X-inactivation on the survival outcome is clearest. Model 1 shows that group membership can be a significant predictor of survival. Model 2 shows that X-inactivation can also serve as a significant survival predictor. The bimodal nature of so many of the posterior distributions from Model 3 deserves detailed examination.

Effectively, one mode corresponds to the effect that the investigators expected. Group 1 mice were bred specifically to have higher X-inactivations, and that higher X-inactivation would result in greater expression of the mutant allele. Greater expression of the mutant allele was expected to produce more severe disease. Although the investigators did not expect mice to die, it is reasonable to expect that the mice with the most severe disease would have the highest risk of death. Also, except for the mutant *Col4a5* allele, there was no reason to expect Group 1 mice would have more health problems than Group 2 mice. That is, the positive mode for α_X and the α_G mode near zero correspond to the expected results of the experiment.

The other modes in Model 3 suggest that either there were key problems with the implementation of the experimental design, or that the mouse model for XLAS is seriously flawed. When α_X is negative, high expression of the mutant *Col4a5* allele is protective, even though it should result in more severe disease and higher risk of death. Values of η_G near 0.05 suggest that Group 1 mice did not show much preference for activation of the disease allele carrying X-chromosome. Although Model 3 shows that this possibility is consistent with the data, there is no other reason to believe that the design or implementation of the experiment was incorrect. Although the negative mode for α_X does appear in other models not discussed in this chapter, it is only in Model 3 that the negative mode for α_X is large. Since Model 3 includes group membership

and X-inactivation as correlated predictors throughout the model, it is the model that is structured to have the greatest uncertainty about X-inactivation and its effects. This suggests that the counterintuitive modes in Model 3 are more likely the result of an ambiguous model structure than an indication of experimental problems.

It is also important to note that the bimodal posteriors for Model 3 do not represent a conflict between the results of Models 1 and 2. The first scenario for Model 3 is indeed much like the results from Model 2, but the second scenario is not like the results from Model 1. Model 3 suggests that the effect of X-inactivation on survival is nearly always important. The first case describes a harmful effect of high X-inactivation and the second scenario describes a somewhat nonsensical protective effect for high X-inactivation. The posterior distribution of α_X for Model 3 places little density near zero. Model 3 suggests that survival is mediated by X-inactivation.

The Bayesian joint modeling strategy used in this chapter was clearly successful in achieving the goal of performing inference using X-inactivation itself, rather than using a gene expression measure as an X-inactivation index. Of course, since X-inactivation was not measured directly, it was a latent variable. Yet this approach allows all of the available measurements, not just a single gene expression, to inform the posterior distribution of each mouse's X-inactivation. This approach also allows the effect of X-inactivation on survival to be examined. Measurements of gene expression were missing for all the dead mice, which made performing a standard frequentist logistic regression or survival analysis impossible. Model 2 and Model 3–Constrained both demonstrate that there is a strong relationship between the modeled risk of death and X-inactivation.

This chapter's modeling strategy also was successful in using the data from all of the mice, rather than just the survivors, for inference. The effect of including the dead mice is clearest in the posterior distributions of η_G. In Model 1, there are no non-missing measurements that can inform the distributions of the X-inactivations for the dead mice. In the other models, non-missing measurements are used to infer the X-inactivations of the dead mice. All of those models place substantial posterior weight on values of η_G around 0.09, suggesting that η_G estimates based only on the data for surviving mice are biased.

Outcome	Parameter	Model 1	Model 2
X-Inactivation	η_0	0.458 (0.433, 0.482)	0.468 (0.442, 0.494)
	η_G	0.077 (0.035, 0.122)	0.097 (0.055, 0.143)
	σ_U	0.066 (0.046, 0.089)	0.068 (0.043, 0.094)
Gene Expression	λ_{0A}	−0.015 (−0.275, 0.175)	−0.044 (−0.352, 0.153)
	λ_{0R}	0.033 (−0.251, 0.275)	−0.001 (−0.347, 0.268)
	λ_{0S}	0	0
	λ_{1A}	1.101 (0.707, 1.640)	1.162 (0.753, 1.801)
	λ_{1R}	1.010 (0.512, 1.597)	1.081 (0.527, 1.796)
	λ_{1S}	1	1
	σ_{XA}	0.053 (0.019, 0.075)	0.057 (0.031, 0.077)
	σ_{XR}	0.109 (0.087, 0.135)	0.110 (0.088, 0.136)
	σ_{XS}	0.056 (0.029, 0.075)	0.059 (0.041, 0.078)
Protein Creatinine	θ_0	1.734 (1.408, 2.058)	1.685 (1.412, 1.958)
	θ_1	0.269 (0.195, 0.343)	0.287 (0.223, 0.350)
	θ_{0G}	−0.041 (−0.544, 0.461)	–
	θ_{1G}	0.065 (−0.055, 0.185)	–
	θ_{0X}	–	−0.115 (−3.505, 3.270)
	θ_{1X}	–	0.587 (−0.261, 1.524)
	σ_L	0.716 (0.653, 0.787)	0.704 (0.639, 0.776)
PUN	ϕ_0	3.722 (3.542, 3.903)	3.822 (3.648, 3.990)
	ϕ_G	0.468 (0.156, 0.780)	–
	ϕ_X	–	3.742 (1.144, 7.026)
	σ_N	0.618 (0.523, 0.737)	0.606 (0.505, 0.728)
Premature Death	α_0	−2.914 (−4.332, −1.816)	−4.464 (−5.479, −2.692)
	α_G	2.360 (1.055, 3.904)	–
	α_X	–	0.383 (0.190, 0.644)
Fit Statistic	DIC	248.6	249.4

Table 3.1: Posterior means and 95% credible intervals for Model 1 and Model 2.

Outcome	Parameter	Model 3	Model 3–Constrained
X-Inactivation	η_0	0.461 (0.427, 0.492)	0.469 (0.443, 0.496)
	η_G	0.075 (0.017, 0.136)	0.093 (0.048, 0.142)
	σ_U	0.068 (0.043, 0.096)	0.067 (0.036, 0.096)
Gene Expression	λ_{0A}	−0.067 (−0.411, 0.155)	−0.080 (−0.431, 0.154)
	λ_{0R}	0.003 (−0.341, 0.273)	−0.018 (−0.377, 0.268)
	λ_{0S}	0	0
	λ_{1A}	1.209 (0.748, 1.919)	1.236 (0.751, 1.959)
	λ_{1R}	1.072 (0.517, 1.785)	1.117 (0.527, 1.857)
	λ_{1S}	1	1
	σ_{XA}	0.052 (0.014, 0.076)	0.053 (0.011, 0.080)
	σ_{XR}	0.110 (0.088, 0.137)	0.110 (0.087, 0.137)
	σ_{XS}	0.059 (0.037, 0.079)	0.060 (0.039, 0.082)
Protein Creatinine	θ_0	1.729 (1.399, 2.059)	1.723 (1.397, 2.049)
	θ_1	0.272 (0.196, 0.347)	0.274 (0.199, 0.349)
	θ_{0G}	−0.022 (−0.713, 0.617)	−0.135 (−0.836, 0.531)
	θ_{1G}	0.042 (−0.106, 0.190)	0.045 (−0.118, 0.201)
	θ_{0X}	−1.228 (−7.229, 4.417)	0.578 (−4.247, 5.650)
	θ_{1X}	0.483 (−0.670, 1.693)	0.419 (−0.780, 1.727)
	σ_L	0.710 (0.645, 0.782)	0.708 (0.641, 0.781)
PUN	ϕ_0	3.729 (3.539, 3.920)	3.746 (3.561, 3.935)
	ϕ_G	0.317 (−0.075, 0.710)	0.252 (−0.184, 0.651)
	ϕ_X	2.094 (−1.094, 5.545)	3.176 (0.221, 7.542)
	σ_N	0.610 (0.510, 0.730)	0.603 (0.499, 0.724)
Premature Death	α_0	−4.463 (−5.481, −2.644)	−4.483 (−5.481, −2.671)
	α_G	2.107 (−2.786, 7.388)	−0.287 (−3.887, 2.264)
	α_X	0.083 (−0.586, 0.714)	0.433 (0.103, 0.894)
Fit Statistic	DIC	–	230.8

Table 3.2: Posterior means and 95% credible intervals for Model 3 and Model 3–Constrained.

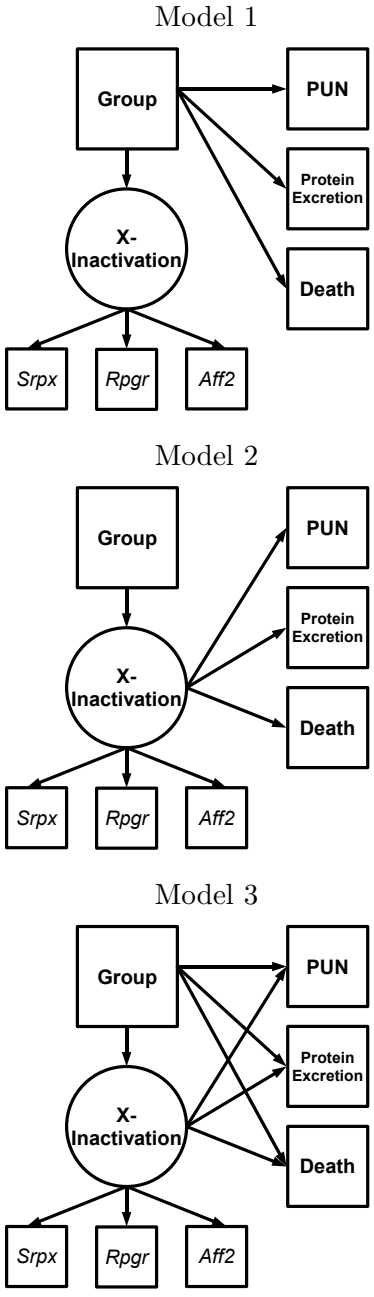

Figure 3.1: Diagram of the general modeling framework. The Group effect on protein excretion, PUN, and survival may be mediated by X-inactivation, which is measured by *Srpx*, *Rpgr* and *Aff2* gene expressions.

