Models for Survival and Longitudinal Data in Cancer Research

by
Ning Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Professor Jeremy M.G. Taylor, Chair
Associate Professor Bin Nan
Assistant Professor Brisa N. Sanchez
Associate Professor John Thomas Wei

This dissertation is dedicated to my dear sister who bravely fought cancer for over two years and slipped away before seeing me finishing this work; may her soul rest in perpetual peace; and to my husband for his loving comfort during the most difficult times of my life and his endless support to make this dissertation possible.

ACKNOWLEDGEMENTS

Foremost, I would like to express my most sincere thanks to my advisor Dr. Jeremy Taylor who motivated me to pursue a PhD and guided me through my years in the the doctoral program at the University of Michigan. I owe him the deepest gratitude for patiently supporting me, believing in me, and assuring me with his expertise and insight throughout my various endeavors over the years. I have been incredibly blessed with an outstanding advisor, mentor, and role model.

I greatly appreciate Dr. Bin Nan, Dr. Brisa Sanchez, and Dr. John Wei for their accessibility in their busy schedules. Their insightful comments and helpful suggestions helped make this dissertation possible. I am also indebted to Dr. Donna Ankerst for her input in the computing aspect of the Bayesian algorithm.

Additionally, I would like to thank the department of Biostatistics for creating such a rigorous yet friendly academic environment. I would also like to acknowledge the faculty, staff and my fellow graduate students for their various forms of support during my graduate study.

Most importantly, none of this would have been possible without the love and patience of my family. My husband and my sister, to whom this dissertation is dedicated to, have been a constant source of love, support and strength all these years. My parents have always believed in diligence, science, and the pursuit of academic excellence. My other family members have aided and encouraged me throughout this endeavor. I would like to express my heart-felt gratitude to all of them.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

In this research, we investigate several topics related to longitudinal and survival analysis for cancer data. Many studies in cancer research involve following patients or subjects over time and monitoring the occurance or progression of the disease. While standard methods can sometimes be used for analyzing such data, in certain cases the structure of the data is such that special models or methods need to be developed to help understand or interpret the data. In this dissertation, we describe and evaluate some novel approaches for survival and longitudinal analysis.

First, we present our development of a family of cure models which provides a unifying framework for survival analysis with cured subjects. Then, we discuss the joint modeling of longitudinal biomarker and survival data based on penalized B-splines to accommodate the nonlinearity of the longitudinal trajectories with a great degree of flexibility. We apply this model to a prostate cancer study. Last, we predict and evaluate the conditional survival probability based on the joint model previously developed, and compare with simpler alternatives.

1.1 The Family of Cure Models

Many medical studies generate time-to-event data, and survival models such as the Cox proportional hazard model or the accelerated failure time model are widely

used to analyze these data. It is typically assumed that all of the study subjects will eventually experience the event of interest if they are followed long enough. However, in reality, the event does not occur with some subjects, even with very long follow-up periods. A reasonable assumption is that these patients have been cured, and consequently cure rate models have been formulated to analyze these situations.

In the literature, there are two formulations of cure rate models: the mixture model and the bounded cumulative hazard model. The mixture model assumes that the study cohort is composed of susceptible subjects and cured subjects, and thus the marginal survival probability is the weighted average of the conditional survival probabilities of the susceptible and the cured subjects. The other cure rate model, the bounded cumulative hazard model, assumes a latent Poisson distribution and derives the marginal survival probability based on a bounded cumulative hazard function.

In past studies, the relative merits of both models have been studied and compared. In this research, we develop a general family of cure models, indexed by a Box-Cox type transformation parameter, that includes both the mixture model and the bounded cumulative hazard model as special cases. This will create a unified framework in which the most appropriate cure rate model can be determined by the data.

1.2 Joint Non-parametric Models of Longitudinal and Survival Data

Along with time-to-event data, many medical studies also collect longitudinal biomarkers. Instead of analyzing them separately, joint modeling of longitudinal and survival data has attracted great attention.

In the literature, two approaches have been proposed to carry out the joint mod-

eling: a two-stage model and a likelihood based joint model. It has been shown that by sharing information between the longitudinal model and the survival model, the likelihood based approach has the advantage of smaller bias and more efficiency in the parameter estimates. Therefore, we focus on the likelihood based joint model in our research.

While applying the joint model to a real data set, we often observe nonlinear trends in the longitudinal trajectories and also a large variation in their shapes. Traditional data-driven parametric longitudinal models face difficulty when trying to accurately model these trends. To accomodate the nonlinearity with a greater degree of flexibility, we adopt the penalized B-splines approach proposed by Eiler and Marx to model the longitudinal biomarker process, and link it with the risk of failure by the Cox proportional hazard model. A Bayesian approach is developed to estimate the posterior distributions of the parameters with MCMC chains. Metropolis-Hastings algorithm is used to implement the MCMC chains. This model is applied to prostate cancer data.

1.3 Evaluation of the Predicted Conditional Survival Estimates Using Absolute Distance Measures

From the perspective of clinical practice, it is of interest to predict the cancer survival outcome. With the joint model that we developed in the previous chapter, we focus on predicting conditional survival probabilities in this chapter. It is important to evaluate the survival predictions before the model is used on any future patient. In the literature, various measures of predictive accuracy have been proposed and compared. In this chapter, we present the absolute distance based measures to assess the predictive accuracy. Simulation studies are carried out to compare our model with three other approaches: a standard Cox proportional hazard model named the

naive approach, a Cox proportional hazard model with the last observed longitudinal biomarkers as a covariate, and a two-stage model in which a mixed model is used for the longitudinal biomarkers and a time-dependent Cox proportional hazard model is fit for the survival data. We derive the predicted conditional survival estimates based on these four approaches, and we obtain the absolute distance measures using these predicted conditional survival probabilities. We compare these absolute distance measures among these four approaches, and show that our model has a relatively higher predictive power than the other three alternatives.

This thesis is organized as follows. In Chapter 2, we present our propsed family of cure rate models with simulation studies and data applications. Chapter 3 develops the joint modeling of the longitudinal biomarkers and survival data based on penalized B-splines with applications to a prostate cancer study. In Chapter 4, we evaluate the predicted conditional survival probabilities derived from the joint model in Chapter 3 using simulation studies. Chapter 5 gives concluding remarks.

CHAPTER II

The Family of Cure Models

2.1 Introduction

In medicine and public health research, survival models are widely used to analyze time-to-event data from studies in which subjects are followed over a certain time period and the time until the occurence of a pre-defined event is recorded. For example, a study may analyze the time from treatment to infection in patients with renal insufficiency, or the time from bone marrow transplant to recurrence of disease in Leukemia patients, or the time from surgery to recurrence of tumor in breast cancer patients. In these examples, the event of interest is defined to be the onset of infection or the recurrence of disease. We typically assume that every study subject will experience the event of interest before the end the study. In reality, the event may not occur with some subjects for various reasons such as death (due to other causes), geographic relocation, or subjects simply staying healthy for an extended period of time. These patients are considered to be censored in traditional survival analysis, and it is assumed that the event of interest will eventually occur to them if they were followed for a long enough time. The Cox proportional hazard model and the accelerated failure time model are frequently used to analyze these time-to-event data.

However, thanks to modern medical practices, some subjects will never experience the event of interest even with long follow-ups, and therefore are reasonably believed to be medically cured. For example, in breast or prostate cancer research, the event of interest is often the recurrence of tumor or death from cancer after treatment. However, many subjects will never experience these events. In this situation, the subjects are not censored in the traditional sense. We can confidently assume they are cured, and therefore the traditional Cox proportional hazard model and the accelerated failure time model are no longer well suited to analyze this type of data. Consequently, cure rate models have been formulated to address this issue.

In the literature, there are two formulations of cure rate models: the mixture model and the bounded cumulative hazard model. The mixture model was first proposed by Berkson and Gage (1952), and assumes that the study cohort is a mixture of susceptible and cured subjects. The susceptible subjects are those who have a chance to experience the event of interest after the start of the follow-up period, and the cured subjects are those who will never experience the event of interest, even with a complete follow-up. The population survival distribution for this study cohort is the weighted average of the conditional survival probability given being susceptible and the conditional survival probability given being cured (which is always 1), with weights p, the probability of cure, and $1 - p$. This mixture model has been extensively studied by Farewell (1977, 1982), Taylor (1995), Sy and Taylor (2000), Peng and Dear (2000), Li and Taylor (2002), and others.

The other cure rate model, the bounded cumulative hazard model, was first proposed by Yakovlev and Tsodikov (1996). In this model, the cumulative hazard is assumed to be bounded, and the population survival distribution is derived accordingly. This model is biologically motivated and is developed by assuming a latent

Poisson distribution on the number of metastasis-competent tumor cells that remain active after cancer treatment and modeling the promotion time to a detectable metastatic condition. The bounded cumulative hazard model has been examined by Tsodikov (1998), Chen, et al. (1999), Ibrahim, et al.(2001) and Tsodikov (2002).

Both cure models have been widely studied and applied in medical research. For example, Brown and Ibrahim (2003) extended the bounded cumulative hazard model to include longitudinal covariates. Yu, et al. (2006) built a joint longitudinal, survival and cure model in which the mixture model was used for the cure rate. Still, it is not always clear which model is more appropriate to use for a given dataset. Chen, et al. (1999) and Ibrahim, et al. (2001) compared the two models to determine which one was superior.

Instead of weighing the relative merits of each model against each other, it is more beneficial to have one unified model that incorporates both of these two formulations and is flexible enough to generate additional intermediate formulations. With this goal in mind, we developed a family of cure models indexed by a Box-Cox type transformation parameter that encompasses the above two cure rate models as special cases. By estimating the index parameter, the optimum formulation from this family of cure rate models is produced. This way, we no longer have to decide a priori which model to use; instead, the data determines the most suitable formulation of the cure model. To estimate this general family of cure rate models, we developed numerical algorithms to obtain the maximum likelihood estimates of the parameters.

The rest of the chapter is organized as follows. In section 2, we present the background information about cure rate models. In section 3, we introduce our new family of cure models and derive the associated likelihood functions. In section 4, simulation studies are presented. In section 5, we apply this model to a bone marrow

transplant study and a tonsil cancer study. Section 6 gives some concluding remarks.

2.2 Background

In this section, we prepare our discussion of the family of cure rate models with some mathematical background information about the two existing formulations and their extensions.

2.2.1 The Mixture Model

The mixture model assumes that the study cohort is a mixture of susceptible and cured subjects. The susceptible subjects are those who may experience the event of interest after the start of the follow-up period, and the cured subjects are those who will never experience the event of interest even with a complete follow-up. Let binary variable D take value 1 if a study subject is susceptible, and value 0 if the subject is cured. We can define the probability of a study subject being cured to be $Pr(D = 0) = p$, and the probability of being susceptible is hence $Pr(D = 1) = 1 - p$. Conditional on being susceptible, the survival probability of this subject is $S_0(t \mid D = 1)$, where t is the follow-up time. Conditional on being cured, the survival probability of this subject is always 1. Based on the law of total probability, the marginal survival probability $S(t)$ for this subject can be written as

$$\begin{aligned} S(t) &= Pr(D = 0) \times S_0(t \mid D = 0) + Pr(D = 1) \times S_0(t \mid D = 1) \\ &= p \times 1 + (1 - p) \times S_0(t \mid D = 1) \\ &= p + (1 - p)S_0(t|D = 1). \end{aligned} \tag{2.1}$$

As $t \to \infty$, $S_0(t \mid D = 1) \to 0$, and thus $S(t) \to p + (1 - p) \times 0$ which is p, the cure rate. That is, the overall marginal survival probability will level off to the probablity of cure if a long enough follow-up period is allowed.

This mixture model was first proposed by Berkson and Gage (1952), and later was examined and extended by various authors. Farewell (1977, 1982) used logistic regression for the probability of cure p and a Weibull distribution for the conditional survival probability given being susceptible $S_0(t \mid D = 1)$. That is,

$$Pr(D = 0; Z_1) = p(Z_1) = \frac{exp(Z_1\alpha)}{1 + exp(Z_1\alpha)}, \tag{2.2}$$

and

$$S_0(t; Z_2, D = 1) = exp(-t^{\gamma} exp(Z_2\beta)), \tag{2.3}$$

where Z_1 is the covariate that is associated with the cure rate through parameter α, Z_2 is the covariate that is associated with the survival probability given being susceptible through parameter β, and γ is the shape parameter for the Weibull distribution. Yamaguchi (1992) assumed the generalized gamma distribution as an alternative parametric formulation for $S_0(t; Z_2, D = 1)$. Kuk and Chen (1992), Sy and Taylor (2000), and Peng and Dear (2000) chose the semiparametric Cox proportional hazard model for $S_0(t; Z_2, D = 1)$. That is,

$$h_0(t; Z_2, D = 1) = h_0(t; D = 1) exp(Z_2\beta), \tag{2.4}$$

where $h_0(t; Z_2, D = 1)$ is the hazard function conditional on being susceptible. The expectation-maximization (EM) algorithm was used to estimate the parameters. Li and Taylor (2002) generalized it by using the semiparametric accelerated failure time model.

2.2.2 The Bounded Cumulative Hazard Model

The second formulation of cure rate models, the bounded cumulative hazard model, assumes that the cumulative hazard $H(t)$ is bounded. That is, $H(t) = \theta F(t)$

with $\theta \geq 0$ and $F(t)$ being the cumulative distribution function of a non-negative random variable. Thus the overall population survival distribution $S(t)$ is given by

$$S(t) = exp(-\theta F(t)). \tag{2.5}$$

This model tries to model the promotion time to a detectable metastatic condition by assuming a latent Poisson distribution on the number of metastasis-competent tumor cells that remain active after cancer treatment. Specifically, let N denote the number of metastatic-competent tumor cells that remain active after treatment. N is assumed to have a Poisson distribution with a non-negative mean θ. Let T_l be the promotion time of tumor cell l. Conditional on N, the set of promotion times $\{T_l, l = 1 \ldots N\}$ are assumed to be independent and identically distributed as $F(t) = 1 - S(t)$. Then the time to the relapse of cancer can be defined as $T = min(T_l), 1 \leq l \leq N$ with survival probability given by

$$\begin{aligned} S(t) &= P(N = 0) + P(T_1 > t, \ldots, T_N > t, N \geq 1) \\ &= exp(-\theta) + \sum_{k=1}^{\infty} S(t)^k \frac{\theta^k}{k!} exp(-\theta) \\ &= exp(-\theta + \theta S(t)) = exp(-\theta F(t)), \end{aligned} \tag{2.6}$$

As $t \to \infty$, $S(t) \to P(N = 0) = exp(-\theta)$, and thus $exp(-\theta)$ is the cure rate in this model. Here, it is often assumed $\theta = exp(Z\beta)$ where Z is the covariate that is associated with the cure rate through parameter β. Chen and Ibrahim (2001) used the EM algorithm to estimate the parameters, whereas Chen, et al. (2002) developed the Bayesian algorithm with proper posterior distributions for parameter estimates.

2.2.3 Extensions to Cure Models

In the past, these two formulations of cure models have been actively compared (Chen, et al. (1999), and Ibrahim, et al. (2001)). The mixture model has the

advantage of being very intuitive. However, with covariates, the population survival probability $S(t)$ no longer maintains the proportional hazard structure. As for the bounded cumulative hazard model, the hazard function can be calcuated as $h(t) = \theta F(t)$, and thus the bounded cumulative hazard model keeps the proportional hazard structure. Note that the bounded cumulative hazard model gives an improper population survival function. This can be easily seen that as $t \to \infty$, $S(t) \to exp(-\theta) > 0$.

To extend the model to include proper population survival distributions, Yin and Ibrahim (2005) developed a general class of survival models with a cure rate through a transformation on the population hazard function, and implemented the Bayesian Markov chain Monte Carlo (MCMC) scheme based on Gaussian quadrature to estimate the parameters. Specifically, this class of survival models is given by

$$h_{pop}(t; Z) = (f(t)^{\gamma} + \gamma Z\beta)^{1/\gamma}, \tag{2.7}$$

where $h_{pop}(t; Z)$ is the population hazard function, Z is the covariate, $f(t)$ is a probability density function, and γ is an index parameter. It can be shown that as $\gamma \to 0$, this model reduces to the bounded cumulative hazard model. When $\gamma = 1$, this model reduces to an additive hazard model in the form of $h_{pop}(t; Z) = f(t) + Z\beta$ which leads to a proper population survival function. The simulation study suggested that the parameter estimates are reasonably close to true values.

In 2006, Yin and Ibrahim proposed another unified approach to model the cure rate based on the transformation of the population survival function in the form of

$$S_{pop}(t; Z, X) = (1 - \gamma\theta(\gamma, Z)F(t; X))^{1/\gamma}, \tag{2.8}$$

where Z and X are the covariate vectors, $F(t; X)$ is a cumulative distribution function, and γ is the index parameter. As $\gamma \to 0$, this model becomes $S_{pop}(t; Z, X) =$

$exp(-\theta(0,Z)F(t;X))$, which is the bounded cumulative hazard model. As $\gamma \to 1$, this model becomes $S_{pop}(t;Z,X) = 1 - \theta(1,Z)F(t;X)$, which is the mixture model. This unified approach links the two cure rate models using an index parameter. However, it can be shown that as $t \to \infty, S_{pop}(t;Z,X) \to (1-\gamma\theta(\gamma,Z))^{1/\gamma}$, which means the cure rate is dependent on the index parameter. Yin (2008) extended the cure rate model to multivariate failure time data, and again the Bayesian approach was used for model estimation and selection.

Zeng, et. al. (2006) proposed a class of semiparametric transformation models for survival data with a cure rate by transforming the population survival function in the form of

$$S_{pop}(t;Z,X) = (1+\gamma\theta(Z)F(t;X))^{-1/\gamma}, \tag{2.9}$$

where again, Z and X are the covariate vectors, $F(t;X)$ is a cumulative distribution function, and γ is the index parameter. As $\gamma \to 0$, this transformation yields $S_{pop}(t;Z,X) = exp(-\theta(0,Z)F(t;X))$, which is the bounded cumulative hazard model. When $\gamma = 1$, this transformation yields a proportional odds type of cure model. The authors proposed treating the index parameter γ as a known fixed value and using the maximum likelihood estimation method to estimate the other parameters.

Past research has demonstrated that both models are good representations of the data, and the simulation studies showed that the parameter estimates are unbiased and well-behaved. It sometimes can be difficult to determine which model is more suitable for a specific research study. The aim of this chapter is to develop one unified model that can incorporate both of these two cure model formulations and be flexible enough to generate more intermediate formulations.

2.3 A General Family of Cure Models

In this section, we develop a family of cure models, indexed by the Box-Cox transformation parameter, such that the two formulations of cure rate models are both special cases of this family. With the Box-Cox transformation parameter taking different values, additional intermediate cure rate models can be derived and estimated for a better understanding of the association between covariates and survival probabilities.

2.3.1 The Family of Cure Models

In traditional linear regression models, the Box-Cox transformation (Box and Cox, 1964) is often used to transform the outcome variables. The original Box-Cox transformation takes the form of

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ log(y) & \text{if } \lambda = 0, \end{cases}$$

where λ is the Box-Cox transformation parameter, and $\lambda \in (-\infty, \infty)$.

Instead of imposing the Box-Cox transformation on the observed outcome variables, we instead impose it on the population survival probability $S(t; Z)$ and cure rate p. Thus we propose the general family of cure models indexed by λ

$$\frac{S(t;Z)^\lambda - 1}{\lambda} = \frac{p(Z)^\lambda - 1}{\lambda}(1 - S_0(t;Z)), \tag{2.10}$$

where $S(t; Z)$ is the population survival distribution, $0 \leq p(Z) \leq 1$, $0 \leq S_0(t; Z) = 1 - F_0(t; Z) \leq 1$ is a proper survival distribution function, and λ is the Box-Cox transformation parameter $\in (-\infty, \infty)$. By collecting terms in both sides of the above equation, we can rewrite the population survival distribution as

$$S(t;Z) = [p(Z)^\lambda + (1 - p(Z)^\lambda)S_0(t;Z)]^{1/\lambda},$$

and the population hazard function is given by

$$\begin{aligned} h(t;Z) &= -(\frac{1}{\lambda})\frac{(1-p(Z)^{\lambda})\frac{dS_0(t)}{dt}}{p(Z)^{\lambda}+(1-p(Z)^{\lambda})S_0(t;Z)} \\ &= \frac{(1-p(Z)^{\lambda})(S_0(t;Z)h_0(t;Z))}{\lambda S(t;Z)^{\lambda}}. \end{aligned}$$

(2.11)

As $t \to \infty$, $S_0(t;Z) \to 0$, and

$$\begin{aligned} S(\infty;Z) &= (p(Z)^{\lambda}+(1-p(Z)^{\lambda}) \times S_0(\infty;Z))^{1/\lambda} \\ &= (p(Z)^{\lambda}+(1-p(Z)^{\lambda}) \times 0)^{1/\lambda} \\ &= p(Z). \end{aligned}$$

(2.12)

This suggests that $p(Z)$ is the probability of cure which has the desirable property that it does not depend on the index parameter λ. Also, the conditional survival distribution for the susceptible subjects can be written as

$$S(t;Z,susceptible) = \frac{(p(Z)^{\lambda}+(1-p(Z)^{\lambda}) \times S_0(t;Z))^{1/\lambda}-p}{1-p}.$$

When $\lambda = 1$, our model reduces to

$$S(t;Z) = p(Z) + (1-p(Z))S_0(t;Z),$$

which is the mixture model presented in (2.1). When $\lambda = 0$, equation (2.9) becomes

$$log(S(t;Z)) = log(p(Z))(1-S_0(t;Z)),$$

and hence, our model reduces to

$$S(t;Z) = exp(log(p(Z))(1-S_0(t;Z))),$$

which is the bounded cumulative hazard model in the form of (2.5) with $F_0(t;Z) = 1 - S_0(t;Z)$ and $\theta = log(p(Z))$.

It can be easily seen that our model not only has these two formulations of cure models as special cases, but is also general enough to include intermediate formulations as λ varies along the real line. To illustrate this feature, we considered a hypothetical example where $S_0(t) = exp(-0.6t^{0.9})$, $p = 0.2$, and $\lambda = [0, 0.5, 1]$. The population survival distributions are calculated based on $t \in [0, 10]$ and plotted in Figure 2.1. Different λ values generate different survival probabilities, and a statistical test will be needed to draw inference on them.

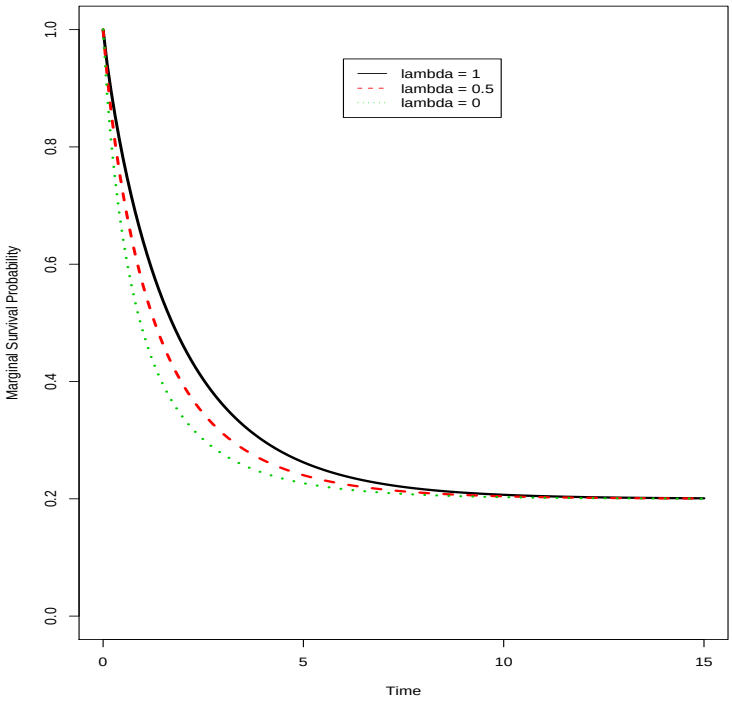

Figure 2.1: Hypothetical example of population survival functions varying with λ

2.3.2 The Covariate Structure

Along with time-to-event data, baseline information is often collected as well. For example, in cancer studies, besides the time to cancer recurrence, we may also have

baseline information such as age, gender, and tumor stage. Therefore, it is necessary to introduce covariates into our family of cure models.

Let Z_1 and Z_2 be two overlapping subsets of covariates or distinct subsets of covariates from the entire set of covariates Z. To understand the association between covariates Z and the cure rate p, we assume a log-log link for p as

$$p_i(Z_1) = exp(-exp(Z_1\alpha)). \tag{2.13}$$

We chose this form of the link function because in the special case when $\lambda = 0$, the population survival function takes the form of

$$\begin{aligned} S(t;Z) &= exp(log(p(Z))(1 - S_0(t;Z))) \\ &= exp(log(exp(-exp(Z\alpha)))(1 - S_0(t;Z))) \\ &= exp(-exp(Z\alpha)F_0(t;Z)). \end{aligned} \tag{2.14}$$

That is, θ in equation 2.5 takes a nice form of $\theta = exp(Z\alpha)$.

To allow the association between covariates Z and the time-to-event distribution, we assume a Weibull form for $S_0(t)$,

$$S_0(t;Z_2) = exp(-\tau t^{\gamma} exp(Z_2\beta)). \tag{2.15}$$

2.3.3 The Likelihood Function

Let T_i and C_i be the event time and censoring time, respectively, of subject i for $i = 1, \ldots, n$. The observed survival time of subject i is given by $t_i = min(T_i, C_i)$, and the corresponding censoring indicator is $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. Let Z_i be the covariates measured from subject i. T_i and C_i are assumed to be independent conditional on Z_i, and $(T_i, C_i, Z_i), i = 1 \ldots n$ are assumed to be

independent and identically distributed. Therefore, the likelihood function is

$$\begin{aligned} L(\alpha,\beta,\gamma,\tau,\lambda;t,\delta,Z) &= \prod_{i=1}^{n}(h(t_i;Z_{2i})S(t_i;Z_{2i}))^{\delta_i}S(t_i;Z_{2i})^{1-\delta_i} \\ &= \prod_{i=1}^{n}(\frac{(1-p(Z_{1i}))^{\lambda})S_0(t_i;Z_{2i})h_0(t_i;Z_{2i})}{\lambda})^{\delta_i}S(t_i;Z_{2i})^{1-\lambda\delta_i}, \end{aligned} \tag{2.16}$$

where $h_0(t)$ is the hazard function corresponding to $S_0(t)$.

Due to the complexity of the model, we were not able to obtain an analytical solution to maximize this likelihood function. A numerical maximization approach such as the Newton-Raphson algorithm can be used instead, and the maximum likelihood estimates of $(\alpha,\beta,\gamma,\tau,\lambda)$ and the variance-covariance matrix can be subsequently obtained. We derived the first order partial derivatives of the log-likelihood function with respect to the parameters and the information matrix, and the details are included in Appendix A.

In the situation where the likelihood function is rather flat, the Newton-Raphson algorithm might encounter convergence issues. When there is not enough information in the data to precisely estimate λ, we can alternatively use the profile likelihood approach to estimate of the parameters. We can consider the profile likelihood approach as a two-stage scheme. In stage one, we fix λ to a finite number of values, such as 0, .5, or 1, or a finer grid of values over a range. We can then apply the Newton-Raphson algorithm to estimate the rest of the parameters to maximize the likelihood function given these fixed λ values. In stage two, based on each set of parameter estimates and these fixed λ values, we can calculate the corresponding likelihood and obtain the maximum of these likelihood values. The corresponding parameter estimates and the λ value are the global maximum likelihood estimators. The likelihood ratio test and profile likelihood confidence intervals can be used for the statistical inference on the parameters (including λ). In fact, past research (Chen,

et al. (1999) and Ibrahim, et al. (2001)) has paid great attention to comparing the mixture model (the special case when $\lambda = 1$) and the bounded cumulative hazard model (the special case when $\lambda = 0$). The profile likelihood approach will serve this purpose by estimating the model fixing λ to 1 or 0 and drawing inference based on the parameter estimates.

2.4 Simulation Studies

We conducted two simulation experiments to investigate the properties of estimates from this model. In the first simulation study, the data were simulated based on the models

$$p = exp(-exp(\alpha_0 + \alpha_1 x_1)) \tag{2.17}$$

$$S_0(t) = exp(-\tau t^{\gamma} exp(\beta_1 x_2)) \tag{2.18}$$

with $\alpha_0 = -0.5$, $\alpha_1 = 2$, $\beta_1 = 1$, $\gamma = 3$, and $\tau = 0.1$. X_1 was an iid sample from Uniform(0,1), and X_2 was an iid sample from Normal(0,1) independent of X_1. Two sample sizes (300 and 600) and two λ values (0 and 1) were considered in the simulation with uniform censoring. The average censoring rate across all datasets was about 50%. The means of the parameter estimates based on 500 simulations are presented in Table 2.1 through 2.4.

Table 2.1 presents the simulation results based on 300 observations in each sample and the true λ is 1. When fixing λ to be 0, the parameter estimates are obviously biased. When fixing λ to 1, the parameter estimates are reasonably close to the true values. When relaxing the constraint on λ to be any real number, the maximum likelihood estimate of λ is relatively unbiased, and the parameter estimates are quite close to those from fixing λ to 1. The standard deviations also behave reasonably well. There seems to be a slight increase in the standard deviation estimates due to

Table 2.1: 300 observations, true $\lambda = 1$, average number of events is 139

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	-0.203	-0.528	-0.529	
(-0.5)	(0.198)	(0.231)	(0.237)	0.246
α_1	1.312	2.090	2.094	
(2.0)	(0.286)	(0.495)	(0.511)	0.501
β_1	1.121	1.010	1.016	
(1.0)	(0.128)	(0.112)	(0.121)	0.119
γ	3.350	3.041	3.055	
(3.0)	(0.241)	(0.210)	(0.220)	0.224
τ	0.044	0.099	0.104	
(0.1)	(0.010)	(0.019)	(0.037)	0.038
λ	0	1	1.057	
(1)	—	—	(0.444)	0.468

estimating λ. Based on the observed information matrix, the coverage rate of α_1 is 96.8%, and the coverage rate of β_1 is 95.2%. The power of the likelihood ratio test of $\lambda = 0$ vs $\lambda = 1$ is 0.70, and the size of the likelihood ratio test of $\lambda = 1$ versus $\lambda = \hat{\lambda}_{MLE}$ is 0.04.

Table 2.2 shows the simulation results based on 600 observations in each sample and the true λ is 1. Due to the bigger sample size, the parameter estimates when fixing λ to 1 and when relaxing the constraint on λ are very close to each other and to the true values. The maximum likelihood estimate of λ has a much smaller bias compared to that in Table 2.1, and the increase in variance estimates due to the estimation of λ is also smaller compared to that in Table 2.1. Again, we are able to obtain good coverage rates for the parameters. Based on the observed information matrix, the coverage rate of α_1 is 95.4%, and the coverage rate of β_1 is 96.2%. The larger sample size also increases the power of the likelihood ratio test. Compared to 0.70 from the previous simulation results, the power of the likelihood ratio test of $\lambda = 0$ versus $\lambda = 1$ when the sample size is 600 is 0.90, and the size of the likelihood ratio test of $\lambda = 1$ versus $\lambda = \hat{\lambda}_{MLE}$ is 0.04.