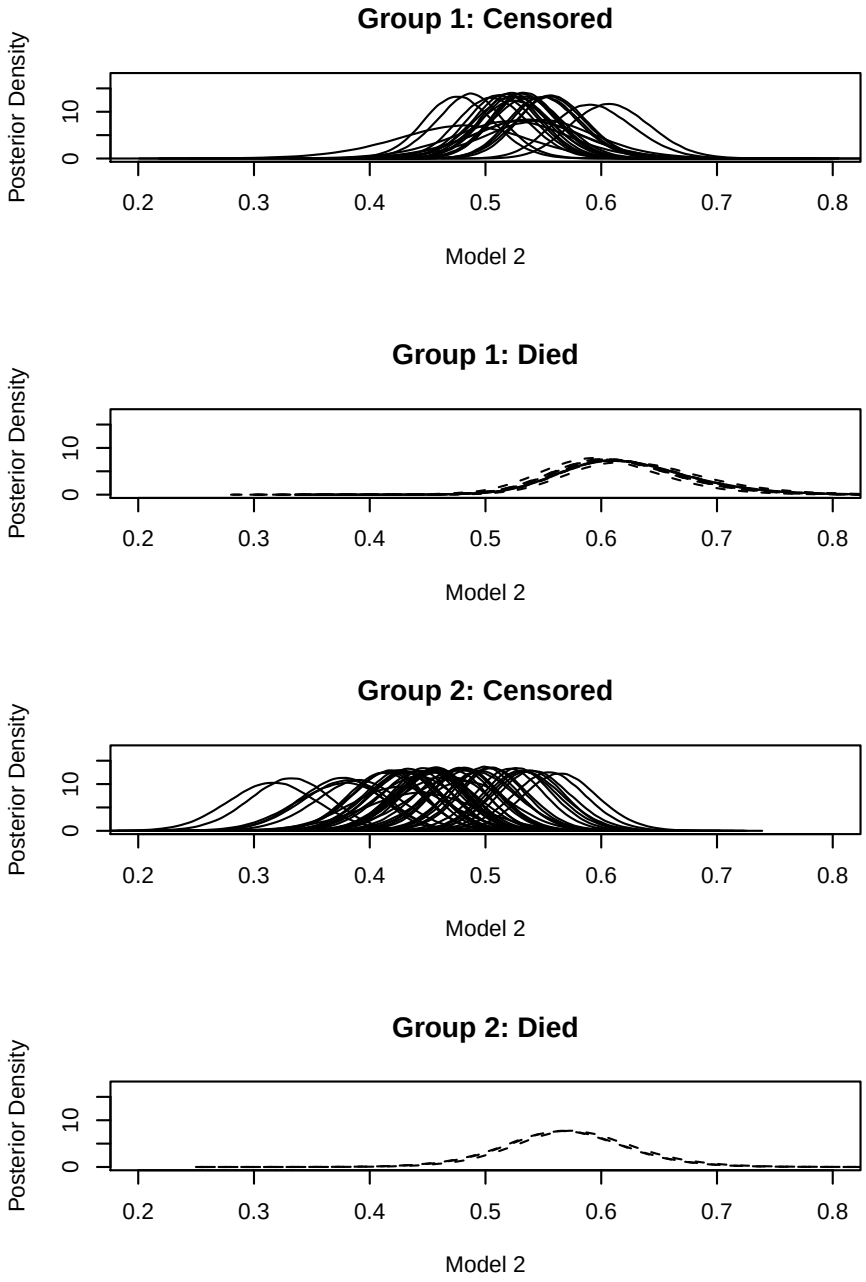

Figure 3.2: Posterior distributions of X-inactivations from Model 2.

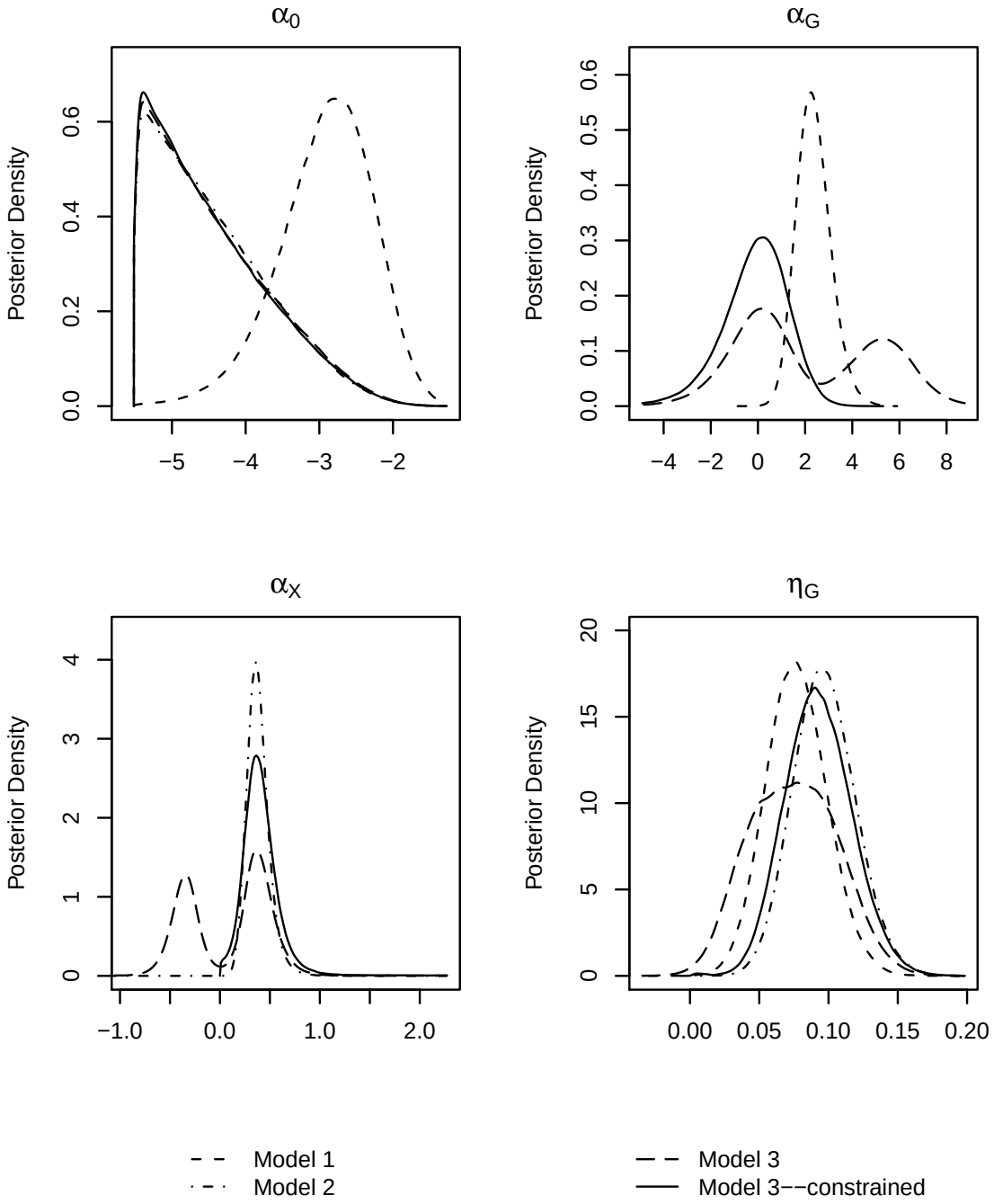

Figure 3.3: Posterior distributions for α_0, α_G, α_X, and η_G for the four models.