Table 2.2: 600 observations, true $\lambda = 1$, average number of events is 278

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	-0.188	-0.503	-0.503	
(-0.5)	(0.142)	(0.167)	(0.172)	0.172
α_1	1.302	2.048	2.048	
(2.0)	(0.203)	(0.352)	(0.355)	0.345
β_1	1.114	1.002	1.006	
(1.0)	(0.089)	(0.078)	(0.081)	0.082
γ	3.315	3.009	3.020	
(3.0)	(0.162)	(0.142)	(0.154)	0.155
τ	0.043	0.100	0.101	
(0.1)	(0.007)	(0.012)	(0.025)	0.025
λ	0	1	1.009	
(1)	—	—	(0.300)	0.304

Table 2.3: 300 observations, true $\lambda = 0$, average number of events is 157

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	-0.507	-0.662	-0.529	
(-0.5)	(0.206)	(0.242)	(0.240)	0.225
α_1	2.017	2.302	2.083	
(2.0)	(0.317)	(0.491)	(0.457)	0.413
β_1	1.012	0.861	1.016	
(1.0)	(0.116)	(0.101)	(0.130)	0.129
γ	3.044	2.650	3.057	
(3.0)	(0.196)	(0.165)	(0.254)	0.247
τ	0.099	0.241	0.102	
(0.1)	(0.017)	(0.028)	(0.036)	0.033
λ	0	1	0.002	
(0)	—	—	(0.317)	0.296

Table 2.4: 600 observations, true $\lambda = 0$, average number ofevents is 315

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	-0.501	-0.639	-0.505	
(-0.5)	(0.137)	(0.159)	(0.157)	0.156
α_1	2.010	2.245	2.026	
(2.0)	(0.201)	(0.294)	(0.277)	0.282
β_1	1.012	0.58	1.016	
(1.0)	(0.082)	(0.073)	(0.088)	0.091
γ	3.026	2.631	3.038	
(3.0)	(0.128)	(0.110)	(0.171)	0.172
τ	0.098	0.241	0.100	
(0.1)	(0.012)	(0.022)	(0.023)	0.023
λ	0	1	-0.011	
(0)	—	—	(0.209)	0.205

Similarly, Table 2.3 and 2.4 summarize the simulation results based on 300 and 600 observations in each sample and the true λ is 0. These two tables give the same conclusions as the previous two. When fixing λ to 0 the parameter estimates are very close to those estimates from relaxing the constraint on λ, and they are all reasonably unbiased. Larger sample size decreases the bias and standard deviations. Good coverage rates are reached based on the observed information matrix. When the sample size is 300, the coverage rate of α_1 is 94.8%, and the coverage rate of β_1 is 95.2%. When the sample size is 600, the coverage rate of α_1 is 97.2%, and the coverage rate of β_1 is 96.4%. In terms of the power of the likelihood ratio test of $\lambda = 0$ versus $\lambda = 1$, when sample size is 300, the power is 0.88. And when sample size is 600, the power increased to 0.99. The size of the likelihood ratio test of $\lambda = 0$ versus $\lambda = \hat{\lambda}_{MLE}$ is 0.07 when sample size is 300 and 0.05 when sample size is 600.

In the previous simulations, the cure rate p was assumed to be associated with X_1 only, and the other component of the model S_0 was assumed to be associated with X_2 only, so that there was no overlap between the two sets of covariates associated with these two components of the model. To investigate the effect of overlapping

Table 2.5: 300 observations, true $\lambda = 1$, average number of events is 105

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	0.011	0.037	0.032	
(0)	(0.244)	(0.299)	(0.251)	0.243
α_1	1.474	1.515	1.523	
(1.5)	(0.231)	(0.261)	(0.246)	0.238
α_2	-1.001	-1.045	-1.056	
(-1.0)	(0.416)	(0.436)	(0.437)	0.417
β_1	0.656	1.044	0.998	
(1.0)	(0.319)	(0.257)	(0.421)	0.400
β_2	2.104	1.550	1.645	
(1.5)	(0.557)	(0.432)	(0.613)	0.608
γ	3.352	3.082	3.131	
(3.0)	(0.283)	(0.264)	(0.332)	0.328
τ	0.051	0.098	0.107	
(0.1)	(0.020)	(0.030)	(0.073)	0.072
λ	0	1	1.110	
(1)	—	—	(1.161)	1.185

covariates, we conducted another simulation study where the data were simulated based on the models

$$p = exp(-exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)) \tag{2.19}$$

$$S_0(t) = exp(-\tau t^{\gamma} exp(\beta_1 x_1 + \beta_2 x_2)) \tag{2.20}$$

with $\alpha_0 = 0$, $\alpha_1 = 1.5$, $\alpha_2 = -1.0$, $\beta_1 = 1$, $\beta_2 = 1.5$, $\gamma = 3$, and $\tau = 0.1$. X_1 was an iid sample from Uniform(-1,1), and X_2 was an iid sample from Uniform(0,1) independent of X_1. Again, two sample sizes (300 and 600) and two λ values (0 and 1) were considered, and uniform censoring was included in the simulation. The average censoring rate is about 66%. The means of the parameter estimates based on 500 simulations are presented in Table 2.5 through 2.8. The coverage rate of the parameter estimates based on the information matrix are included in Table 2.9, along with the size and power of the likelihood ratio tests.

From Table 2.5 through 2.8, we can see that the parameter estimates when fixing λ to the true value are very close to those when relaxing the constraint on λ, and they

Table 2.6: 600 observations, true $\lambda = 1$, average number of events is 211

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	-0.002	0.019	0.019	
(0)	(0.166)	(0.172)	(0.172)	0.170
α_1	1.468	1.518	1.521	
(1.5)	(0.153)	(0.163)	(0.165)	0.166
α_2	-0.979	-1.030	-1.032	
(-1.0)	(0.281)	(0.295)	(0.296)	0.290
β_1	0.653	1.022	1.029	
(1.0)	(0.202)	(0.167)	(0.272)	0.292
β_2	2.047	1.522	1.534	
(1.5)	(0.368)	(0.294)	(0.414)	0.422
γ	3.292	3.035	3.037	
(3.0)	(0.197)	(0.178)	(0.230)	0.231
τ	0.052	0.100	0.109	
(0.1)	(0.014)	(0.021)	(0.049)	0.053
λ	0	1	1.137	
(1)	—	—	(0.764)	0.867

Table 2.7: 300 observations, true $\lambda = 0$, average number of events is 110

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	0.027	0.016	0.028	
(0)	(0.226)	(0.227)	(0.231)	0.232
α_1	1.523	1.505	1.523	
(1.5)	(0.241)	(0.240)	(0.244)	0.231
α_2	-1.037	-1.010	-1.039	
(-1.0)	(0.385)	(0.391)	(0.395)	0.397
β_1	1.032	1.410	1.035	
(1.0)	(0.306)	(0.256)	(0.408)	0.394
β_2	1.576	1.022	1.589	
(1.5)	(0.488)	(0.406)	(0.617)	0.632
γ	3.071	2.787	3.102	
(3.0)	(0.227)	(0.208)	(0.305)	0.322
τ	0.098	0.183	0.107	
(0.1)	(0.032)	(0.050)	(0.070)	0.062
λ	0	1	0.024	
(0)	—	—	(0.710)	0.723

Table 2.8: 600 observations, true $\lambda = 0$, average number of events is 220

Parameters (True Values)	Estimate (SD) assuming $\lambda = 0$	Estimate (SD) assuming $\lambda = 1$	Estimate (SD) with λ estimated	Mean of SD's based on observed information matrix
α_0	0.026	0.017	0.027	
(0)	(0.155)	(0.156)	(0.156)	0.163
α_1	1.517	1.500	1.519	
(1.5)	(0.167)	(0.167)	(0.172)	0.161
α_2	-1.034	-1.008	-1.037	
(-1.0)	(0.269)	(0.272)	(0.273)	0.278
β_1	1.009	1.376	1.008	
(1.0)	(0.211)	(0.174)	(0.276)	0.271
β_2	1.567	1.024	1.579	
(1.5)	(0.335)	(0.271)	(0.418)	0.436
γ	3.043	2.762	3.055	
(3.0)	(0.167)	(0.151)	(0.212)	0.219
τ	0.097	0.181	0.100	
(0.1)	(0.022)	(0.034)	(0.038)	0.039
λ	0	1	0.013	
(0)	—	—	(0.425)	0.464

Table 2.9: Coverage rates of parameters

Parameters	Size 300 $\lambda = 1$	Size 600 $\lambda = 1$	Size 300 $\lambda = 0$	Size 600 $\lambda = 0$
α_1	96.40	95.60	94.20	93.60
α_2	91.40	94.40	94.00	95.20
β_1	94.40	95.00	95.60	94.80
β_2	93.00	95.00	94.60	95.80
λ	91.60	95.00	99.40	99.00
Power	15.40	39.60	18.00	44.60
Size	3.40	2.40	4.20	2.60

are all relatively unbiased. There appears to be almost no increase in the variance of $\hat{\alpha}$ due to the estimation of λ, but the inflation in the variance of $\hat{\beta}$ due to estimation of λ is more substantial. The parameter estimate of λ has a larger bias compared to that from the previous simulation study. The variance estimate of $\hat{\lambda}$ is also larger than that in the previous study. Larger sample sizes help decrease the bias and the variance estimate of $\hat{\lambda}$. This suggests that when there are overlapping covariates associated with the two components of the model, it will be difficult to obtain precise estimates of λ unless the sample size is large.

2.5 Application

For illustration we applied our family of cure models to two data sets. One is a small data set from the bone marrow transplant study analyzed by Kersey, et al. in 1987. The other is a larger data set from a tonsil cancer study. The bone marrow transplant data set consists of 91 patients, and the tonsil cancer data set has 672 patients.

2.5.1 Bone Marrow Transplant Study

In the bone marrow transplant study, 91 refractory acute lymphoblastic leukemia patients were followed for 5 years, and their times of recurrence were recorded along with their treatment methods. Among them, 46 patients were in the allogeneic treatment group, and 45 were in the autologous treatment group. Sixty-eight patients in total had recurrence, 33 in the allogeneic group and 35 in the autologous group. The Kaplan-Meier plot in Figure 2.2 shows that the allogeneic group levels off at around 2 years after treatment, and the autologous group levels off at around 1 year after treatment. There are about 26% patients in the allogeneic group and about 20% in the autologus group that did not experience recurrence by the end of the study. The absence of the event of interest for these patients suggests that a cure model may be appropriate to analyze these data.

We fit our general family of cure models to these data. The profile likelihood approach was used to estimate the parameters. The allogeneic group was used as the reference group. Survival time was scaled into years. We searched in the range for $\lambda \in [-10, 5]$ with 0.01 intervals, and we obtained the maximum likehood at $\hat{\lambda} = -3.6$. Table 2.10 presents the parameter estimates along with their standard errors at different values of λ.

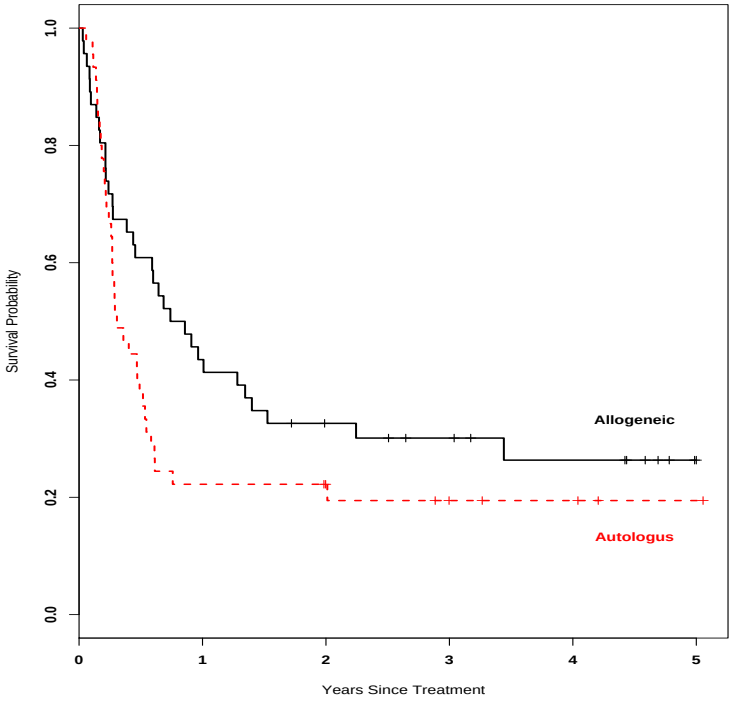

Figure 2.2: Bone marrow transplant study: Kaplan-Meier plot

Table 2.10: Bone marrow transplant study: parameter estimates (full model)

Parameters	$\lambda = 0$ (SE)	$\lambda = 1$ (SE)	$\lambda = \hat{\lambda}$ (SE with λ fixed)	$\lambda = \hat{\lambda}$ (SE with λ estimated)
α_0	0.30	0.26	0.46	
	(0.19)	(0.19)	(0.15)	(0.87)
α_1	0.19	0.22	0.17	
Autologous vs Allogeneic	(0.29)	(0.26)	(0.21)	(0.95)
β_1	0.69	0.75	1.25	
Autologous vs Allogeneic	(0.33)	(0.26)	(0.89)	(1.00)
γ	1.30	1.14	2.21	
	(0.11)	(0.10)	(0.22)	(0.99)
τ	1.01	1.43	0.15	
	(0.23)	(0.27)	(0.10)	(0.61)
λ	0	1	-3.6	
	—	—	–	(0.99)
$logL$	-66.21	-67.26	-65.15	

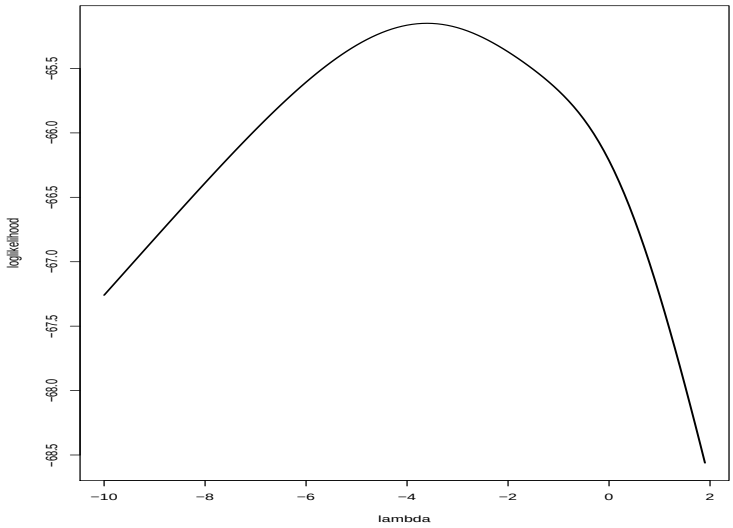

Figure 2.3: Bone marrow transplant study: : log likelihood vs λ (full model)

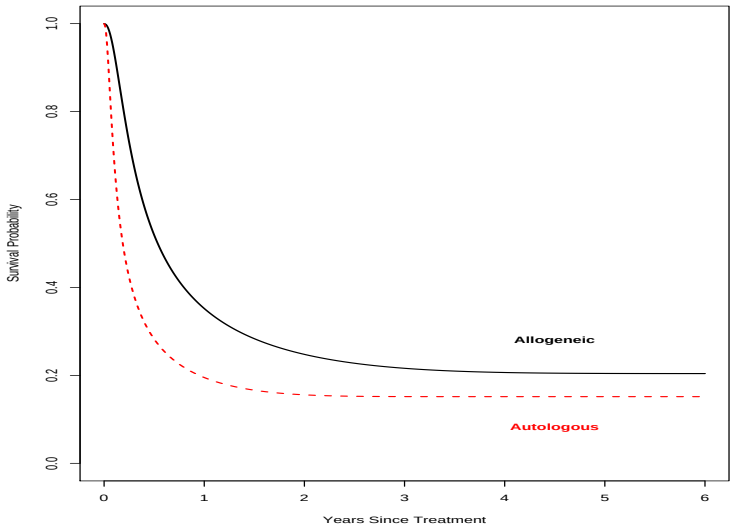

Figure 2.4: Bone marrow transplant study: predicted population survival probability (full model) at $\hat{\lambda}$

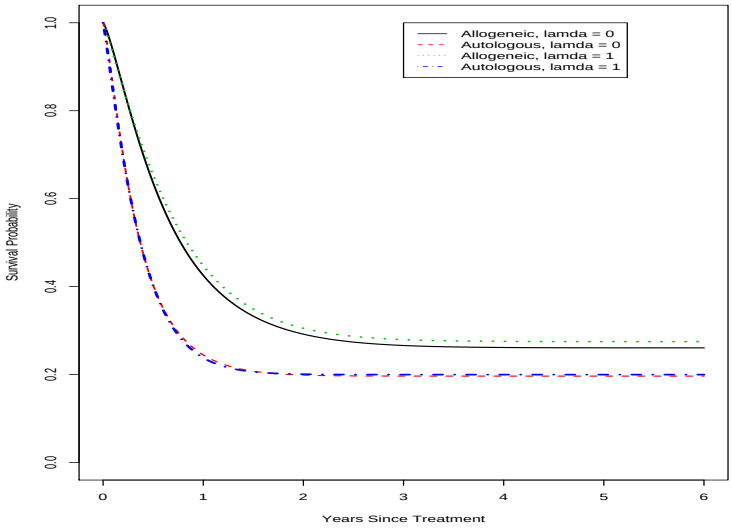

Figure 2.5: Bone marrow transplant study: predicted population survival probability (full model) for a fixed λ

The parameter estimates suggest that the autologous treatment group has a higher risk of recurrence compared to the allogeneic group. However, this difference is not statistically significant based on the Wald test. Furthermore, the chance of being cured does not differ statistically between these two treatment methods. Fixing λ to either 0 or 1 does not change the above conclusions. We observe a big increase in terms of variance estimates due to estimating the index parameter λ. The likelihood ratio test of $\lambda = 1$ versus $\lambda = 0$ yields $LR = -2((-67.2591)(-66.2149)) = 2.0884 < 3.84$, and therefore an insignificant inference on the difference of the mixture model versus the bounded cumulative hazard model (Figure 2.3 through 2.5).

This lack of significance may be due to the strong correlation among parameters γ, τ, and λ. The estimated correlation coefficient can be as high as 0.83 between $\hat{\gamma}$ and $\hat{\lambda}$. To address this issue, we considered the exponential distribution (a special case of Weibull distribution) for $S_0(t)$. Specifically, we fixed parameter γ to be 1, and refit our model with this assumption.

Again, the profile likelihood approach was used to estimate the model parameters.

Table 2.11: Bone marrow transplant study: parameter estimates (reduced model)

Parameters	$\lambda = 0$ (SE)	$\lambda = 1$ (SE)	$\lambda = \hat{\lambda}$ (SE with λ fixed)	$\lambda = \hat{\lambda}$ (SE with λ estimated)
α_0	0.29	0.27	0.26	
	(0.20)	(0.19)	(0.19)	(0.54)
α_1	0.18	0.21	0.21	
Autologous vs Allogeneic	(0.28)	(0.27)	(0.27)	(0.85)
β_1	0.48	0.65	0.67	
	(0.39)	(0.25)	(0.25)	(0.28)
τ	0.95	1.43	1.53	
	(0.25)	(0.28)	(0.29)	(0.47)
λ	0	1	1.20	
	—	—	–	(0.76)
$logL$	-69.71	-68.28	-68.25	

The allogeneic group was used as the reference group. Survival time was scaled into years. We searched in the range for $\lambda \in [-10, 5]$ with 0.01 intervals and obtained the maximum likehood at $\hat{\lambda} = 1.2$. Table 2.11 presents the parameter estimates along with their standard errors at different values of λ.

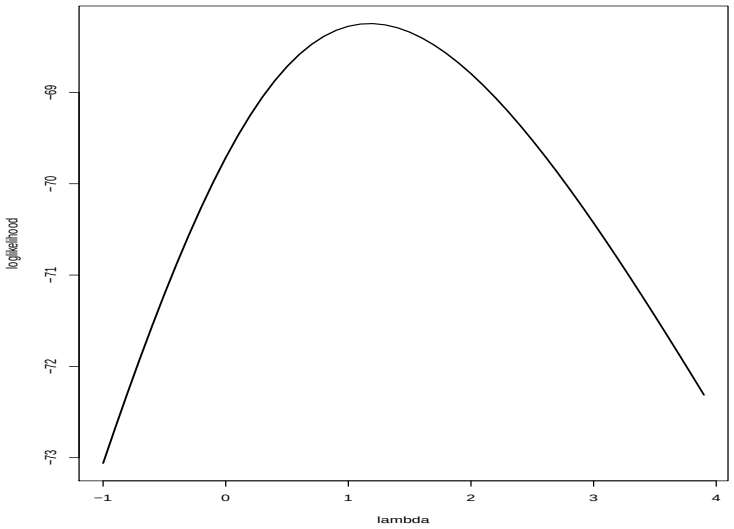

Figure 2.6: Bone marrow transplant study: log likelihood vs λ (reduced model)

The results suggest that there is no statistically significant difference in terms of the cure rate between these two treatment groups. However, the autologous

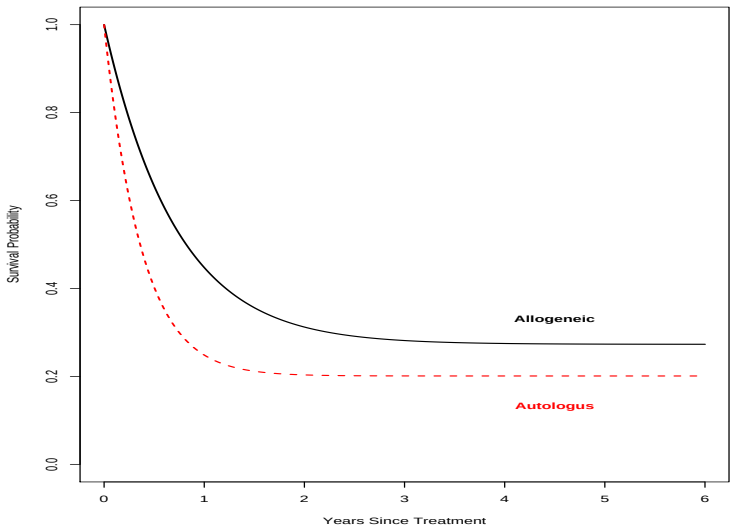

Figure 2.7: Bone marrow transplant study: predicted population survival probability (reduced model) at $\hat{\lambda}$

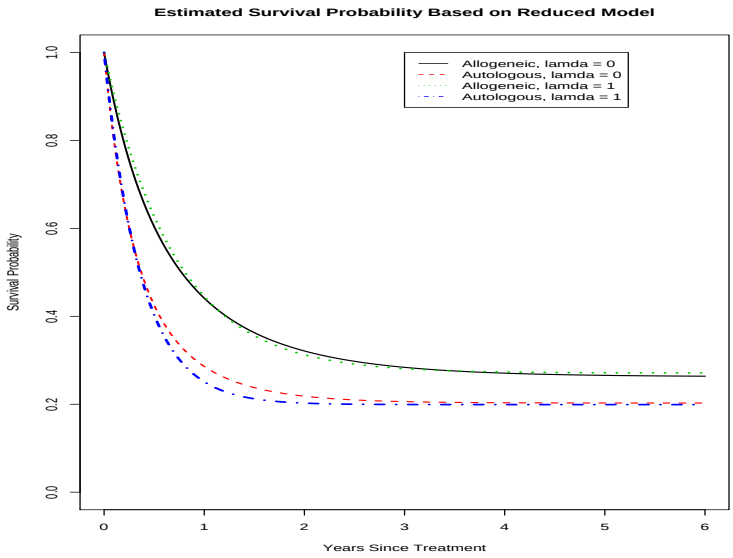

Figure 2.8: Bone marrow transplant study: predicted population survival probability (reduced model) at fixed λ

Table 2.12: Tonsil cancer Study: patient demographics part 1

Covariates	Mean	S.D.
Total Dose	58.18	9.36
Age (10 yr)	6.01	1.11
Duration	3.69	1.34

Table 2.13: Tonsil cancer study: patient demographics part 2

Covariates		Frequency	Percent
Gender	Female	201	29.91
	Male	471	70.09
Node	No	286	42.56
	Yes	386	57.44
T-Stage	1	96	14.29
	2	259	38.54
	3	254	37.80
	4	63	9.38

treatment group has a significantly higher risk of recurrence compared to the allogeneic group. The likelihood ratio test of $\lambda = 1$ versus $\lambda = 0$ yields $LR = 2((-68.2752)(-69.7141)) = 2.8778 < 3.84$, and therefore the difference between the two cure models is insignificant. Again we are not able to distinguish the mixture model from the bounded cumulative hazard model based on our model (Figure 2.6 through 2.8). Since our data set only has 91 patients, it is very difficult to obtain a precise estimate for the Box-Cox transformation parameter λ based on such a small sample. A larger sample size will be needed if λ is one of the parameters of interest.

2.5.2 Tonsil Cancer Study

To see how this model works in a larger data set, we also applied it to a tonsil cancer study which had 672 patients with 206 events. Patients were followed up to 15 years, and baseline information including age, gender, node, tumor stage, total dose, and treatment duration was collected along with the time from treatment to cancer recurrence. The patient demographics are presented in Table 2.12 and 2.13.

The Kaplan-Meier plot in Figure 2.9 shows that the survival distribution starts

to level off at around 3 years, and that there are about 70% patients who did not experience an event even with a 15-year follow-up period. It will be appropriate to use a cure model to fit this data.

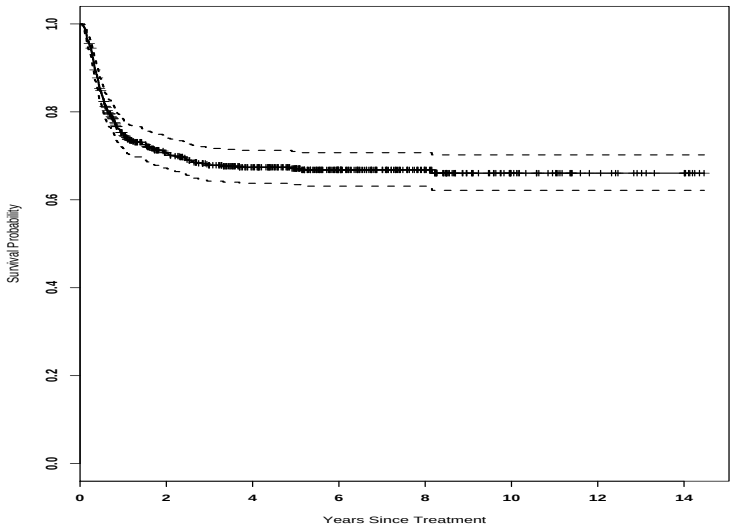

Figure 2.9: Tonsil cancer study: Kaplan-Meier plot

After dropping the insignificant covariates, we are left with the model presented in Table 2.14. The results show that compared to the T-stage 1 group, the T-stage 3 and 4 groups are significantly lower in the cure rate. Also higher total dose will increase the cure rate. There is a statistically significant association between node and cure rate as well. On average age has a negative effect on tonsil cancer survival probability. These findings are presented graphically in Figure 2.10 through Figure 2.14. If we limit $\lambda \in [0, 1]$, then the likelihood ratio test of $\lambda = 1$ versus $\lambda = 0$ yields $LR = 2(-507.868 - (-513.434)) = 11.13$, suggesting that, statistically speaking, the bounded cumulative hazard model is a better fit to these data compared to the mixture model, as shown in Figure 2.15. Instead of increased variance estimates due to estimating λ, we observe a decrease in variance estimates for the parameters in $p(Z)$ and stable variance estimates for the parameters in $S_0(t)$. We believe this is

Table 2.14: Tonsil cancer study: parameter estimates

Parameters	$\lambda = 0$ (SE)	$\lambda = 1$ (SE)	$\lambda = \hat{\lambda}$ (SE with λ fixed)	$\lambda = \hat{\lambda}$ (SE with λ estimated)
p				
Intercept	-1.02	-0.93	-1.17	
	(1.00)	(1.00)	(0.66)	(0.40)
T-stage 2	0.61	0.64	0.45	
	(0.58)	(0.94)	(0.46)	(0.29)
T-stage 3	1.42	1.43	1.17	
	(0.55)	(0.88)	(0.46)	(0.29)
T-stage 4	1.91	1.87	1.47	
	(0.68)	(1.00)	(0.47)	(0.30)
Node	0.43	0.40	0.35	
	(0.98)	(1.00)	(0.12)	(0.12)
Total Dose	-0.02	- 0.02	-0.01	
	(0.02)	(0.03)	(0.006)	(0.005)
S_0				
Age (per 10 yr)	-0.27	-0.25	-0.33	
	(0.04)	(0.11)	(0.05)	(0.06)
γ	1.17	1.09	1.52	
	(0.06)	(0.21)	(0.11)	(0.12)
τ	4.73	5.03	3.30	
	(1.00)	(1.00)	(1.00)	(1.01)
λ	0	1	-3.81	
	—	—	–	(0.99)
$logL$	-507.87	-513.43	-500.86	

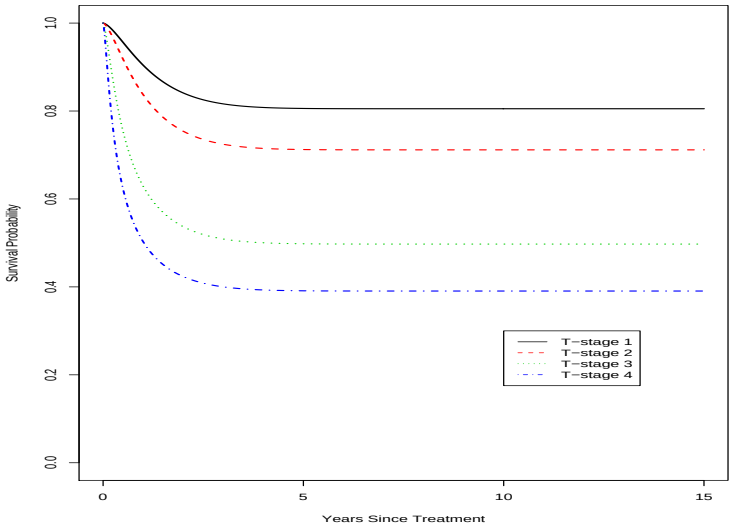

Figure 2.10: Tonsil cancer study: predicted population survival probability, λ = MLE, node = 1, dose = 58, age = 60

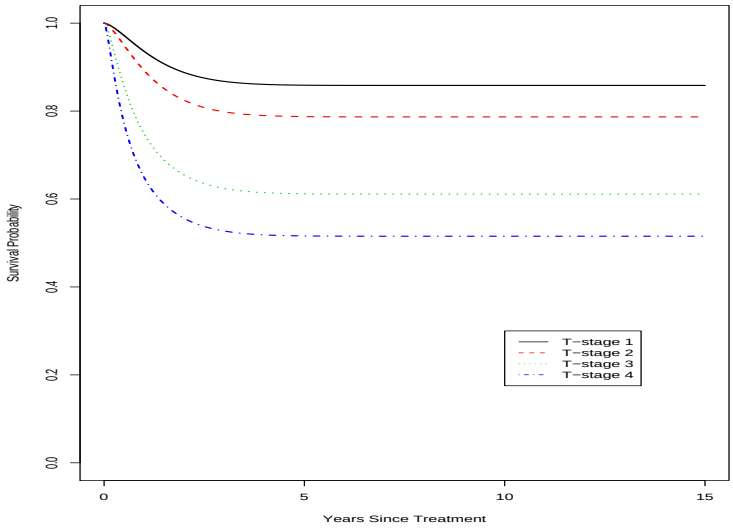

Figure 2.11: Tonsil cancer study: predicted population survival probability, λ = MLE, node = 0, dose = 58, age = 60