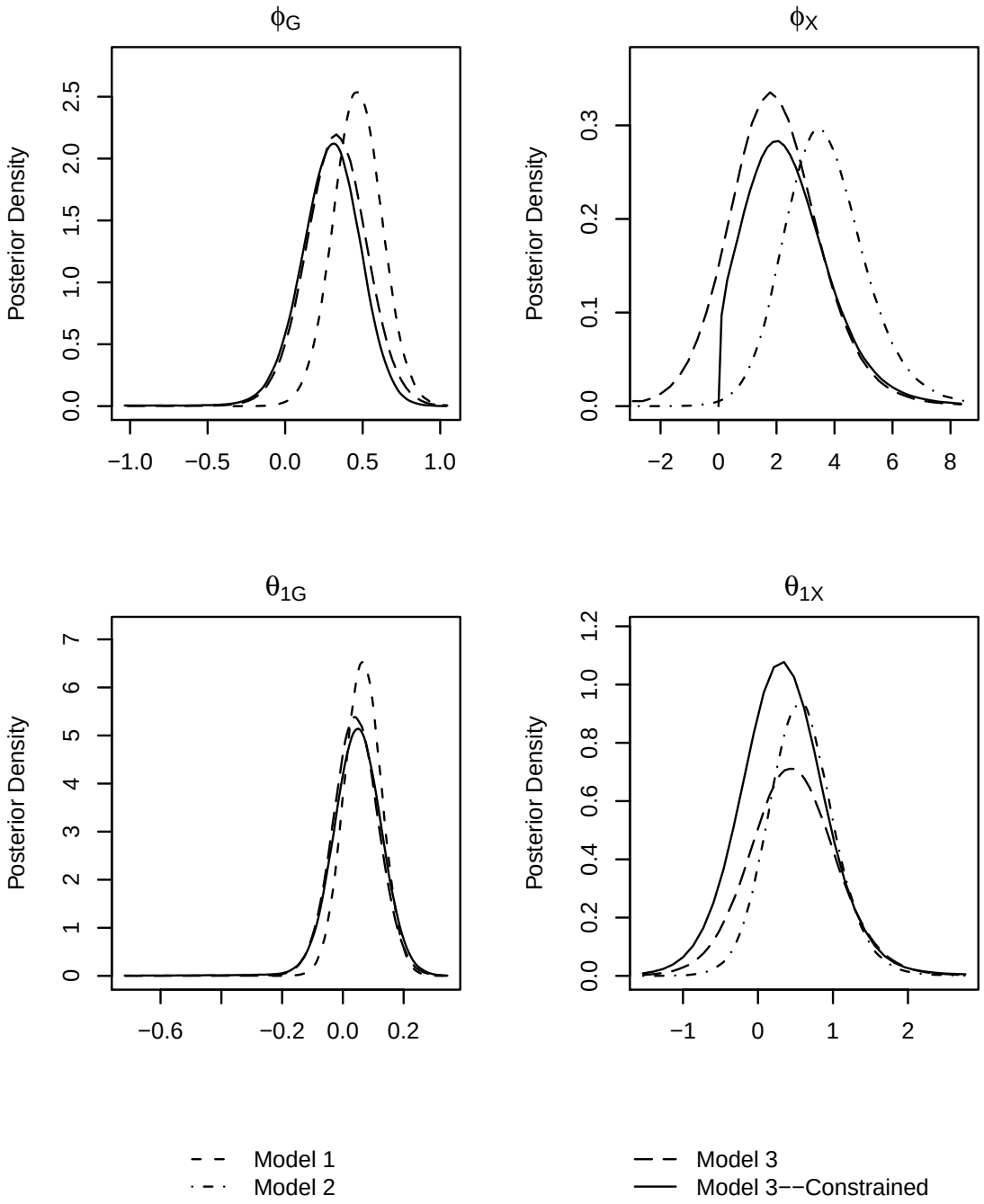

Figure 3.4: Posterior distributions for ϕ_G, ϕ_X, θ_{1G}, and θ_{1X} for the four models.

Chapter 4

Semiparametric Proportional Hazards Regression with Vector-Valued Time Scales

4.1 Introduction

Cox regression models[13] are among the most commonly used models for survival analysis. Like other survival models, these models require a time scale, though the time scale may not technically be a time. For example, Aalen et al.[14] provide an example of an analysis using Cox regression using "drive for thinness" as a time scale using data from Skårderud et al.[15]. More typical choices of time scale include the time since observation began (time on study), age, and the time since a fixed date (chronological time).

The baseline hazard is assumed to be a function of time as measured by the time scale. The time scale used for the analysis, therefore, should be related to the hazard. Often, however, there is more than one time scale available. In a typical clinical trial, the date that each participant enters the study is recorded, along with his or her age. The date that each participant fails or is censored is also recorded. Time on study, age, and chronological time are all possible time scale choices. Time on study is typically used for clinical trials, but the hazard may be more strongly related to age than to time

on study.

Other study designs face similar time scale choices. Kom et al.[16] conclude that age is a better choice of time scale than time since the survey (i.e.,time on study) for health survey data. Thiebaut and Benichou[17] used simulations to demonstrate that age is preferable to time on study for epidemiological cohort data. The choice of time scale is not limited to human health studies. The time to failure for an automobile part could be modeled using either time since manufacture (i.e.,age) or the cumulative distance the automobile traveled.

Suppose that the hazard is a function of more than one time scale. In a clinical trial, time on study and age may both be related to the baseline hazard. The risk of failure for an automobile part may be a function of both age and mileage. In many analyses, one time scale is chosen, and other potentially important time scales are treated as time-varying covariates or are discretized and used as stratification variables. Both of these approaches have limitations. In order to use a time scale as a time-varying covariate, a functional form must be found to satisfy the proportional hazards assumption. It may be very difficult to find such a function. Choosing too many strata could reduce efficiency.

Another approach is to create a composite time scale. Farewell and Cox[18] suggest a procedure to generate a single time scale from two or more available time scales. Oakes[19] suggests combining multiple time scales and proposes a collapsibility condition to evaluate whether the combined scale is fully informative. Gertsbakh and Kordonsky[20], Kordonsky and Gertsbakh[21], and Duchesne and Lawless[22][23] developed procedures to find composite time scales. Composite time scales, however, may not be easy to interpret.

Still another approach is to parametrically model the hazard as a function of more than one time scale. Efron[24] develops a two-way proportional hazards model using a Poisson generalized linear model. This approach is well-suited to situations where the hazard itself is a key focus of investigation. When the effects of the covariates are of primary interest, the Cox regression model is often preferred because it allows the baseline hazard function to be treated as a nuisance. Yet traditional Cox regression models use only a single time scale variable, forcing other time scales to be modeled

parametrically instead of as part of the nuisance baseline hazard. The goal of developing a vector-valued time scale extension of Cox proportional hazards regression is to allow more covariates to be included in the baseline hazard to avoid the parametric assumptions about the functional form of the relationship of covariates to the hazard.

Cox[13], in his classic paper, suggests generalizing his semiparametric proportional hazards model to vector-valued time, but does not explore the problem. This chapter will extend these models to vector-valued time scales by providing one possible strategy for defining tied events and risk sets. In Section 4.2, I will define a vector-valued time scale, define risk sets, and explain how to fit a vector-valued time scale Cox regression model using standard software. An analysis of a data set will be shown in Section 4.3. Section 4.4 will discuss simulations and results to compare this chapter's modeling approach to other modeling options, including Cox regression models with time-varying covariates and stratified models. Discussion will follow in Section 4.5.

4.2 Vector-valued Time Scales and Associated Risk Sets

Survival data is typically recorded according to chronological time. Typically, observation of unit i begins at time s_i and continues until time e_i, and δ_i indicates whether unit i failed at e_i. On the chronological time scale, s_i and e_i are specific dates and times. Duchesne and Lawless[22] describe a time scale as non-negative function of time and covariates that is a non-decreasing function of chronological time for all possible covariates. Simply, a time scale is a functional that can increase in value over chronological time, but never decreases. Let $\tau_i(t)$ be the time scale function for unit i. Then $\tau_i(t) \leq \tau_i(t')$ whenever $t \leq t'$. The i subscript is included to take into account unit i's covariates, which are assumed to be external. Since units i and i' can have different covariates $\tau_i(t)$ may not equal $\tau_{i'}(t)$. Returning to the automobile part failure example, the cumulative distance traveled by automobile i is a non-decreasing function of chronological time, though chronological time alone is not enough to determine the cumulative distance traveled for any particular auto. The position of the car at each time must be also be known. The cumulative distance traveled cannot decrease as chronological time increases, and two different autos can have different cumulative distances traveled at any particular date and time on the chronological time scale.

An r-vector time scale is a vector whose elements are scalar time scales. If $\vec{\tau}_i(t)$ is a r-vector time scale for unit i, then there are r scalar time scales $\tau_{1i}(t), \tau_{2i}(t), \ldots, \tau_{ri}(t)$ such that $\vec{\tau}_i(t) = (\tau_{1i}(t), \tau_{2i}(t), \ldots, \tau_{ri}(t))'$.

Vector-valued time scales present both philosophical and practical difficulties in survival analysis. On a scalar time scale, a unit either survives until, or fails before, a particular time. On a vector-valued time scale, a unit's survival status can be different for the different component time scales. If, for example, the vector-valued time scale is composed of age and time on study, a unit may survive beyond a particular point on the age scale, but not the time on study scale. Another difficulty involves tied times. On a scalar time scale, it is common for several units to be at risk at precisely the time of an event. On a vector-valued time scale, it is often very unlikely for other units to be at risk at exactly the vector-valued time that a unit fails. This chapter will suggest that a well-planned weighted Cox proportional hazards regression model can address these challenges. The current method focuses on models that include only baseline (i.e.,not time-varying) covariates. Extensions which include time-varying covariates are discussed for future research.

Cox proportional hazards regression models[13] are among the most popular models used for survival analysis. Following the notation of Grambsch and Therneau[25], each unit is an independent counting process $\{N_i(t), t \geq 0, i = 1, \ldots, n\}$ with the following intensity function:

$$Y_i(t) \exp\left\{\vec{\beta}'\vec{z}_i\right\} d\Lambda_0(t), \tag{4.1}$$

where $Y_i(t)$ is the at-risk process for the ith subject, $\vec{\beta}$ is a p-vector of regression parameters, and $\vec{z}_i$ is a p-vector of baseline covariates, and $d\Lambda_0(t)$ is the baseline hazard function. $Y_i(t)$ is equal to 1 when the ith subject is at risk, and is zero otherwise. Coefficient estimates for $\vec{\beta}$ are usually found by maximizing the log partial likelihood:

$$\log L_p = \sum_{i=1}^{n} \int_0^{\infty} \left[Y_i(t)\vec{\beta}'\vec{z}_i - \log\left\{ \sum_{j=1}^{n} Y_j(t) \exp\left[\vec{\beta}'\vec{z}_j\right] \right\} \right] dN_i(t). \tag{4.2}$$

The log partial likelihood for a Cox proportional hazard regression model depends on the at-risk process $Y_i(t)$ and the counting processes $N_i(t)$ for each unit, as well as the values of the covariate vector for each unit at each event time.