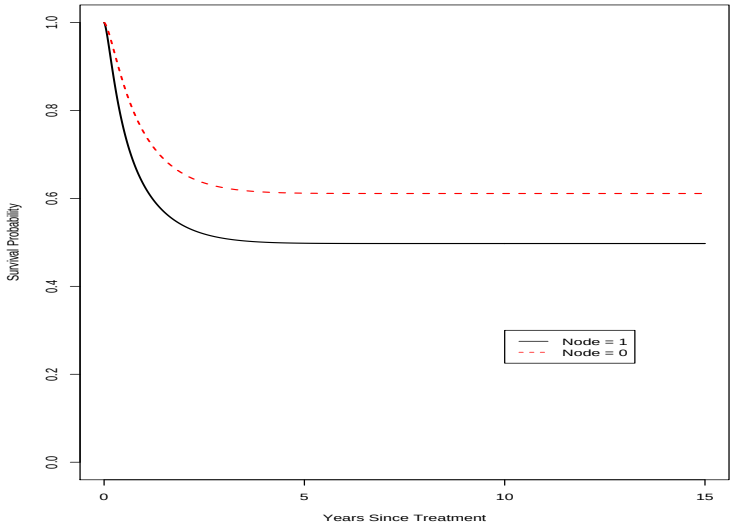

Figure 2.12: Tonsil cancer study: predicted population survival probability, λ = MLE, dose = 58, T-stage = 3, age = 60

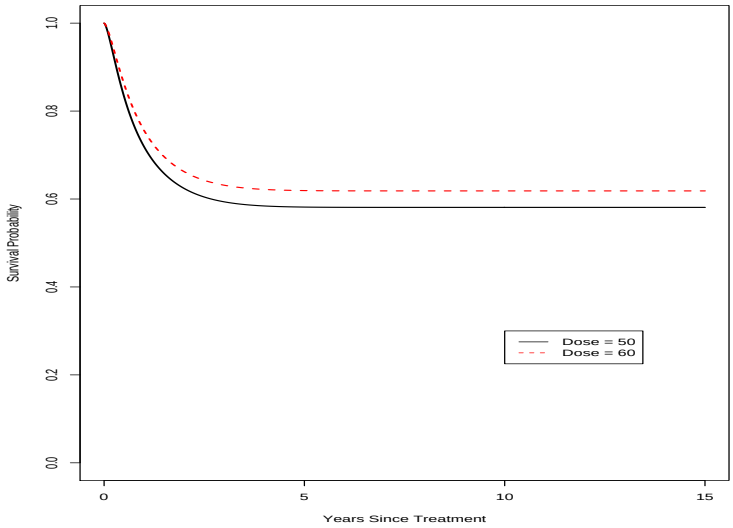

Figure 2.13: Tonsil cancer study: predicted population survival probability, λ = MLE, node = 0, T-stage = 3, age = 60

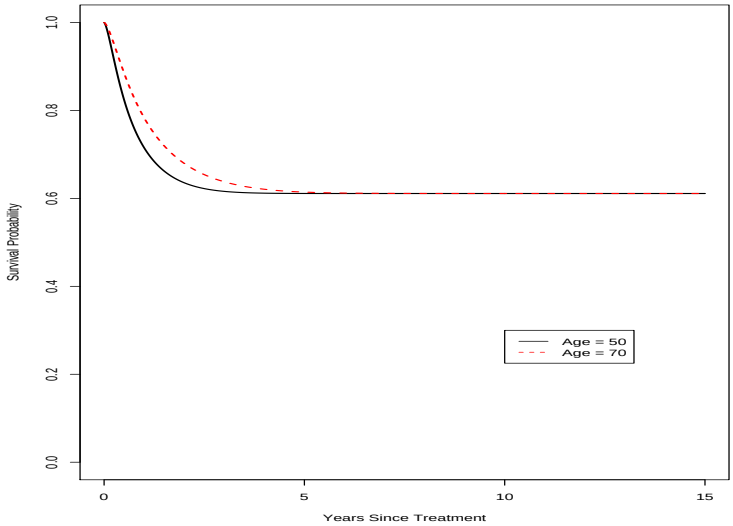

Figure 2.14: Tonsil cancer study: predicted population survival probability, λ = MLE, node = 0, T-stage = 3, dose = 58

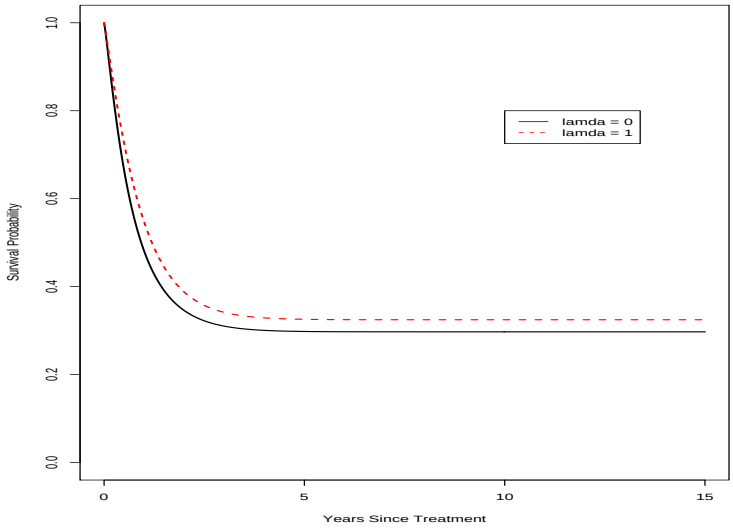

Figure 2.15: Tonsil cancer study: predicted population survival probability, λ = MLE, node = 1, T-stage = 4, dose = 58, age = 60

due to the fact that the covariates in these two parts of the model do not overlap. We observe the similar behavior in the simulation results (Table 2.1 - 2.4).

2.6 Discussion

In this chapter, we developed a general family of cure models indexed by a Box-Cox type transformation parameter λ. At different values of λ, the mixture model formulation and the bounded cumulative hazard model become special cases of this general family of cure model. By varying the values of λ, different intermediate cure models can be considered and estimated. We showed that unbiased parameter estimates can be obtained using the profile likelihood approach. The likelihood ratio test can be used to compare these two existing formulations of cure models.

The aim of our research is the development of a general family of cure models to include both the mixture model and the bounded cumulative hazard model. With the same general purpose, Yin and Ibrahim (2006) proposed a unified approach to model cure rate based on the transformation of the population survival function.

Similar to our model, Yin's approach also linked these two different formulations of cure models based on one index parameter. However, the cure rate p derived from Yin's approach is a function of the index parameter, and sometimes it is difficult to obtain a precise estimate of the index parameter. In contrast, we formulated our model in the way that the cure rate p does not depend on the index parameter λ, and λ can be considered as a nuissance parameter in certain studies. Hence, our estimate of the cure rate p is more robust to the different choices of λ.

When applying our model to the two real data sets, we observe that the maximum likelihood estimates of λ fall outside of the range [0,1]. When $|\lambda|$ becomes very large, it may introduce instability to the model estimation, and may also be difficult to interpret the results. An alternative to the choice of λ is to restrict λ to the range of [0,1], and thus we still have both the mixture model and the bounded cumulative hazard model as the two special cases (i.e., $\lambda = 0$ or 1), and we also have the intermediate cure models obtained from λ values between 0 and 1.

Our experience also indicates that for small samples, the likelihood function as a function of λ is often quite flat. Thus, obtaining an accurate estimate of λ requires a relatively large sample size, and may not be practical in many medical studies. However, when we do have relatively large sample sizes, we showed in our simulation studies that the profile likelihood approach can yield unbiased parameter estimates, and estimating λ does not necessarily cause a significant increase in the variance estimates of the parameters of interest.

Our family of cure models assumes a log-log link for the cure rate p. Future studies could explore the possibility of logit link for p. Our model also assumes a Weibull form for $S_0(t)$. It will be interesting to see other forms for $S_0(t)$ such as semiparametric or nonparametric forms. For parameter estimation, we chose the maximum likelihood

approach, but future investigations using Bayesian analysis could be performed to obtain the posterior distributions for the parameters.

CHAPTER III

Joint Non-parametric Models of Longitudinal and Survival Data

3.1 Introduction

In recent years, there have been an increasing number of medical studies collecting both longitudinal biomarkers and survival information. In medicine, a biomarker is frequently used as an indicator of a particular state within disease progression. Longitudinal biomarkers are those markers measured repeatedly over a certain time period. For example, in prostate cancer research patients are often followed starting from cancer treatment until an event such as tumor recurrence or death occurs, or the patients are lost to follow-up. Here, the PSA values are the biomarkers that are measured repeatedly over the follow-up period. Another well-cited example is in AIDS research where biomarkers such as CD4 lymphocyte counts or RNA viral loads are measured routinely during patients' visits over the follow-up period. Along with these biomarkers, we also know the survival information such as the length of follow-up for each patient and whether each patient experiences the event. This information is typically referred to as survival data. Depending on the research question, a mixed effects model is often used to model the longitudinal biomarkers, and a separate Cox proportional hazard model is used for the survival data analysis.

Instead of modeling the longitudinal biomarkers and survival data separately, joint

modeling of the longitudinal biomarkers and survival data has gained growing interest recently. In the literature, two approaches have been proposed to carry out the joint modeling: a two-stage model (Raboud et al. 1993, Tsiatis et al. 1995, Bycott and Taylor 1998, Dafni and Tsiatis 1998) and a likelihood based joint model(DeGruttola and Tu 1994, Faucett and Thomas 1996, Wulfsohn and Tsiatis 1997, Xu and Zeger 2001a, Xu and Zeger 2001b, Wang and Taylor 2001). It can be shown that the likelihood based joint model effectively corrects bias and improves the efficiency of parameter estimates, and therefore is preferrable to the two-stage model.

One important aspect of the joint modeling is the accurate representation of the longitudinal biomarkers. It is typically assumed that for each patient there is a true underlying biomarker process (curve) that is unobservable, and the observed longitudinal biomarkers are the true values plus some noise such as measurement errors. Past research has made extensive use of a linear mixed effects model with random intercepts and slopes to model the biomarker process. However, when applying the joint model to real medical data, we often observe a nonlinear trend in the longitudinal biomarker trajectories. To accomodate this behavior, the likelihood based joint model was extended to include various nonlinear parametric functional forms to better represent the biomarker process(Pauler and Finkelstein 2002, Wu 2003, Yu et al. 2004). These parametric assumptions are typically data driven, and often times are not flexible enough to capture the curvatures of longitudinal trajectories in different medical studies.

To address the issue of nonlinearity with a great degree of flexibility, nonparametric smoothing techniques have been proposed to model the longitudinal trajectories (Brown, et al. 2005). Eiler and Marx (1996) proposed a penalized B-spline approach (P-splines) based on a number of equally spaced B-splines, in which the variability

of the associated coefficients is controlled by a penalty. In this research, we develop a joint model in which the longitudinal biomarkers are modeled using penalized B-splines following Eiler and Marx (1996) and Wen (2007) and linked with the survival data using a proportional hazard model with time dependent covariates. Following Fahrmeir, et al. (2004), a Bayesian algorithm is developed to fit the joint model. This model provides a more parsimonious covariance structure for the random effects in the longitudinal mixed model, and choosing the location of knots for the spline basis is no longer an issue. The proposed model is applied to a prostate cancer study.

The rest of the chapter is organized as follows. In section 2, we provide some background information. In section 3, we present the model and develop the Bayesian estimation methods. In section 4, simulation studies are presented. In section 5, we apply this model to a prostate cancer study. Section 6 gives some concluding remarks.

3.2 Background

In this section, we discuss the two main approaches used for joint modeling of the longitudinal and survival data. We describe the two-stage model and the likelihood based model, and we illustrate some of the difficulties in modeling the longitudinal biomarker trajectories.

3.2.1 Two-stage Model

The main goal of a two-stage model is to fit the survival model (usually a Cox model) with the true underlying longitudinal biomarker as a time-dependent covariate. The issue here is that we do not know what the true longitudinal biomarker is due to the periodic measurements of the biomarker and the measurement error. Therefore, in the first stage, a mixed effects regression model is used to represent the true longitudinal trajectory. Then at each event time, the biomarkers are assumed

to be missing data and are imputed using this mixed model for all patients in the risk set. This imputation is carried out across all event times and all risk sets. In the second stage, these imputed biomarker values are treated as the true values at the time of event and incorporated in the Cox model to estimate the risk of failure.

Tsiatis, et al. (1995) used this model on an AIDS study in which CD4 counts were the longitudinal biomarker and death was the event of interest. In stage one, a linear mixed model was used to model the logarithm of the CD4 counts and subsequently derive the imputed values. In stage two, a Cox regression model with time dependent covariates was fitted. They concluded that CD4 may not be a good surrogate marker for death in AIDS patients. Bycott and Taylor (1998) used a random intercept plus a Brownian motion error term to fit the CD4 data, and then performed the imputation on the event times. This research suggested that the two-stage model can reduce the bias in the parameter estimates in the survival model. However, it can be shown that this method does not use the survival information in modeling the longitudinal data, and therefore there is a loss of information and bias in parameter estimates.

3.2.2 Likelihood Based Joint Model

To better use the data, a likelihood based joint model has been proposed where the likelihood of both longitudinal biomarkers and survival times is maximized simultaneously. Let $Y_i(t)$ and $Z_i(t)$ be the observed and true biomarker values, respectively, for subject i at time t, $i = 1, ..., n$. Let m_i be the number of repeated measurements for subject i, so that $Y_i = (Y_{i1}, ..., Y_{im_i})$ is the vector of observed longitudinal biomarker values. Let T_i and Δ_i be the observed survival time and survival indicator for subject i. That is, $T_i = min(D_i, C_i)$ where D_i is the event time and C_i is the censoring time for subject i. Δ_i is 1 if subject i has an event, and 0 otherwise. Let X_i be the baseline covariates for subject i.

For the longitudinal biomarkers, it is often assumed that

$$Y_i(t) = Z_i(t) + e_i(t)$$

$$Z_i(t) = Z_i(t; \beta_i).$$

Often a random effects model is used to model the longitudinal data with the assumptions that $e_i(t) \sim N(0, \sigma_e^2)$ and $\beta_i \sim N(\beta, \Sigma)$.

For the survival data, the survival times T_i are linked with the longitudinal biomarkers through

$$\lambda(t) = g(Z_i(t), X_i; \theta),$$

where $\lambda(t)$ can be hazard function in the case of porportional hazard model or a one-to-one transformation of the survival times as in the accelerated failure time model.

With the above assumptions, the joint likelihood function can be written as

$$\prod_{i=1}^{n} \prod_{t=t_1}^{t_{m_i}} \int f(y_{it}|\beta_i, \sigma_e^2) f(\beta_i|\beta, \Sigma) f(T_i, \Delta_i|\theta, \beta_i) d\beta_i,$$

where $f(y_{it}|\beta_i, \sigma_e^2)$ is the normal density function of $Y_i(t)$, $f(\beta_i|\beta, \Sigma)$ is the density of the random effects which takes a multivariate normal form here, and $f(T_i, \Delta_i|\theta)$ is the density function derived from the survival model. By maximizing this joint likelihood function, we can simultaneously estimate the parameters that control the longitudinal biomarker process and those that describe the risk of failure, especially the parameter that links the these two processes together. By borrowing the information from each other, this joint modeling approach can correct bias and improve the efficiency of parameter estimates, and therefore lead to a more accurate estimation of the relationship between the longitudinal biomarker process and the risk of failure.

To apply this likelihood based joint model, DeGruttola and Tu (1994) analyzed an AIDS study data set using a linear mixed model with single-knot quadratic regression

splines to model the CD4 counts and a multiple linear regression model on survival data. Faucett and Thomas (1995) considered a linear mixed model for longitudinal biomarkers and proportional hazard model for the survival data, and the Gibbs sampling technique was used for parameter estimation. Wulfsohn and Tsiatis (1997) developed an EM algorithm to estimate this model. Xu and Zeger (2001a) adopted a latent variable approach instead of a linear mixed model and used the MCMC algorithm to estimate the parameters. Xu and Zeger (2001b) extended their latent variable model to include multiple longitudinal biomarkers, and again the MCMC algorithm was implemented for estimation. Wang and Taylor (2001) proposed to model the longitudinal data with an integrated Ornstein-Uhlenbeck stochastic process such that the mixed effects model and the Brownian motion structure are both special cases, and an MCMC algorithm was developed to estimate the model.

These studies showed that the maximum likelihood estimates of the parameters using the likelihood based method are efficient and unbiased, whereas those using two-stage model are biased and inefficient. Due to the advantages of the likelihood based joint model, we focus on this approach throughout the rest of this chapter.

3.2.3 Nonlinearity and Variation in the shape of Longitudinal Biomarker Trajectories

One important aspect of joint modeling is to model the longitudinal biomarkers. As shown in Section 3.2.2, it is typically assumed that for each patient there is a true underlying biomarker process (curve) that is unobservable to us, and the observed longitudinal biomarkers are the true values plus some noise such as measurement error. In the past, a linear mixed effects model with random intercepts and slopes has been extensively used to model the biomarker process. However, when applying the joint model to real medical data, we often observe a nonlinear trend in the longitudinal biomarker trajectories. For example, in prostate cancer studies, the PSA

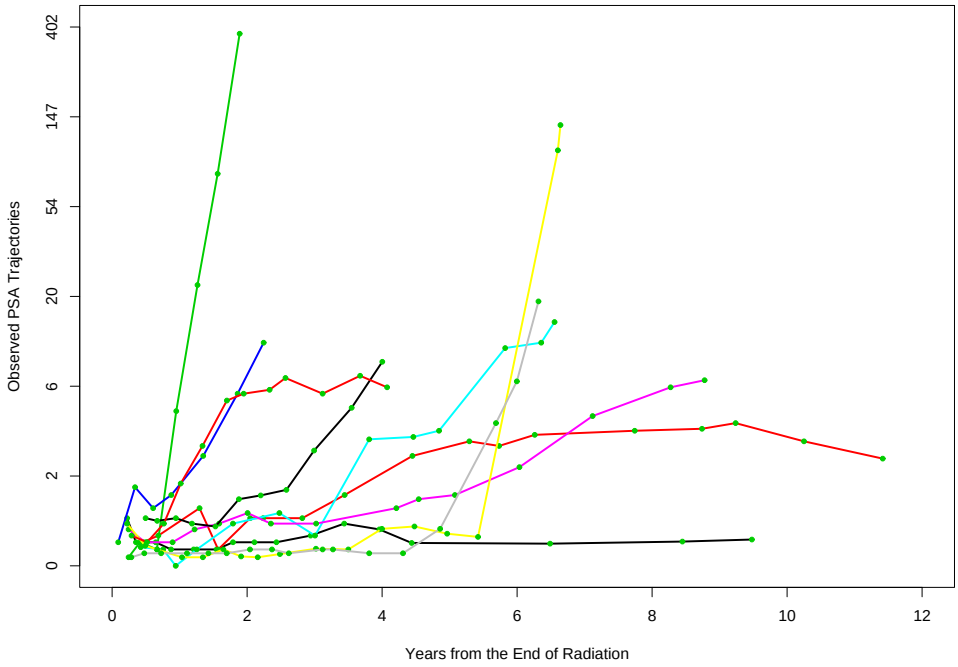

Figure 3.1: Ten sample patients with different nonlinear PSA profiles

values for a patient may have a nonlinear trend. Figure 3.1 shows the longitudinal trajectories of PSA following radiation therapy for 10 random patients. All of these ten patients did not receive hormonal therapy initially, but a few of them received hormonal therapy later in the course of the disease progression. We can clearly see the nonlinear pattern of the PSA trajectories for some of these patients in this graph.

To deal with the nonlinearity, Pauler and Finkelstein (2002) modeled the PSA trajectory with a two-piece linear spline within a non-linear hierarchical Bayesian approach. Yu, et al. (2004) used a double exponential mixed effect model to describe the nonlinear PSA profile. Similarly, Wu (2003) used a nonlinear (multiphase exponential decay) mixed effects model to describe the CD4 counts decline after HIV infection.

By introducing nonlinearity into the longitudinal models, these approaches suc-

cessfully modeled the biomarker process for their specific data sets. However, their parametric assumptions are typically data driven, and may not be flexible enough to capture the curvatures of longitudinal trajectories when there is a large variation in the shape of the curves. Again using Figure 3.1 as an example, we can see that the PSA trajectories typically start low due to the effect of radiation therapy. As time passes, some patients see an exponential increase in their PSA values, while other patients' trajectories remain steadily low. In other words, the shape of PSA trajectories can vary significantly among different patients. In this situation, the previously mentioned longitudinal models may not be adequate to model the biomarker process. Therefore, a flexible model is required to accomodate the nonlinearity and variation of the longitudinal trajectories.

To address the issue of nonlinearity with a greater degree of flexibility, nonparametric smoothing techniques have been proposed to model the longitudinal trajectories. There are many different types of smoothing methods, including spline smoothers, kernel smoothers, local regression methods, wavelets, running medians, etc. Brown, et. al. (2005) proposed to use cubic B-splines to model the profiles of CD4 counts and a Cox model to link the longitudinal biomarkers to the risk of failure. The number of knots for the cubic B-splines was selected using the Conditional Predictive Ordinate (CPO) and the Deviance Information Criterion (DIC).

While the cubic B-splines provide a flexible and robust estimation of the longitudinal model, the random effects can have a large number of parameters in the covariance structure, and it may be difficult to determine the number and the location of knots for the B-splines. Eiler and Marx (1996) proposed a penalized B-spline approach (P-splines) based on a number of equally space B-splines, in which the variability of the associated coefficients is controlled by a penalty. Specifically, let $(x_i, y_i), i = 1, ..., n$,

be the observed data. In B-spline context, we write $\hat{y}(x) = \sum_{j=1}^{m} \hat{\alpha}_j B_j(x;q)$ where $B_j(x;q)$ is the jth B-spline of degree q. The least square objective function with a continuous smoothing penalty is

$$S = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} \alpha_j B_j(x;q))^2 + \lambda \int_{x_{min}}^{x_{max}} (\sum_{j=1}^{m} \alpha_j B_j''(x;q))^2 dx.$$

This continuous second order derivative can be approximated by the finite discrete difference function. That is,

$$h^2 \sum_{j=1}^{m} \alpha_j B_j''(x;q) = \sum_{j=1}^{m} \Delta^2 \alpha_j B_j(x;q-2),$$

where h is the distance between knots, and $\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-1}$, a second order difference function. Therefore, the least square objective function with the order k finite difference penalty can be written as

$$S = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} \alpha_j B_j(x;q))^2 + \lambda \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2,$$

where k is the kth order of the difference function Δ. In the same way, the penalized likelihood function is

$$logL = logL(y;\alpha) + \frac{\lambda}{2} \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2.$$

In the frequentist setting, we can maximize this penalized likelihood function to obtain the maximum penalized likelihood estimates (MPLE) of the parameters. Wen (2007) applied this P-spline method to joint modeling of longitudinal and survival data and derived the MPLE this way. On the other hand, from the Bayesian perspective, we can specify a prior distribution for the parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ in the form of

$$p(\alpha|\tau^2) \propto exp(\frac{\alpha K \alpha'}{2\tau^2}),$$

where K is a penalty matrix that controls for the smoothness or the abrupt changes in the objective function, and hence the matrix K serves the same purpose as the Δ^k in the penalized likelihood function above. Here τ^2 is like the smoothing parameter in the frequentist setting. Fahrmeir, et. al. (2004) extended the penalized B-splines model to space-time data and developed empirical Bayesian algorithm to estimate the parameters.

In this chapter, we develop a joint model by adopting the Penalized B-spline method to model the longitudinal biomarkers and link with the survival data using proportional hazard model with time dependent covariates. A Bayesian algorithm is developed to derive the posterior distribution, and the MCMC technique is used to estimate the parameters.

3.3 Model and Method

3.3.1 Longitudinal Penalized B-Spline Model

Let $Y_i(t)$ be the observed biomarker value for subject i at time t, and $Z_i(t)$ be the true underlying longitudinal trajectory for subject i, $i = 1, ..., n$. We define

$$Y_i(t) = Z_i(t) + e_i(t) \tag{3.1}$$

where $e_i(t)$ is the measurement error which is assumed to be independently distributed from $N(0, \sigma_e^2)$. Following Rice and Wu (2001),

$$Z_i(t) = \sum_{k=1}^{q_1} B_{1k}(t)\beta_k + \sum_{k=1}^{q_2} B_{2k}(t)b_{ik} \tag{3.2}$$

where B_{1k} and B_{2k} are, respectively, the q_1 and q_2 dimensional B-spline basis functions on [0, T] with some fixed equally spaced knots. Here T is the maximum follow-up time for all subjects. The quantity $\sum_{k=1}^{q_1} B_{1k}(t)\beta_k$ defines the fixed effects representing the population mean curve, and the quantity $\sum_{k=1}^{q_2} B_{2k}(t)b_{ik}$ defines the

random effects capturing the individual variation. In matrix notation,

$$Z_i(t) = B_1(t)\beta + B_2(t)b_i, \tag{3.3}$$

where β is a column vector of $(\beta_1, \beta_2, ..., \beta_{q_1})$, and b_i is a column vector of $(b_{i1}, b_{i2}, ..., b_{iq_2})$.

To control the variability of the coefficients β and b_i and hence the smoothness or curvature of the population and individual trajectories, we apply a penalty term to these coefficients. Let $K_1 = D_1'D_1$ and $K_2 = D_2'D_2$ be the penalty matrix where D_1 and D_2 are the second order difference matrices in the form of

$$D = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & & & & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

We can then decompose the coefficients β and b_i into unpenalized and penalized parts. That is,

$$\beta = \Psi_1^{UN}\beta^{UN} + \Psi_1^P\beta^P \tag{3.4}$$

$$b_i = \Psi_2^{UN}b_i^{UN} + \Psi_2^P b_i^P \tag{3.5}$$

where the matrix Ψ_1^{UN} is a $q_1 \times 2$ matrix with the first column being constant 1 and the second column containing the null space of the penalty matrix K_1, the matrix Ψ_1^P is a $q_1 \times (q_1 - 2)$ matrix defined by $\Psi_1^P = D_1'(D_1D_1')^{-1}$, β^{UN} is a vector of dimension 2×1, and β^P is a vector of dimension $(q_1 - 2) \times 1$. Similarly, the matrix Ψ_2^{UN} is a $q_2 \times 2$ matrix containing the null space of the penalty matrix K_2, the matrix Ψ_2^P is a $q_2 \times (q_2 - 2)$ matrix defined by $\Psi_2^P = D_2'(D_2D_2')^{-1}$, b_i^{UN} is a vector of dimension 2×1, and b_i^P is a vector of dimension $(q_2 - 2) \times 1$. Clearly, β^{UN} and b_i^{UN} represent the parts of β and b_i that are not penalized by the penalty matrix K_1 and K_2, whereas

β^P and b_i^P represent the deviations of the parameters β and b_i from the null space of the penalty matrix K_1 and K_2. In other words, the unpenalized part controls the linear trend of the longitudinal trajectory through intercept and slope terms, and the penalized part represents the curvature and smoothness of the longitudinal trajectory through the penalties to the coefficients. Later in this chapter, we will develop a Bayesian algorithm where prior densities will be placed on the parameters β^P, b_i^{UN}, and b_i^P to control the smoothness of both the population and the individual trajectories.

Let X_i be the p baseline covariates including the intercept. We further specify that the fixed effects β^{UN} depend on these baseline covariates, that is,

$$\beta_i^{UN} = X_i\gamma, \tag{3.6}$$

where γ is some $p \times 2$ matrix of fixed effects parameters. Therefore, the complete longitudinal model based on the penalized B-splines is

$$Y_i(t) = B_1(t)\Psi_1^{UN}X_i\gamma + B_1(t)\Psi_1^P\beta^P + B_2(t)\Psi_2^{UN}b_i^{UN} + B_2(t)\Psi_2^P b_i^P + e_i(t) \tag{3.7}$$

3.3.2 Survival Model

Let T_i and Δ_i be the observed survival time and survival indicator for subject i. That is, $T_i = min(D_i, C_i)$ where D_i is the event time and C_i is the censoring time for subject i. Δ_i is 1 if subject i has an event, and 0 otherwise. We assume the longitudinal data is linked to the risk of failure through a proportional hazard model with time dependent covariates,

$$\lambda(t) = \lambda_0(t)exp(\alpha Z_i(t) + X_i\phi). \tag{3.8}$$

Here, $\lambda_0(t)$ is the baseline hazard function, α is the coefficient that links the true longitudinal curve to the risk of failure, and ϕ is the parameter that links the baseline covariates to the risk of failure.

3.3.3 Joint Model

Let $\theta = (\gamma, \beta^P, \sigma_e^2)$, and $\omega_i = (b_i^{UN}, b_i^P)$.