In order to generalize Equation 4.2 to vector-valued time scales, $Y_i(t)$ needs to be generalized. But first, it is useful to consider some intermediate functions. An indicator for the time scales on which a unit i survives beyond time $\vec{t}$ is needed to define the at-risk process. Define $a_{ki}\left(\vec{t}\right)$ as follows:

$$a_{ki}\left(\vec{t}\right) = \mathrm{I}\left\{\tau_{ki}(e_i) \geq t_k\right\} \tag{4.3}$$

where $\vec{t} = (t_1, t_2, \ldots, t_r)'$ is any r-vector in $\mathcal{R}^{r+}$. $a_{ki}\left(\vec{t}\right)$ indicates that unit i is at risk after t_k on the kth time scale. Similarly, an indicator that unit i was at risk before time $\vec{t}$ is needed. Now define:

$$b_{ki}\left(\vec{t}\right) = \mathrm{I}\left\{\tau_{ki}(s_i) \leq t_k\right\} \tag{4.4}$$

In this way $b_{ki}\left(\vec{t}\right)$ indicates that unit i is at risk before t_k on the kth time scale. If $a_{ki}\left(\vec{t}\right) = b_{ki}\left(\vec{t}\right) = 1$, then there must be some time t where $\tau_{ki}(t) = t_k$. When $a_{ki}\left(\vec{t}\right) = b_{ki}\left(\vec{t}\right) = 1$ define:

$$c_{ki}\left(\vec{t}\right) = \max\left\{t \in [s_i, e_i] \text{ such that } \tau_{ki}(t) \leq t_k\right\} \tag{4.5}$$

That is, $c_{ki}\left(\vec{t}\right)$ is the last chronological time when unit i is at risk on the kth time scale.

If the baseline hazard is a uniformly continuous function on the vector-valued time scale, then the hazard for nearby points will be approximately the same, by definition. In a broader sense, even if uniform continuity does not hold, it is often reasonable to assume that nearby points of vector-valued time have similar hazards. This suggests that including units that are at risk near an event time may be one approach to generating plausible risk sets when the only units at risk at precisely the event times are the failed units themselves. Let $d_i\left(\vec{t}, t\right)$ be an indicator function that equals 1 whenever unit i is near $\vec{t}$ at chronological time t. One possible definition for $d_i\left(\vec{t}, t\right)$ is:

$$d_i\left(\vec{t}, t\right) = \begin{cases} \prod_{k=1}^{r} \mathrm{I}\left\{\mid \tau_{ki}(t) - t_k \mid \leq \gamma_k\right\} & t \text{ defined} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

where $\gamma_1, \gamma_2, \ldots, \gamma_r$ are nonnegative constants. In this definition, $d_i\left(\vec{t}, t\right)$ is equal to one only if $\vec{\tau}_i(t)$ is close to $\vec{t}$ on all the scalar time scales.

Now, the raw weights can be defined. The raw weights describe the fraction of the r time scales on which a unit is at risk and nearby. Let $W_i^*(\vec{t})$ be defined:

$$W_i^*\left(\vec{t}\right) = \frac{1}{r}\left[\prod_{k=1}^{r} b_{ki}\left(\vec{t}\right)\right] \sum_{k=1}^{r} a_{ki}\left(\vec{t}\right) d_i\left(\vec{t}, c_{ki}\left(\vec{t}\right)\right) \tag{4.7}$$

The at-risk indicator can now be defined:

$$Y_i^*\left(\vec{t}\right) = \mathrm{I}\left\{W_i^*\left(\vec{t}\right) > 0\right\} \tag{4.8}$$

$Y_i^*\left(\vec{t}\right)$ is equal to one if unit i is at risk before $\vec{t}$ on all the times scales and unit i survives to $\vec{t}$ or beyond on at least one time scale near $\vec{t}$. Figure 4.1 provides some example trajectories to demonstrate the risk set definition.

Generalizing the counting process, define:

$$N_i^*\left(\vec{t}\right) = \max\left(N_{1i}(t_1), N_{2i}(t_2), \ldots, N_{ri}(t_r)\right), \tag{4.9}$$

where $N_{ki}(s_k)$ is the counting process for the kth time scale in the vector-valued time scale vector. The events for units j and j' are tied when $\vec{\tau}_j(e_j) = \vec{\tau}_{j'}(e_{j'}$ and $\delta_j = \delta_{j'} = 1$.

Substituting the generalizations of the at-risk process and counting process into the log partial likelihood in equation 4.2 produces:

$$\log L_p^* = \sum_{i=1}^{n} \int_{\mathcal{R}^{r+}} \left[Y_i^*\left(\vec{t}\right)\vec{\beta}'\vec{z}_i - \log\left\{\sum_{j=1}^{n} Y_j^*\left(\vec{t}\right)\exp\left[\vec{\beta}'\vec{z}_j\right]\right\}\right] dN_i^*\left(\vec{t}\right). \tag{4.10}$$

This quantity has the same form as the standard partial likelihood, only the definitions of the at-risk process and the counting process have changed. As a result, this function can be maximized using the same algorithms used to maximize the log partial likelihood (Equation 4.2).

It is worth noting that the standard log partial likelihood is returned when $r = 1$, that is when $\vec{t} = t_1$. In one dimension, the at-risk process $Y_i^*\left(\vec{t}\right)$ is only equal to one when the raw weight $W_i^*\left(\vec{t}\right)$ is equal to one. The raw weight is only equal to one if $a_{1i}\left(\vec{t}\right) = 1$, $b_{1i}\left(\vec{t}\right) = 1$ and $d_i\left(\vec{t}, c_{1i}\left(\vec{t}\right)\right) = 1$. If unit i is at risk at t then $a_{1i}\left(\vec{t}\right) = 1$ and $b_{1i}\left(\vec{t}\right) = 1$. Also, $d_i\left(\vec{t}, c_{1i}\left(\vec{t}\right)\right) = 1$, since any at risk individual must pass through the exact event time on the only time scale, $\mid \tau_{1i}(c_{1i}) - t_1 \mid = 0 \leq \gamma_1$.

Weights will be used to include the number of dimensions in which a unit survives beyond a time point. Functions to fit Cox regression models, such as the `coxph` function in R [26], often allow different weights to be applied to different observations. This chapter will consider weights of the form:

$$w_i^*\left(\vec{t}, v\right) = \left[W_i^*\left(\vec{t}\right)\right]^v \tag{4.11}$$

where $v > 0$. Consider the limit of $w_i^*\left(\vec{t}, v\right)$ as v increases to $+\infty$. All the weights will go to zero, except for observations where $w_i^*\left(\vec{t}, v\right) = 1$. The limit of $w_i^*\left(\vec{t}, v\right)$ as v decreases to 0^+ will go to 1 for all observations that survive beyond $\vec{t}$ near $\vec{t}$ in any dimension.

The `coxph` function in R can fit Cox regression models for left-truncated right-censored data using counting process coded data sets. This function can easily be adapted to fit the vector-valued time scale generalization by creating a specially coded data set for the analysis. Let $\left\{\vec{\tau}_{(1)}(e_{(1)}), \vec{\tau}_{(2)}(e_{(2)}), \ldots, \vec{\tau}_{(q)}(e_{(q)})\right\}$ be the q unique observed failure time vectors in any order. If unit i is at risk at $\vec{\tau}_{(j)}(e_{(j)})$, create a record where start=$j-1$, stop=j,
the failure indicator $\delta_i = \mathrm{I}\left\{\vec{\tau}_i(e_i) = \vec{\tau}_{(j)}(e_{(j)})\right\}$, $w_i^*\left(\vec{\tau}_{(j)}(e_{(j)}), v\right)$, and the covariate vector z_i. Since the covariate values and the risk sets defined by this counting process coded data set are the same as the covariate values and risk sets used in equation 4.10, `coxph` will be able to generate the appropriate parameter estimates.

4.3 Example

The Mayo Clinic Primary Biliary Cirrhosis Data provides an opportunity to apply a vector-valued time scale analysis to a real data set (Therneau and Grambsch[27]). This data set, distributed as part of the survival package of R[28], contains data of 312 primary biliary cirrhosis (PBC) patients involved in a randomized placebo-controlled trial of D-penicillamine conducted at the Mayo Clinic between 1974 and 1984. The age of each participant is included in the data set, which allows a vector-valued time scale analysis using both age and time since randomization as the two time scales, using $v = 1$. This analysis will be compared to a more conventional analysis on the time since randomization time scale. In this case, including baseline age or including time-varying age as covariates are equivalent. Only a subset of the possible covariates were used in the analyses shown here.

Table 4.1 shows the results from the two analyses. Hazard ratios are generally quite similar for the two analyses. Since there is such broad agreement between the two analyses, there is little evidence for the superiority or inferiority of either. Of course, if the effect of age on the hazard is of particular interest to investigators, then the

conventional analysis allows statistical inference on that particular question. If the age effect is included to improve inference on other parameters, then both analyses appear to be roughly equivalent. The agreement between the two fitted models suggests that the effect of age was modeled well parametrically in the conventional analysis.

4.4 Simulations

All the simulations involved two time scales, which will be labeled scale 1 and scale 2. For convenience, the time units for each scale will be the same, and will be called "years." In each simulation, half the units were assigned to each of two hypothetical treatments, which were coded as -1 and 1. Each unit also had a continuous covariate X which was generated from a standard normal distribution. Four baseline hazard functions were considered:

$$\lambda_A(t_1, t_2) = 2 \times \left(\phi\left(\sqrt{t_1^2 + t_2^2} - 3 \right) + \phi\left(\sqrt{t_1^2 + t_2^2} - 7 \right) \right) \tag{4.12}$$

$$\lambda_B(t_1, t_2) = (4 - (t_1 - t_2)) \times \phi(t_2 - 3) + (t_1 - t_2) \times \phi(t_2 - 7) \tag{4.13}$$

$$\lambda_1(t_1, t_2) = 2 \times (\phi(t_1 - 6) + \phi(t_1 - 8)) \tag{4.14}$$

$$\lambda_2(t_1, t_2) = 2 \times (\phi(t_2 - 3) + \phi(t_2 - 7)) \tag{4.15}$$

In each of these equations, $\phi()$ is the probability density function for the standard normal distribution.

In each simulation scenario, 5000 data sets of 250 observations each were analyzed, under both the null hypothesis ($\beta_{trt} = \beta_X = 0$) and an alternative ($\beta_{trt} = \beta_X = 0.25$). Ten regression models were fit for each data set. Three models used scale 2 as the time scale. One of the three ignored scale 1. Another included scale 1 as a time-varying covariate, and the third was stratified on a trajectory variable that was relevant to the hazard. Two models used scale 1 as the time scale. One of the two ignored scale 2, and the other included scale 2 as a time-varying covariate. Five regression models used a two-dimensional time scale. Each of these five models used a different value

for v, the exponent that modifies the raw weight: $+\infty$, 2, 1, 0.5, and 0^+. In each two-dimensional time scale, a unit was considered nearby if it was less than 0.5 years away. Each simulation scenario also featured censoring which followed a exponential distribution with rate chosen to achieve approximately 55% censored observations.

In the first simulation scenario, all units start at the time origin and follow linear trajectories at different angles from the scale 1 axis using baseline hazard $\lambda_A(t_1, t_2)$. This baseline hazard depends only on the distance from the origin, and it peaks at distances of 3 and 7 units from the origin. The trajectory variable used was the angle from the scale 1 axis. Table 4.2 shows the results under the null and alternative hypotheses. The simulations under the null indicate acceptable Type 1 error rates for all values of $v \leq 2$ in the two-dimensional hazard models. The one-dimensional hazard models also had acceptable Type 1 error rates except when the other time scale was included as a misspecified time-varying covariate in which case the Type 1 error rate was over two times the nominal 0.05 rate. Under the alternative, all the two-dimensional hazard models had higher power than the one-dimensional hazard models without the time-varying covariate. It appears that MSE is minimized when v is less than or equal to 1.

In the next three simulation scenarios, all units followed parallel linear trajectories, but had different starting points along scale 1. These starting points were chosen at random from a Uniform$[0, 4]$ distribution. The starting point on scale 1 was used as the trajectory variable for stratification. These simulations are similar in structure to using time on study as scale 2 and age as scale 1.

The second scenario used baseline hazard $\lambda_B(t_1, t_2)$. This hazard produces peak hazards at different points on scale 2, depending on the starting point on scale 1. Type 1 error rates were high for $v = +\infty$, but acceptable for $v \leq 2$. Including a misspecified time-varying covariate led to elevated Type 1 error rates when either scale 1 or scale 2 was used as the timescale for analysis. Stratification on scale 1 starting time yielded results comparable to the two-dimensional analysis. The power of the two-dimensional models was higher when $v = 1$ compared to $v < 1$. Table 4.3 summarizes these results.

The third scenario uses baseline hazard $\lambda_1(t_1, t_2)$. Under the null hypothesis, two-dimensional hazard models with $v \leq 1$ had acceptable Type 1 error rates. Under the alternative hypothesis, the two-dimensional hazard model with $v = 1$ had somewhat

less power than the one-dimensional hazard model using scale 1 as the time scale, the true model, but substantially more power than the one-dimensional hazard model using scale 2. Table 4.4 summarizes the results from these simulations.

The fourth scenario uses baseline hazard $\lambda_2(t_1, t_2)$. Under the null hypothesis, two-dimensional hazard models with $v \leq 1$ had acceptable Type 1 error rates. Under the alternative hypothesis, the two-dimensional hazard model with $v = 1$ had noticeably more power than the one-dimensional hazard model using scale 1 as the time scale, but not quite as much power as the one-dimensional model using scale 2 as the time scale, the true model. Table 4.5 summarizes the results from these simulations.

4.5 Discussion

It is worthwhile to consider whether a vector-valued time scale analysis has clear advantages over alternative analyses when conditions for a vector-valued time scale analysis are favorable. The results from the first two simulation scenarios, under hazards $\lambda_A(t_1, t_2)$ and $\lambda_B(t_1, t_2)$, demonstrate that the vector-valued time scale analysis with $v = 1$ has either higher power or lower Type 1 error rates than all the one-dimensional time scale analyses, except perhaps the stratified analyses. The stratification, however, was performed on baseline variables strongly related to trajectory. In this way, the stratified analyses and the vector-valued time scale analyses both eliminated individuals whose trajectories were far from an event time from its risk set.

It is also useful to consider how poorly a vector-valued time scale analysis performs when a one-dimensional analysis is correct. The results from simulations using hazards $\lambda_1(t_1, t_2)$ and $\lambda_2(t_1, t_2)$ illustrate this situation. When the baseline hazard is only a function of time scale 1, the vector-valued time scale analysis with $v = 1$ has a bit less power than the analysis on time scale 1 (0.68 vs. 0.74, and 0.69 vs. 0.73), but higher power than the analysis on time scale 2 (0.49 and 0.50). The vector-valued time scale analysis with $v = 1$ performed similarly to the analysis on time scale 2 that was stratified on starting time on time scale 1. Also, when the baseline hazard is only a function of time scale 2, the vector-valued time scale analysis with $v = 1$ has a bit less power than the analysis on time scale 2 (0.68 vs. 0.73, and 0.68 vs. 0.73), but higher power than the analysis on time scale 1 (0.58 and 0.57).

It is noteworthy that the vector-valued time scale analyses with $v = 1$ had acceptable Type 1 error rates under all the simulation scenarios (0.0398 to 0.0516), even when the hazard was only a function of one time scale. Analyses using a single time scale and a misspecified time-varying covariate, however, had high Type 1 error rates, sometimes over 0.1. So, a vector-valued analysis can be viewed as relatively robust.

These results suggest two potential conclusions. First, when the baseline hazard varies substantially as a function of more than one time scale, a vector-valued time scale analysis may yield lower Type 1 errors or higher power than a one-dimensional time scale analysis that fails to model the effects of the other time scales on the hazard parametrically. Second, when the baseline hazard only depends on a single time scale, the vector-valued time scale analysis may have power between a one-dimensional analysis on the correct time scale and a one-dimensional analysis on another time scale.

This suggests that performing a vector-valued time scale analysis could complement a one-dimensional time scale analysis. If the results of the two analyses are in general agreement, with perhaps somewhat higher p-values in the vector-valued time scale analysis, then that result suggests that the major effects of alternative time scales are properly modeled in the one-dimensional analysis. If the results of the two analyses diverge, then this suggests that there the effects of alternative time scales are poorly modeled in the one-dimensional analysis.

Although the simulation results with $v = 1$ appear successful, a general recommendation concerning an optimal value of v may not be warrented. The choice of v could be related to the selection of nearness criteria, and is a possible topic of future study. Future research could also include extending this method using different weighting schemes, or extending it to include time-varying covariates.

Covariate	Conventional	Vector-Valued Time Scale
Age (Years)	1.026 [1.006, 1.046]	–
Serum albumin (mg/dl)	0.452 [0.262, 0.779]	0.449 [0.258, 0.780]
Serum bilirunbin (mg/dl)	1.107 [1.068, 1.148]	1.107 [1.065, 1.152]
Urine copper (ug/day)	1.003 [1.001, 1.005]	1.005 [1.002, 1.007]
Untreated or successfully treated edema	1.021 [0.550, 1.896]	1.073 [0.545, 2.116]
Edema despite diuretic therapy	2.560 [1.291, 5.077]	2.028 [0.898, 4.576]
Standardised blood clotting time	1.279 [1.029, 1.590]	1.216 [0.969, 1.527]
Histologic stage of disease=2	4.351 [0.536, 35.290]	3.534 [0.441, 28.315]
Histologic stage of disease=3	5.814 [0.759, 44.552]	4.359 [0.560, 33.901]
Histologic stage of disease=4	8.034 [1.052, 61.359]	7.801 [1.009, 60.319]
Active Treatment	0.864 [0.572, 1.303]	1.197 [0.777, 1.846]

Table 4.1: Hazard ratios with 95% confidence intervals for the conventional analysis and the vector-valued time scale analysis of the Mayo Clinic Primary Biliary Cirrhosis Data.