For subject i, the contribution to the likelihood from the longitudinal model is

$$\begin{aligned} L(\theta, \omega_i | y_i(t)) &= \frac{1}{\sqrt{2\sigma_e^2}} exp(-\frac{1}{2\sigma_e^2}[y_i(t) - B_1(t)\Psi_1^{UN} X_i\gamma - B_1(t)\Psi_1^P\beta^P \\ &\quad -B_2(t)\Psi_2^{UN} b_i^{UN} - B_2(t)\Psi_2^P b_i^P]^2) \end{aligned} \tag{3.9}$$

The contribution to the likelihood from the survival model is

$$\begin{aligned} L(\alpha, \phi | T_i, \Delta_i) &= \lambda_0(T_i)^{\Delta_i} exp(\Delta_i[\alpha Z_i(T_i) + X_i\phi]) \\ &\quad exp(-\int_0^{T_i} \lambda_0(u) exp(\alpha Z_i(u) + X_i\phi)\, du) \end{aligned} \tag{3.10}$$

The likelihood function conditional on the data from the joint longitudinal and survival model is

$$\begin{aligned} L(\theta, \omega_1, ...\omega_n, \alpha, \phi | Y, T, \Delta) &= \prod_{i=1}^{n} \prod_{t=t_1}^{t_{m_i}} \frac{1}{\sqrt{2\sigma_e^2}} exp(-\frac{1}{2\sigma_e^2}[y_i(t) - B_1(t)\Psi_1^{UN} X_i\gamma \\ &\quad -B_1(t)\Psi_1^P\beta^P - B_2(t)\Psi_2^{UN} b_i^{UN} - B_2(t)\Psi_2^P b_i^P]^2) \\ &\quad \lambda_0(T_i)^{\Delta_i} exp(\Delta_i[\alpha Z_i(T_i) + X_i\phi]) \\ &\quad exp(-\int_0^{T_i} \lambda_0(u) exp(\alpha Z_i(u) + X_i\phi)\, du). \end{aligned} \tag{3.11}$$

3.3.4 Priors and Posterior Distributions

To control the smoothness of both the population and the individual longitudinal trajectories, we apply the Bayesian approach to the joint model. The prior distributions for the longitudinal model parameters are assumed to be $\gamma \sim$ *constant*, $b_i^{UN} \sim N(0, \Sigma)$, $\beta^P \sim N(0, \sigma_\beta^2)$, and $b_i^P \sim N(0, \sigma_b^2)$, where Σ is a 2×2 matrix in the form of

$$\Sigma = \begin{pmatrix} \sigma_{b1}^2 & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b2}^2 \end{pmatrix}.$$

Compared to Brown et. al. (2005), these assumptions yield a more parsimonious covariance structure for the random effects. For the other parameters in the joint model, we assume the following prior densities: $\sigma_e^2 \sim IG(a, b)$, $\alpha \sim N(0, \sigma_\alpha^2)$, and $\phi \sim N(0, \sigma_\phi^2 I)$. Furthermore, we also assume conjugate hyperpriors for $\sigma_\beta^2 \sim IG(\kappa_1, \kappa_2)$, $\sigma_b^2 \sim IG(\upsilon_1, \upsilon_2)$, and $\Sigma \sim IW_\nu(V)$. Here, we choose noninformative priors where a = 0.001, b = 0.001, $\sigma_\alpha^2 = 100$, $\sigma_\phi^2 = 100$, $\kappa_1 = 0.001$, $\kappa_2 = 0.001$, $\upsilon_1 = 0.001$, $\upsilon_2 = 0.001$, $\nu = 2$,

$$V = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

The posterior distribution is proportional to the product of the likelihood function and the prior densities of the parameters, that is,

$$\begin{aligned} &Ł(\theta, \omega_1, ...\omega_n, \alpha, \phi | Y, T, \Delta) \\ &\times f(\sigma_e^2) f(\gamma) f(b_i^{UN}|\Sigma) f(\beta^P|\sigma_{\beta^P}^2) f(b_i^P|\sigma_{b^P}^2) f(\alpha) f(\phi) f(\sigma_{\beta^P}^2) f(\sigma_{b^P}^2)][\Sigma) \end{aligned} \tag{3.12}$$

3.3.5 MCMC Estimation

Some parameters have closed form conditional posterior distributions.

$$\begin{aligned} [\sigma_e^2 \mid \bullet] &\sim IG(a + \frac{N}{2}, b + \frac{\sum_i \sum_t (y_{it} - Z_{it})^2}{2}) \\ [\sigma_{\beta^P}^2 \mid \bullet] &\sim IG(\kappa_1 + \frac{N}{2}, \kappa_2 + \frac{N(\beta^P)'(\beta^P)}{2}) \\ [\sigma_{b^P}^2 \mid \bullet] &\sim IG(\upsilon_1 + \frac{N}{2}, \upsilon_2 + \frac{\sum_i (b_i^P)'(b_i^P)}{2}) \\ [\Sigma \mid \bullet] &\sim IW_{N+\nu}(\sum_i (b_i^{UN})(b_i^{UN})' + V) \end{aligned}$$

For the rest of the parameters, since it is not feasible to sample directly from the posterior conditional distribution, we implement the Metropolis-Hastings (MH) algorithm to build a Markov chain of random samples from the posterior distributions. The MH algorithm requires sampling from a known proposal density and uses MH

steps to update the MCMC chain. Here, we use a normal proposal density with the accepted parameter values from the previous step as the new mean, and the variance of the proposal normal density is tuned to achieve a certain acceptance rate (20% and above). Due to the large number of parameters, we use Gibbs sampling algorithm to update each parameter from the conditional distribution given all the other parameter estimates from the previous step. The autocorrelation of the MCMC chain is taken into account by identifying the appropriate burn-in period and only keeping every tenth value in the chain. Trace plots are visually examined for proper mixing of chains. Different starting values are used to examine the convergence.

3.4 Application

We applied our method to prostate cancer data that we obtained from the Radiation Therapy Oncology Group and the BC Cancer Agency in Vancouver. The data included both repeated longitudinal measurements of PSA values and survival information such as time of tumor recurrence (local, nodal, and distant metastasis), death from prostate cancer, and loss to follow-up. Also included was baseline information such as age, tumor stage, baseline PSA value, Gleason score, radiation dose, and radiation duration. We focus our attention to those patients who received both radiation therapy and planned hormonal therapy as the primary treatment method because the longitudinal PSA trajectories from these patients had a much larger variation in shape compared with those from the patients who did not receive initial hormonal therapy. Due to the heterogeneity in the shapes of the PSA trajectories, the parametric models have great difficulty fitting these data.

3.4.1 The Data

The data we used in our application consisted of 2434 patients in total, 280 from study RTOG9406, 560 from study RTOG9202, 517 from the BC Cancer Agency, and 1077 from study RTOG9413.

RTOG9202 study was a phase III clinical trial to evaluate the effectiveness of elective versus therapeutic androgen deprivation and disease progression. Study subjects recruited for this trial were late stage prostate cancer patients. They had a clinical T-stage ranging from T2C, T3 and T4. About 20% of these subjects had a Gleason score between 2 and 5, about 21% had a Gleason score of 6, and 59% had a Gleason score of 7 to 10. The logarithm transformed baseline PSA values of these subjects had an average of 3.1 with a range of 0.12 to 5.53. All patients in this study were administered hormonal therapy before radiation. About half of them got a short period of planned hormonal therapy which ended at the end of radiation, and the other half got a longer period of planned hormonal therapy which ended about two years after the end of radiation therapy. The 560 patients we chose to include in our application were those who received the shorter schedule of planned hormonal therapy for a larger variation of the shapes of the longitudinal PSA trajectories.

Study RTOG9406 was a phase I/II clinical trial with a dose escalation schedule intended to find the maximum tolerable dose. The study subjects in this trial were primarily early and mid stage prostate cancer patients. Roughly 46% of the patients had a tumor stage of 1, 44% had a tumor stage of 2, and only 10% had a tumor stage of 3 and 4. About 11% of these patients had a Gleason score between 2 and 5, 78% had a Gleason socre of 6 and 7, and only 11% had a Gleason score of 8 to 10. The logarithm transformed baseline PSA values of these subjects had an average of 3.26 with a range of 0.09 to 4.25. For the same purpose as above, we chose 280

patients who had induction hormonal therapy ending within one year from the end of radiation to include in our application.

Study RTOG9413 was a phase III clinical trial to compare the survival probability among radiation treated patients with different types of planned hormonal therapies. The study subjects recruited into this trial were mostly mid to late stage prostate cancer patients. Among these patients, about 14% have a tumor stage of 1, 52% had a tumor stage of 2, and 33% had a tumor stage of 3 and 4. In terms of the Gleason score, 2% patients had a Gleason score between 2 and 5, 25% had a Gleason score of 6, 44% had a Gleason score of 7, and 29% had a Gleason score of 8 to 10. The logarithm transformed baseline PSA values of these subjects had an average of 3.1 with a range of 1.1 to 4.6. Again, for the same reason, 1077 patients with planned hormonal therapy ending within one year from the end of radiation were chosen to be included in our application.

We also received prostate cancer data from the BC Cancer Agency. The study subjects recruited into this cancer registry were mostly early to mid stage prostate cancer patients. Among these patients, about 22% have a tumor stage of 1, 61% had a tumor stage of 2, and 17% had a tumor stage of 3 and 4. In terms of the Gleason score, 39% patients had a Gleason score between 2 and 5, 32% had a Gleason score of 6, 22% had a Gleason score of 7, and 6% had a Gleason score of 8 to 10. The logarithm transformed baseline PSA values of these subjects had an average of 2.6 with a range of 0.3 to 5.4. The 517 study subjects we included in our study all received planned hormonal therapy ending within one year from the end of radiation therapy.

By combining these four cohorts together, we had a well balanced sample of patients with all stages of cancer progression. While the hormonal therapy tended to

keep PSA values low while it was in effect, the restriction that planned hormonal therapy had to end within one year of the end of radiation therapy ensured a wide range of trends in the longitudinal PSA trajectories. After the end of planned hormonal therapy, some patients saw a fast exponential increase in PSA levels whereas other patients were able to maintain a steady low PSA profile, possibly due to the lingering effects of the initial hormonal therapy. As the residual effect from the initial hormonal therapy wore off, the PSA values of some patients started to grow higher after the plateau, whereas others remained low and leveled off all the way until the last day of observation. For these patients, the radiation therapy may have cured the patients by eliminating the cancer cells. Figure 3.1 shows the different patterns of some sample patients.

The patients with follow-up times of less than one year were excluded from the study. Out of these 2434 patients, we randomly sampled 1934 patients to create our training dataset, and the rest of the 500 patients became our test dataset. We used the training set to build our joint model and used the test set to assess the performance of the joint model. The training set consisted of 19527 PSA measurements. For the test set, patients were arbitrarily censored at 3 years or 5 years. Those who did not have follow-up time long enough to reach 3 years or 5 years were excluded from the test sets. This left 417 patients for the 3-year test set with 2943 PSA measurements, and 333 patients for the 5-year test set with 3275 PSA measurements.

Baseline information including T-stage, Gleason score, baseline PSA, radiation dose, radiation duration, and age were recorded together with salvage hormonal therapy information. The event of interest was defined to be clinical tumor recurrence, which included local recurrence, regional recurrence, and distant metastasis,

Table 3.1: Patients Demographics: T-stage, Gleason Score, Hormonal Therapy

Covariate		Training set	3-year test set	5-year test set
T-stage	1 or 2	1156 (59.77%)	256 (61.39%)	209 (62.76%)
	3 or 4	778 (40.23%)	161 (38.61%)	124 (37.24 %)
Gleason score	< 7	731 (37.79%)	156 (37.41%)	133 (39.94 %)
	7+	1203 (62.21%)	261(62.59%)	200 (60.06 %)
salvage hormonal therapy	Yes	244 (12.62%)	63 (15.11%)	50 (15.02%)
	No	1690 (87.38%)	354 (84.89%)	283 (84.99%)
Event	Yes	389 (20.11%)	57 (13.67%)	36 (10.81%)
	No	1545 (79.89%)	360 (86.33%)	297 (89.19%)

Table 3.2: Patients Demographics: Baseline PSA, Total Dose, Age, Treatment Duration

Covariate	Mean (S.D.)		
	Training set	3-year test set	5-year test set
log(Baseline PSA + 1)	2.93 (0.79)	2.86 (0.78)	2.89 (0.79)
Total Dose	70.16 (3.79)	70.11 (3.96)	70.00 (3.90)
Age	69.21 (6.58)	69.51 (6.26)	69.52 (6.33)
Treatment Duration (days)	53.95 (6.18)	54.08 (6.28)	53.86 (5.26)

or death from prostate cancer. Patients who did not have a clinical recurrence or did not die from prostate cancer were censored at the last contact date or date of death from other causes. Only PSA measurements prior to an event or censoring were used in the model. For patients who have later salvage hormonal therapy, PSA measurements after the date of hormonal therapy were excluded from modeling.

Table 3.1 presents the frequencies of the baseline T-stage, Gleason score, salvage hormonal therapy, and events. Table 3.2 presents the summary statistics of the continuous variables.

3.4.2 The Model Built on the Training Set

We follow the same joint model framework as described in the Section 3.3 to fit the training set using all available patients and all PSA measurements in this set. For simplicity, $B_1(t)$ and $B_2(t)$ are assumed to be the same B-spline basis with the same equally spaced knots, and therefore are both denoted as $B(t)$. We chose to have 11 equally spaced knots between 0 and 12.27 years which is the maximum follow-up

time in our training set. The baseline hazard was assumed to be a step function. All baseline covariates enter the longitudinal and survival models. In addition, the slope of the PSA trajectory calculated using the linear combination of the derivatives of the B-spline bases $B(t)$ also entered the proportional hazard model as a time-dependent covariate. An indicator variable was created to indicate the administration of salvage hormonal therapy, and is included in the survival model as well.

Therefore, the longitudinal model in (3.7) becomes

$$log(PSA_{it}+1) = B(t)\Psi^{UN}\beta_i^{UN} + B(t)\Psi^{P}\beta^{P} + B(t)\Psi^{UN}b_i^{UN} + B(t)\Psi^{P}b_i^{P} + e_{it} \tag{3.13}$$

$$\beta_i^{UN} = \begin{pmatrix} \gamma_{10} + \gamma_{11}log(BaselinePSA_i + 1) + \gamma_{12}Tstage_i + \gamma_{13}Gleason_i \\ \gamma_{20} + \gamma_{21}log(BaselinePSA_i + 1) + \gamma_{22}Tstage_i + \gamma_{23}Gleason_i \end{pmatrix}$$

and the survival model in (3.8) becomes

$$\begin{aligned} \lambda(t) &= \lambda_0(t)exp(\alpha_1 log(\hat{PSA}_i(t) + 1) + \alpha_2 \frac{\partial log(\hat{PSA}_i(t) + 1)}{\partial t} \\ &\quad + \phi_1 log(BaselinePSA_i + 1) + \phi_2 Tstage_i + \phi_3 Gleason_i + \phi_4 HT_i(t)). \end{aligned}$$

Here, $log(\hat{PSA}_i(t) + 1)$, the slope, and $HT(t)$ are time-dependent covariates. $HT(t)$ is a binary variable which takes 1 if a patient has had salvage hormonal therapy at time t and 0 otherwise. $Tstage$ takes value 0 if the tumor stage is 1 or 2, and 1 if the tumor stage is 3 or 4. $Gleason$ takes value 0 if the Gleason score is between 2 and 6, and 1 if the Gleason score is between 7 and 9. Note that the treatment duration, total radiation dose, and age are omitted from our model due to their statistically insignificant effects.

All priors are chosen to be noninformative priors as described in Section 3.4. To generate the starting values for the MCMC chains, we did preliminary analyses on

the data using a linear mixed model and a Cox proportional hazard model. We used PROC MIXED to fit the longitudinal model, and took the fixed effects output as the starting values for the MCMC chains for parameters γ and β^P. The starting values for the random effects parameters were set to zero. For some patients, we had to adjust the starting values of the random effects parameters to achieve convergence. We also used PROC PHREG to fit the Cox proportional hazard model with the last observed PSA value ($log(PSA(t)+1)$), the slope of the the PSA trajectory calculated using the last two observed PSA values, and the hormonal therapy status as the time dependent covariates, along with baseline PSA, T-stage and Gleason score in the model. We took the SAS output to be the starting values for the survival model parameters (α and ϕ). In the MH steps, the proposal densities were all normal densities with the accepted values from previous step as the new means, and the variances were selected to ensure an acceptance rate of minimum 20%. For example, the proposal density for γ_1 was $\gamma_1^{(s)} \sim N(\gamma_1^{(s-1)}, \sigma^2_{\gamma_1})$, where s was the s-th step.We ran 20000 iterations with the first 10000 iterations being burn-in period. For the rest of the 10000 iterations, we kept only the draws from every 10th iteration.

Our analysis concluded that age, total dose, and treatment duration did not have a statistical significant effect on the PSA value or risk of event, therefore they were not included in our final model. The parameter estimates from the training set are shown in table 3.3.

This analysis suggested a statistically significant positive association between the slope of PSA and risk of tumor recurrence or death from prostate cancer. Our previous analysis showed that without the slope of PSA in the model, the current PSA value has a strong positive association with the risk of failure. With the inclusion of the slope, the current PSA value is no longer significant. Also, the T-stage and

Table 3.3: Parameter Estimates

Parameter	Estimate	S.D.	Lower 2.5%	Upper 2.5%
Longitudinal Part				
Intercept	-0.05	0.04	-0.10	0.05
log(baseline PSA + 1)	0.14	0.01	0.12	0.16
T-stage	0.08	0.02	0.04	0.13
Gleason	-0.02	0.02	-0.06	0.03
Intercept*t	-0.42	0.04	-0.49	-0.35
log(baseline PSA + 1)*t	0.23	0.01	0.21	0.25
T-stage*t	0.18	0.01	0.16	0.22
Gleason*t	0.25	0.02	0.22	0.29
Survival Part				
T-stage	0.31	0.14	0.04	0.57
Gleason	0.32	0.15	0.04	0.63
log(baseline PSA + 1)	0.14	0.09	-0.02	0.31
PSA(t)	-0.01	0.03	-0.06	0.06
Slope of PSA(t)	1.11	0.12	0.90	1.31
Salvage Hormonal Therapy	0.001	0.17	-0.33	0.33

Gleason score are both significant, and higher T-stage or Gleason score is associated with higher risk of failure. Baseline PSA and salvage hormonal therapy are not significant in our model. One possible explanation is that all patients in our data have had initial planned hormonal therapy, and these patients may have developed a resistance to further hormonal therapy, thereby eliminating the effectiveness of the salvage hormonal therapy in reducing the risk of tumor recurrence or death.