$\beta_{trt} = 0$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.0008	0.0008	0.0968	0.0972	0.0094	0.0506
Scale 2 time-varying Scale 1	0.0003	0.0003	0.0994	0.1254	0.0157	0.1146
Scale 2 Stratified on Angle	0.0009	0.0009	0.1004	0.1020	0.0104	0.0524
Scale 1	-0.0013	-0.0013	0.0968	0.0986	0.0097	0.0520
Scale 1 time-varying Scale 2	-0.0009	-0.0009	0.0995	0.1269	0.0161	0.1248
Two-Dimensional $v = +\infty$	0.0001	0.0001	0.1128	0.1220	0.0149	0.0676
Two-Dimensional $v = 2$	0.0007	0.0007	0.1034	0.1019	0.0104	0.0460
Two-Dimensional $v = 1$	0.0008	0.0008	0.1021	0.0994	0.0099	0.0438
Two-Dimensional $v = 0.5$	0.0009	0.0009	0.1018	0.0989	0.0098	0.0442
Two-Dimensional $v = 0^+$	0.0009	0.0009	0.1017	0.0989	0.0098	0.0444
$\beta_X = 0$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	0.0010	0.0010	0.0977	0.0987	0.0098	0.0542
Scale 2 time-varying Scale 1	-0.0013	-0.0013	0.1003	0.1225	0.0150	0.1072
Scale 2 Stratified on Angle	-0.0011	-0.0011	0.1014	0.1023	0.0105	0.0492
Scale 1	-0.0020	-0.0020	0.0978	0.0995	0.0099	0.0506
Scale 1 time-varying Scale 2	-0.0012	-0.0012	0.1003	0.1235	0.0153	0.1080
Two-Dimensional $v = +\infty$	-0.0008	-0.0008	0.1139	0.1218	0.0148	0.0584
Two-Dimensional $v = 2$	-0.0013	-0.0013	0.1043	0.1022	0.0105	0.0418
Two-Dimensional $v = 1$	-0.0014	-0.0014	0.1030	0.1000	0.0100	0.0398
Two-Dimensional $v = 0.5$	-0.0015	-0.0015	0.1026	0.0996	0.0099	0.0404
Two-Dimensional $v = 0^+$	-0.0015	-0.0015	0.1025	0.0998	0.0100	0.0408
$\beta_{trt} = 0.25$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.2021	-0.0479	0.0984	0.0982	0.0119	0.5458
Scale 2 time-varying Scale 1	0.2674	0.0174	0.1021	0.1230	0.0154	0.7250
Scale 2 Stratified on Angle	0.2308	-0.0192	0.1023	0.1035	0.0111	0.6150
Scale 1	0.2015	-0.0485	0.0984	0.0982	0.0120	0.5382
Scale 1 time-varying Scale 2	0.2675	0.0175	0.1022	0.1236	0.0156	0.7204
Two-Dimensional $v = +\infty$	0.2688	0.0188	0.1154	0.1233	0.0156	0.6392
Two-Dimensional $v = 2$	0.2446	-0.0054	0.1055	0.1030	0.0106	0.6482
Two-Dimensional $v = 1$	0.2419	-0.0081	0.1041	0.1009	0.0102	0.6518
Two-Dimensional $v = 0.5$	0.2413	-0.0087	0.1038	0.1006	0.0102	0.6510
Two-Dimensional $v = 0^+$	0.2414	-0.0086	0.1037	0.1009	0.0103	0.6514
$\beta_X = 0.25$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	0.2021	-0.0479	0.0992	0.1003	0.0124	0.5286
Scale 2 time-varying Scale 1	0.2621	0.0121	0.1022	0.1215	0.0149	0.7042
Scale 2 Stratified on Angle	0.2325	-0.0175	0.1035	0.1056	0.0115	0.6182
Scale 1	0.2006	-0.0494	0.0991	0.1000	0.0124	0.5284
Scale 1 time-varying Scale 2	0.2618	0.0118	0.1023	0.1234	0.0154	0.6936
Two-Dimensional $v = +\infty$	0.2708	0.0208	0.1170	0.1258	0.0163	0.6376
Two-Dimensional $v = 2$	0.2459	-0.0041	0.1068	0.1057	0.0112	0.6430
Two-Dimensional $v = 1$	0.2428	-0.0072	0.1053	0.1035	0.0108	0.6450
Two-Dimensional $v = 0.5$	0.2421	-0.0079	0.1050	0.1031	0.0107	0.6426
Two-Dimensional $v = 0^+$	0.2419	-0.0081	0.1048	0.1033	0.0107	0.6414

Table 4.2: Mean simulation results with baseline hazard $\lambda_A(t_1, t_2)$ with trajectories radiating from the origin. $\hat{SE}(\hat{\beta_{trt}})$ and $\hat{SE}(\hat{\beta_X})$ are the standard errors estimated by the software using the data. MC SE is the standard error calculated from the simulation coefficient estimates. The rejection rate is the proportion of simulated data sets where the hypothesis that the coefficient is zero is rejected at the $\alpha = 0.05$ level.

$\beta_{trt} = 0$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.0002	0.0002	0.0970	0.0995	0.0099	0.0550
Scale 2 time-varying Scale 1	0.0009	0.0009	0.0980	0.1048	0.0110	0.0632
Scale 2 Stratified on Scale 1	-0.0001	-0.0001	0.1010	0.1025	0.0105	0.0520
Scale 1	0.0002	0.0002	0.0971	0.0869	0.0076	0.0256
Scale 1 time-varying Scale 2	0.0015	0.0015	0.0996	0.1272	0.0162	0.1212
Two-Dimensional $v = +\infty$	-0.0004	-0.0004	0.1018	0.1067	0.0114	0.0606
Two-Dimensional $v = 2$	-0.0003	-0.0003	0.1011	0.1014	0.0103	0.0494
Two-Dimensional $v = 1$	-0.0002	-0.0002	0.1008	0.0979	0.0096	0.0422
Two-Dimensional $v = 0.5$	-0.0001	-0.0001	0.1007	0.0957	0.0092	0.0388
Two-Dimensional $v = 0^+$	0.0000	0.0000	0.1006	0.0933	0.0087	0.0346
$\beta_X = 0$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	0.0000	0.0000	0.0979	0.0997	0.0099	0.0520
Scale 2 time-varying Scale 1	-0.0001	-0.0001	0.0991	0.1054	0.0111	0.0628
Scale 2 Stratified on Scale 1	0.0007	0.0007	0.1019	0.1033	0.0107	0.0530
Scale 1	-0.0003	-0.0003	0.0976	0.0852	0.0073	0.0262
Scale 1 time-varying Scale 2	0.0003	0.0003	0.1006	0.1259	0.0159	0.1182
Two-Dimensional $v = +\infty$	0.0003	0.0003	0.1029	0.1085	0.0118	0.0614
Two-Dimensional $v = 2$	0.0002	0.0002	0.1021	0.1030	0.0106	0.0514
Two-Dimensional $v = 1$	0.0001	0.0001	0.1017	0.0993	0.0099	0.0458
Two-Dimensional $v = 0.5$	0.0001	0.0001	0.1016	0.0970	0.0094	0.0418
Two-Dimensional $v = 0^+$	0.0000	0.0000	0.1015	0.0944	0.0089	0.0390
$\beta_{trt} = 0.25$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.2107	-0.0393	0.0984	0.0996	0.0115	0.5714
Scale 2 time-varying Scale 1	0.2593	0.0093	0.1001	0.1040	0.0109	0.7328
Scale 2 Stratified on Scale 1	0.2513	0.0013	0.1031	0.1039	0.0108	0.6808
Scale 1	0.1635	-0.0865	0.0981	0.0861	0.0149	0.3674
Scale 1 time-varying Scale 2	0.2902	0.0402	0.1027	0.1233	0.0168	0.7802
Two-Dimensional $v = +\infty$	0.2625	0.0125	0.1041	0.1084	0.0119	0.7082
Two-Dimensional $v = 2$	0.2502	0.0002	0.1032	0.1027	0.0105	0.6800
Two-Dimensional $v = 1$	0.2416	-0.0084	0.1028	0.0990	0.0099	0.6568
Two-Dimensional $v = 0.5$	0.2359	-0.0141	0.1026	0.0968	0.0096	0.6394
Two-Dimensional $v = 0^+$	0.2294	-0.0206	0.1025	0.0945	0.0093	0.6166
$\beta_X = 0.25$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	0.2076	-0.0424	0.0992	0.1014	0.0121	0.5574
Scale 2 time-varying Scale 1	0.2569	0.0069	0.1009	0.1066	0.0114	0.7276
Scale 2 Stratified on Scale 1	0.2501	0.0001	0.1042	0.1063	0.0113	0.6798
Scale 1	0.1599	-0.0901	0.0982	0.0852	0.0154	0.3558
Scale 1 time-varying Scale 2	0.2832	0.0332	0.1025	0.1228	0.0162	0.7588
Two-Dimensional $v = +\infty$	0.2623	0.0123	0.1054	0.1113	0.0125	0.6986
Two-Dimensional $v = 2$	0.2497	-0.0003	0.1043	0.1054	0.0111	0.6754
Two-Dimensional $v = 1$	0.2409	-0.0091	0.1038	0.1017	0.0104	0.6510
Two-Dimensional $v = 0.5$	0.2351	-0.0149	0.1036	0.0994	0.0101	0.6308
Two-Dimensional $v = 0^+$	0.2285	-0.0215	0.1035	0.0969	0.0099	0.6090

Table 4.3: Mean simulation results with baseline hazard $\lambda_B(t_1, t_2)$ and different starting points along scale 1. $\hat{SE}(\hat{\beta_{trt}})$ and $\hat{SE}(\hat{\beta_X})$ are the standard errors estimated by the software using the data. MC SE is the standard error calculated from the simulation coefficient estimates. The rejection rate is the proportion of simulated data sets where the hypothesis that the coefficient is zero is rejected at the $\alpha = 0.05$ level.

$\beta_{trt} = 0$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.0002	0.0002	0.0957	0.0986	0.0097	0.0576
Scale 2 time-varying Scale 1	0.0013	0.0013	0.0978	0.1304	0.0170	0.1378
Scale 2 Stratified on Scale 1	0.0010	0.0010	0.0999	0.1037	0.0108	0.0572
Scale 1	0.0008	0.0008	0.0957	0.0994	0.0099	0.0554
Scale 1 time-varying Scale 2	0.0008	0.0008	0.0964	0.1007	0.0101	0.0566
Two-Dimensional $v = +\infty$	0.0007	0.0007	0.1010	0.1116	0.0124	0.0722
Two-Dimensional $v = 2$	0.0007	0.0007	0.1002	0.1055	0.0111	0.0590
Two-Dimensional $v = 1$	0.0007	0.0007	0.0998	0.1014	0.0103	0.0488
Two-Dimensional $v = 0.5$	0.0007	0.0007	0.0997	0.0988	0.0098	0.0446
Two-Dimensional $v = 0^+$	0.0006	0.0006	0.0996	0.0960	0.0092	0.0402
$\beta_X = 0$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	-0.0015	-0.0015	0.0967	0.0987	0.0097	0.0502
Scale 2 time-varying Scale 1	0.0004	0.0004	0.0990	0.1284	0.0165	0.1244
Scale 2 Stratified on Scale 1	0.0004	0.0004	0.1010	0.1047	0.0110	0.0536
Scale 1	-0.0002	-0.0002	0.0970	0.1007	0.0101	0.0542
Scale 1 time-varying Scale 2	-0.0001	-0.0001	0.0976	0.1019	0.0104	0.0570
Two-Dimensional $v = +\infty$	0.0003	0.0003	0.1023	0.1125	0.0127	0.0724
Two-Dimensional $v = 2$	0.0003	0.0003	0.1014	0.1063	0.0113	0.0586
Two-Dimensional $v = 1$	0.0003	0.0003	0.1010	0.1021	0.0104	0.0488
Two-Dimensional $v = 0.5$	0.0003	0.0003	0.1008	0.0994	0.0099	0.0436
Two-Dimensional $v = 0^+$	0.0004	0.0004	0.1007	0.0965	0.0093	0.0362
$\beta_{trt} = 0.25$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.1879	-0.0621	0.0971	0.0981	0.0135	0.4916
Scale 2 time-varying Scale 1	0.2931	0.0431	0.1007	0.1256	0.0176	0.7888
Scale 2 Stratified on Scale 1	0.2455	-0.0045	0.1020	0.1035	0.0107	0.6822
Scale 1	0.2525	0.0025	0.0979	0.0997	0.0099	0.7360
Scale 1 time-varying Scale 2	0.2542	0.0042	0.0986	0.1012	0.0103	0.7332
Two-Dimensional $v = +\infty$	0.2687	0.0187	0.1034	0.1123	0.0130	0.7284
Two-Dimensional $v = 2$	0.2555	0.0055	0.1024	0.1059	0.0112	0.7084
Two-Dimensional $v = 1$	0.2461	-0.0039	0.1020	0.1017	0.0104	0.6836
Two-Dimensional $v = 0.5$	0.2399	-0.0101	0.1018	0.0991	0.0099	0.6644
Two-Dimensional $v = 0^+$	0.2328	-0.0172	0.1017	0.0963	0.0096	0.6432
$\beta_X = 0.25$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	0.1903	-0.0597	0.0978	0.0987	0.0133	0.4968
Scale 2 time-varying Scale 1	0.2902	0.0402	0.1011	0.1228	0.0167	0.7764
Scale 2 Stratified on Scale 1	0.2490	-0.0010	0.1034	0.1043	0.0109	0.6778
Scale 1	0.2558	0.0058	0.0994	0.1008	0.0102	0.7346
Scale 1 time-varying Scale 2	0.2573	0.0073	0.1000	0.1020	0.0105	0.7316
Two-Dimensional $v = +\infty$	0.2735	0.0235	0.1051	0.1126	0.0132	0.7346
Two-Dimensional $v = 2$	0.2599	0.0099	0.1039	0.1063	0.0114	0.7138
Two-Dimensional $v = 1$	0.2502	0.0002	0.1034	0.1023	0.0105	0.6878
Two-Dimensional $v = 0.5$	0.2438	-0.0062	0.1032	0.0998	0.0100	0.6692
Two-Dimensional $v = 0^+$	0.2365	-0.0135	0.1030	0.0971	0.0096	0.6460

Table 4.4: Mean simulation results with baseline hazard $\lambda_1(t_1, t_2)$ and different starting points along scale 1. $\hat{SE}(\hat{\beta_{trt}})$ and $\hat{SE}(\hat{\beta_X})$ are the standard errors estimated by the software using the data. MC SE is the standard error calculated from the simulation coefficient estimates. The rejection rate is the proportion of simulated data sets where the hypothesis that the coefficient is zero is rejected at the $\alpha = 0.05$ level.

$\beta_{trt} = 0$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	-0.0012	-0.0012	0.0968	0.0986	0.0097	0.0536
Scale 2 time-varying Scale 1	-0.0012	-0.0012	0.0973	0.0999	0.0100	0.0548
Scale 2 Stratified on Scale 1	-0.0014	-0.0014	0.1004	0.1023	0.0105	0.0518
Scale 1	-0.0004	-0.0004	0.0969	0.0916	0.0084	0.0390
Scale 1 time-varying Scale 2	-0.0011	-0.0011	0.0984	0.1177	0.0139	0.0952
Two-Dimensional $v = +\infty$	-0.0012	-0.0012	0.1013	0.1088	0.0118	0.0678
Two-Dimensional $v = 2$	-0.0010	-0.0010	0.1007	0.1039	0.0108	0.0592
Two-Dimensional $v = 1$	-0.0009	-0.0009	0.1004	0.1004	0.0101	0.0516
Two-Dimensional $v = 0.5$	-0.0008	-0.0008	0.1003	0.0982	0.0096	0.0468
Two-Dimensional $v = 0^+$	-0.0007	-0.0007	0.1003	0.0958	0.0092	0.0410
$\beta_X = 0$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	-0.0007	-0.0007	0.0980	0.1002	0.0100	0.0508
Scale 2 time-varying Scale 1	-0.0008	-0.0008	0.0986	0.1012	0.0102	0.0502
Scale 2 Stratified on Scale 1	-0.0009	-0.0009	0.1016	0.1039	0.0108	0.0520
Scale 1	-0.0001	-0.0001	0.0978	0.0929	0.0086	0.0356
Scale 1 time-varying Scale 2	0.0005	0.0005	0.0996	0.1175	0.0138	0.0946
Two-Dimensional $v = +\infty$	-0.0008	-0.0008	0.1026	0.1108	0.0123	0.0654
Two-Dimensional $v = 2$	-0.0008	-0.0008	0.1019	0.1057	0.0112	0.0536
Two-Dimensional $v = 1$	-0.0009	-0.0009	0.1016	0.1021	0.0104	0.0466
Two-Dimensional $v = 0.5$	-0.0009	-0.0009	0.1014	0.0998	0.0100	0.0408
Two-Dimensional $v = 0^+$	-0.0009	-0.0009	0.1013	0.0974	0.0095	0.0366
$\beta_{trt} = 0.25$	$\hat{\beta}_{trt}$	Bias	$\hat{SE}(\hat{\beta_{trt}})$	MC SE	MSE	Rejection Rate
Scale 2	0.2515	0.0015	0.0987	0.0984	0.0097	0.7292
Scale 2 time-varying Scale 1	0.2532	0.0032	0.0992	0.0994	0.0099	0.7302
Scale 2 Stratified on Scale 1	0.2512	0.0012	0.1024	0.1019	0.0104	0.6992
Scale 1	0.2103	-0.0397	0.0984	0.0918	0.0100	0.5798
Scale 1 time-varying Scale 2	0.2721	0.0221	0.1009	0.1172	0.0142	0.7456
Two-Dimensional $v = +\infty$	0.2652	0.0152	0.1035	0.1084	0.0120	0.7254
Two-Dimensional $v = 2$	0.2541	0.0041	0.1027	0.1031	0.0107	0.7042
Two-Dimensional $v = 1$	0.2460	-0.0040	0.1023	0.0997	0.0099	0.6800
Two-Dimensional $v = 0.5$	0.2406	-0.0094	0.1022	0.0975	0.0096	0.6626
Two-Dimensional $v = 0^+$	0.2344	-0.0156	0.1021	0.0952	0.0093	0.6418
$\beta_X = 0.25$	$\hat{\beta}_X$	Bias	$\hat{SE}(\hat{\beta_X})$	MC SE	MSE	Rejection Rate
Scale 2	0.2552	0.0052	0.1000	0.1000	0.0100	0.7330
Scale 2 time-varying Scale 1	0.2567	0.0067	0.1006	0.1008	0.0102	0.7330
Scale 2 Stratified on Scale 1	0.2556	0.0056	0.1038	0.1031	0.0107	0.7034
Scale 1	0.2109	-0.0391	0.0988	0.0913	0.0099	0.5746
Scale 1 time-varying Scale 2	0.2706	0.0206	0.1014	0.1135	0.0133	0.7490
Two-Dimensional $v = +\infty$	0.2693	0.0193	0.1050	0.1097	0.0124	0.7282
Two-Dimensional $v = 2$	0.2576	0.0076	0.1041	0.1042	0.0109	0.7032
Two-Dimensional $v = 1$	0.2491	-0.0009	0.1036	0.1006	0.0101	0.6836
Two-Dimensional $v = 0.5$	0.2434	-0.0066	0.1034	0.0983	0.0097	0.6688
Two-Dimensional $v = 0^+$	0.2369	-0.0131	0.1033	0.0959	0.0094	0.6470

Table 4.5: Mean results with baseline hazard $\lambda_2(t_1, t_2)$ and different starting points along scale 1. $\hat{SE}(\hat{\beta_{trt}})$ and $\hat{SE}(\hat{\beta_X})$ are the standard errors estimated by the software using the data. MC SE is the standard error calculated from the simulation coefficient estimates. The rejection rate is the proportion of simulated data sets where the hypothesis that the coefficient is zero is rejected at the $\alpha = 0.05$ level.

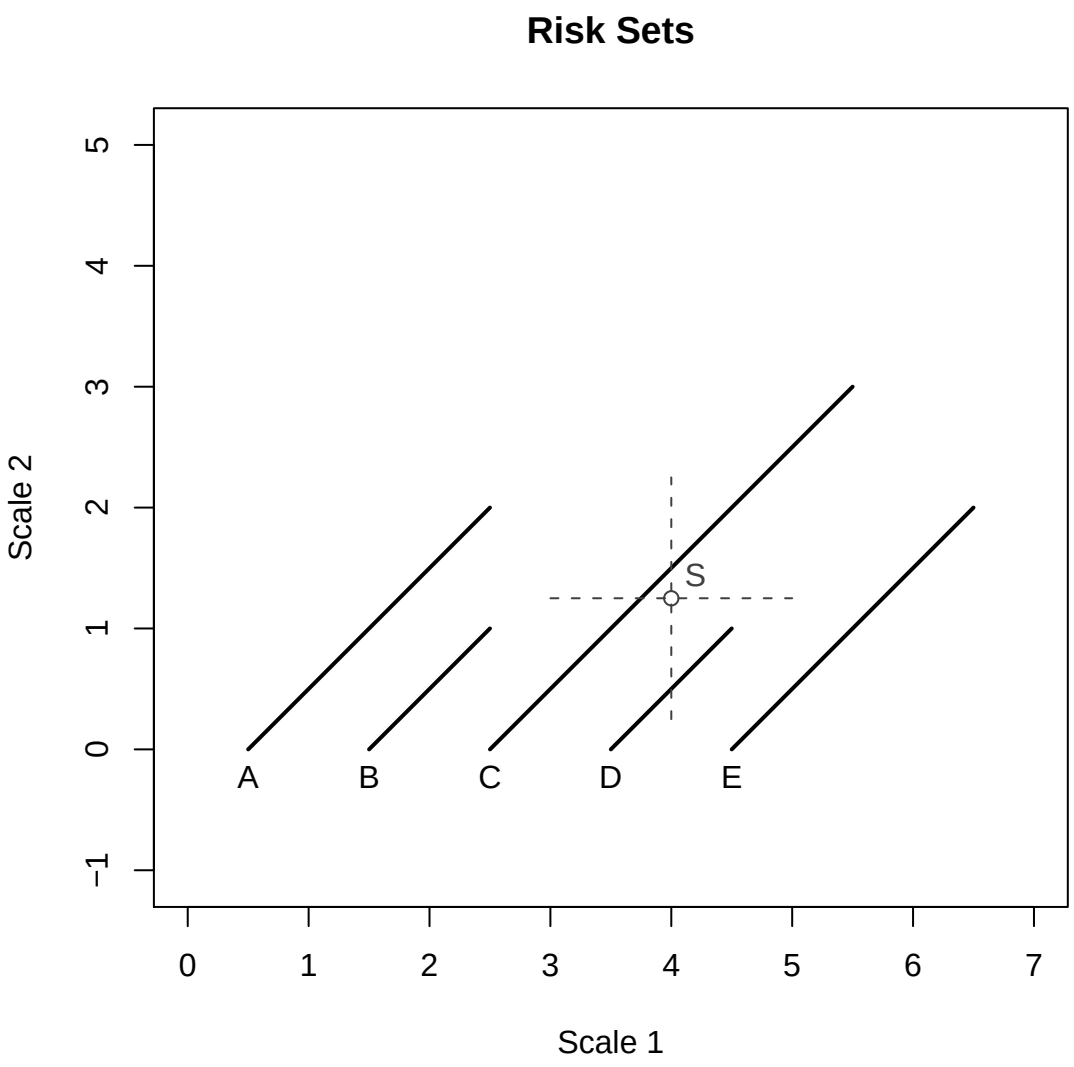

Figure 4.1: The slanted lines marked A, B, C, D, and E represent the paths of four units through the two dimensional time scale, labeled here as Scale 1 and Scale 2. Dashed lines indicate the positions that are considered nearby S. A survives beyond S on time scale 2, but is not in the risk set of S because A is not nearby S. B does not survive beyond S on either scale, and is not in the risk set. C survives beyond S on both time scales and is nearby S on both scales, so C is in the risk set with weight 1^v. D survives beyond S nearby S on only one of two scales, so it is in the risk set with weight 0.5^v. E is not in the risk set because it is not at risk before S on all time scales.

Chapter 5

Conclusions

Joint models for longitudinal and event time data can fail to produce event probability prediction errors smaller than simpler analyses. Simulations described in Chapter 2 suggest that these poor prediction errors increase when the error in the longitudinal measurements increase.

The joint modeling framework seeks to maximize the joint likelihood of both the longitudinal and event time submodels without a preference for one submodel or the other. The fitted joint models strike a balance between fitting the longitudinal data and fitting the event time data. Greater uncertainty in the longitudinal data allows greater flexibility in fitting the longitudinal data while maximizing the likelihood of the survival data. The statistical significance of the model parameters can indicate whether the fitted model suggests that there is a relationship between longitudinal trajectory and hazard, but the joint model framework allows for that to be the result of either trajectory being a good predictor of hazard or hazard being a good predictor of trajectory. Chapter 2 demonstrated that when the trajectory is poorly defined by the longitudinal data alone, the event time data can have great influence on the predicted trajectory.

Asymptotic evaluations of these methods are necessarily incomplete because they will miss such small-sample effects. Future work could include further investigation into the conditions where joint model produces poor event probability predictions. Also, it could be worthwhile to determine whether somewhat different joint model frameworks are prone to poor predictions.

Chapter 3 demonstrated that using a Bayesian joint model can allow for inference

concerning a possible mediator that would not be possible using separate models. Because the measures of X-inactivation were missing for all of the prematurely dead mice, it was not possible to investigate the possibility of a relationship between X-inactivation and survival using a standard logistic regression model. The Bayesian joint models in Chapter 3 worked by allowing using the available data to produce a joint posterior distribution for both X-inactivation and the model paramters.

Chapter 4 showed that vector-valued time scale analyses can have advantages over conventional proportional hazards regression models. When the baseline hazard is a function of more than one time scale, vector-valued time scale analysis can have advantages over scalar-valued time scale analyses in terms of power and Type 1 error rates. When the baseline hazard is only a function of one time scale, vector-valued time scale analyses can perform almost as well at the true scalar-valued times cale hazard models, and better than incorrectly paramterized scalar-valued time scale models.

Future work would include development of improved schemes for risk set inclusion and weighting. Also, it is worthwhile to investigate what conditions are necessary for a vector-valued time scale analysis to provide a substantial improvement over more conventional analyses.

References

[1] R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics (Oxford, England)*, 1(4):465–480, Dec 2000. PUBM: Print; JID: 100897327; ppublish.

[2] AA Tsiatis, V. Degruttola, and MS Wulfsohn. Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, 90(429), 1995.

[3] Michael S. Wulfsohn and Anastasios A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997.

[4] Xu Guo and Bradley P. Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *Amer. Statist.*, 58(1):16–24, 2004.

[5] Edward F. Vonesh, Tom Greene, and Mark D. Schluchter. Shared parameter models for the joint analysis of longitudinal data and event times. *Stat. Med.*, 25(1):143–163, 2006.

[6] Anastasios A. Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statist. Sinica*, 14(3):809–834, 2004.

[7] T.S. Rector, S.H. Kubo, and J.N. Cohn. Patients self-assessment of their congestive heart failure. Part 2: content, reliability and validity of a new measure, the Minnesota Living with Heart Failure questionnaire. *Heart failure*, 3(5):198–209, 1987.

[8] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

[9] R. Schoop, E. Graf, and M. Schumacher. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2):603–610, 2008, http://www.blackwell-synergy.com/doi/pdf/10.1111/j.1541-0420.2007.00889.x.

[10] Timothy E. Hanson, Adam J. Branscum, and Wesley O. Johnson. Bayesian semiparametric methods for joint modeling event time and longitudinal data: A case study illustrating competing approaches. Technical Report 2008-010, Division of Biostatistics, University of Minnesota, 2008.

[11] Michelle N. Rheault, Stefan M. Kren, Linda A. Hartich, Melanie Wall, William Thomas, Hector A. Mesa, Philip Avner, George E. Lees, Clifford E. Kashtan, and Yoav Segal. X-inactivation modifies disease severity in female carriers of murine X-linked Alport syndrome. *Nephrol. Dial. Transplant.*, 25(3):764–769, 2010, http://ndt.oxfordjournals.org/cgi/reprint/25/3/764.pdf.

[12] Michelle N. Rheault, Stefan M. Kren, Beth K. Thielen, Hector A. Mesa, John T. Crosson, William Thomas, Yoshikazu Sado, Clifford E. Kashtan, and Yoav Segal. Mouse Model of X-Linked Alport Syndrome. *J Am Soc Nephrol*, 15(6):1466–1474, 2004, http://jasn.asnjournals.org/cgi/reprint/15/6/1466.pdf.

[13] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[14] O. Aalen, Ø. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view*. Springer Verlag, 2008.

[15] F. Skårderud, P. Nygren, and B. Edlund. 'Bad Boys" Bodies: The Embodiment of Troubled Lives. Body Image and Disordered Eating Among Adolescents in Residential Childcare Institutions. *Clinical Child Psychology and Psychiatry*, 10(3):395, 2005.

[16] E.L. Kom, B.I. Graubard, and D. Midthune. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology*, 145(1):72–80, 1997.

[17] A.C.M. Thiebaut and J. Benichou. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in medicine*, 23(24), 2004.

[18] VT Farewell and DR Cox. A note on multiple time scales in life testing. *Applied Statistics*, pages 73–75, 1979.

[19] D. Oakes. Multiple time scales in survival analysis. *Lifetime Data Analysis*, 1(1):7–18, 1995.

[20] I. Gertsbakh and K.B. Kordonsky. Choice of the best time scale for preventive maintenance in heterogeneous environments. *European Journal of Operational Research*, 98(1):64–74, 1997.

[21] K.B. Kordonsky and I. Gertsbakh. Multiple time scales and the lifetime coefficient of variation: engineering applications. *Lifetime Data Analysis*, 3(2):139–156, 1997.

[22] T. Duchesne and J. Lawless. Alternative time scales and failure time models. *Lifetime Data Analysis*, 6(2):157–179, 2000.

[23] T. Duchesne and J. Lawless. Semiparametric inference methods for general time scale models. *Lifetime Data Analysis*, 8(3):263–276, 2002.

[24] B. Efron. The two-way proportional hazards model. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 899–909, 2002.

[25] P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[26] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[27] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.

[28] Terry Therneau and original R port by Thomas Lumley. *survival: Survival analysis, including penalised likelihood.*, 2009. R package version 2.35-7.