Due to the P-spline formulation of our longitudinal model, it was difficult to interpret each individual parameter estimate associated with the B-spline basis. Instead, we used the Bayesian draws to generate 1000 estimated PSA trajectories for each individual. Figure 3.2 and 3.3 show the median and 95% credible interval for 8 sample longitudinal trajectories. The solid vertical lines are the times of event or censoring. The dashed vertical lines are the times of salvage hormonal therapy. The little stars in the graphs are the value of $log(baselinePSA + 1)$. Table 3.4 summarizes the average percentage of observed PSA values that fall within the 95% credible interval. There are altogether 19527 PSA values in the training set, and 18629(98.40%) of

Table 3.4: percentage of observed PSA values within the 95% CI

	Training set	3-year test set	5-year test set
Total # observed PSA	19527	1633	655
Percentage within 95% CI	95.40%	96.02 %	94.50 %
Total # patients	1934	343	228
Average percentage within 95% CI	94.77 %	94.69 %	95.29 %

them fall within the 95% credible interval. On the individual level, we can calculate the coverage rate of the credible interval for each patient, and then compute the average of these coverage rates over 1934 patients in the training set. On average a patient has about 94.77% of the observed PSA measurements included within the 95% credible interval.

To check the proper mixing of the Markov chains, trace plots are graphed and visually examined. Figures 3.4 show eight samples of the trace plots. We can see that the correlation among consecutive iterations of the Markov chains is not very strong.

3.4.3 Validation Based on the Test Set

We then made predictions for the patients in the test set using the joint model that we built from the training set. We used the samples of fixed effects parameters (γ, β^P, α, ϕ, σ_e^2, σ_b^2, Σ) from the training set model, and ran smaller size MH draws on the test sets to generate MCMC chains for the random effects parameters (b_i^{UN} and b_i^P). We generated 200 MH draws of these random effects parameters based on each fixed effects parameter draw, and kept the last value. This yielded 1000 samples of every parameter in our model. We then calculated the predicted individual PSA trajectory beyond the censoring year (3 or 5 year) until year 12.27 and their credible interval (Figure 3.5 and 3.6) based on the longitudinal model of (3.11) with the same baseline covariates as in the training model. Table 3.4 summarizes the average

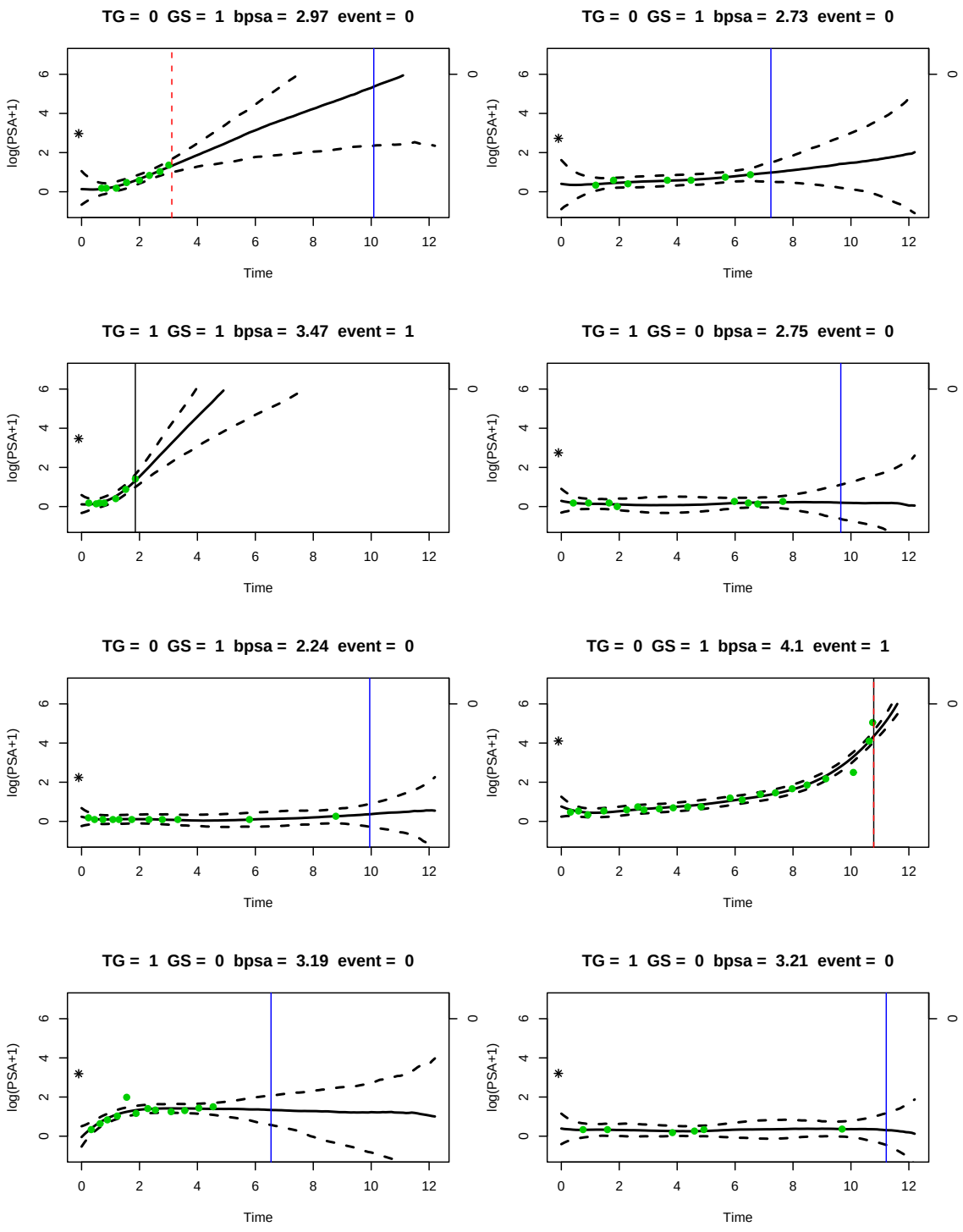

Figure 3.2: Sample patients with observed PSA values and estimated PSA trajectories and their 95% credible intervals

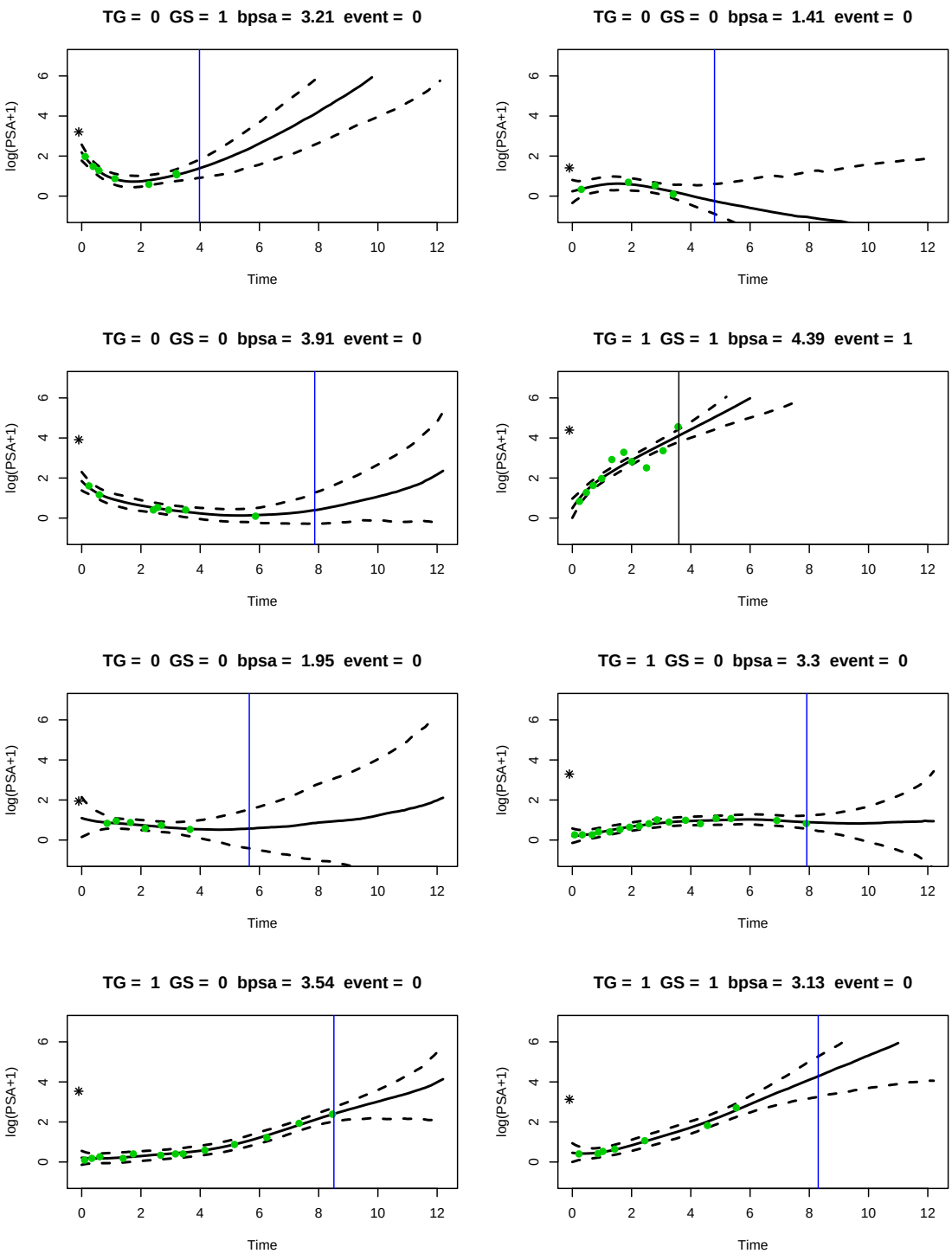

Figure 3.3: Sample patients with observed PSA values and estimated PSA trajectories and their 95% credible intervals

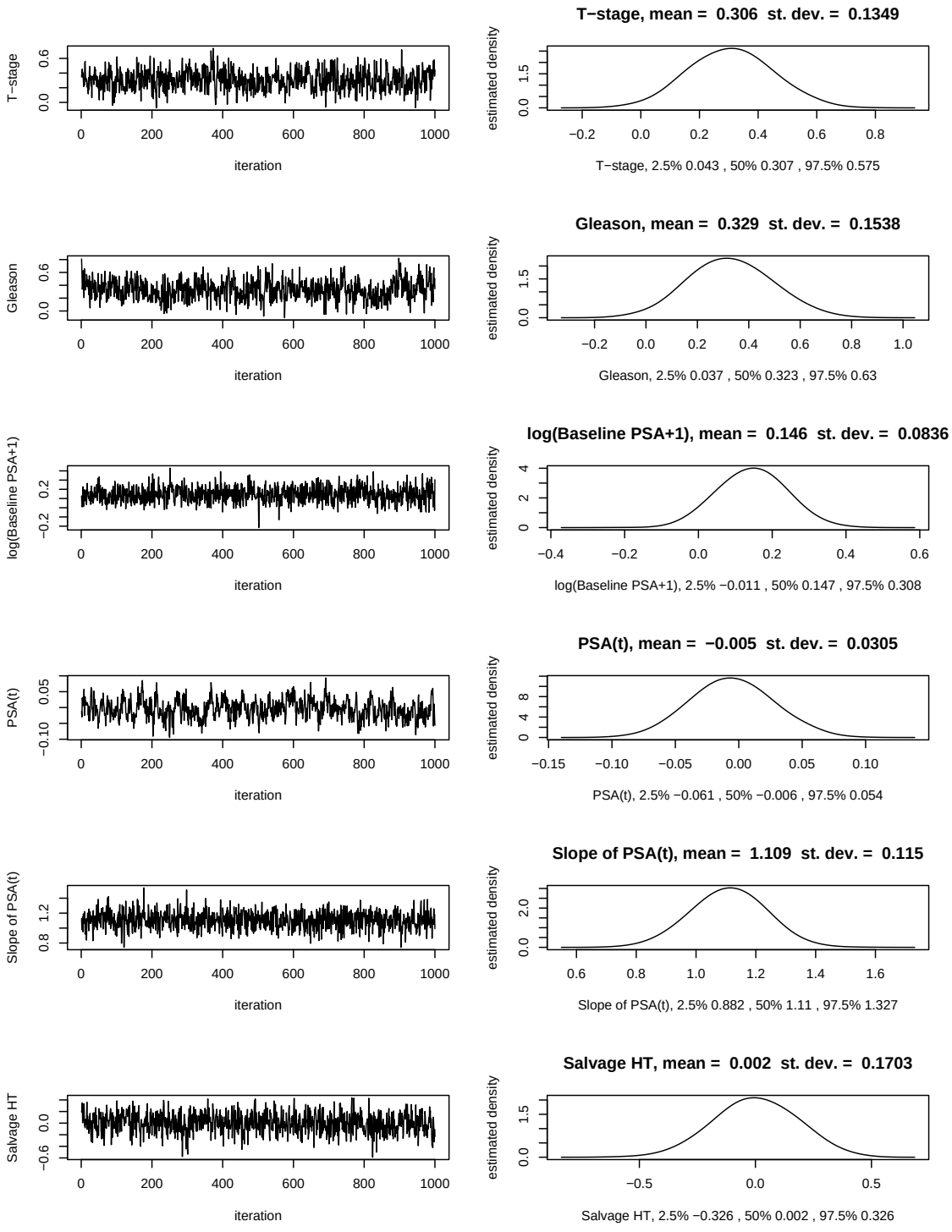

Figure 3.4: Trace plot for the covariate estimates in table 3.3

percentage of observed PSA values that fall within the 95% credible interval over all patients in the test sets. For example, when we censor the test set on year 3, there are altogether 1633 PSA values measured after year 3 from 343 test patients. Among these 1633 observed PSA values, 1568 (96.02%) of them lie within the 95% credible interval. And if we look at the individual-level prediction, on average a patient has about 94.69% of the observed PSA measurements included within the 95% credible interval. We see similar results when we censor the test set on year 5, suggesting a well-constructed 95% credible interval.

To validate the survival model, we calculated the predicted survival probability conditional on information up to the censoring year (3 or 5 year). That is, for patient i,

$$\begin{aligned} P(t > T_j | t > T_i) &= exp(-\int_{T_i}^{T_j} \lambda(u; log(\hat{PSA_i}(u) + 1), \frac{\partial log(\hat{PSA_i}(u) + 1)}{\partial u}, \\ &\quad HT_i(u), X_i, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3, \hat{\phi}_4) du), \end{aligned}$$

where $T_i < T_j <= t$ and $T_i = 3$ or 5 depending on where we censor the test set, and X_i are baseline covariates including T-stage, Gleason and log(baseline PSA+1). Specifically, we have 1000 samples of each of the parameters in this hazard model. We can then construct 1000 estimated survival probability curves for each individual. The median survival probability and its assoociated 95% credible interval are therefore obtained using a sample median and 95% quantiles from these 1000 survival estimates.

Figures 3.7 and 3.8 show the plot of predicted conditional survival probabilities with 95% credible interval for eight sample patients from the two test sets. Table 3.5 shows the predicted number of events obtained from our model versus the expected number of events based on Kaplan-Meier estimates in the two test sets. We can see

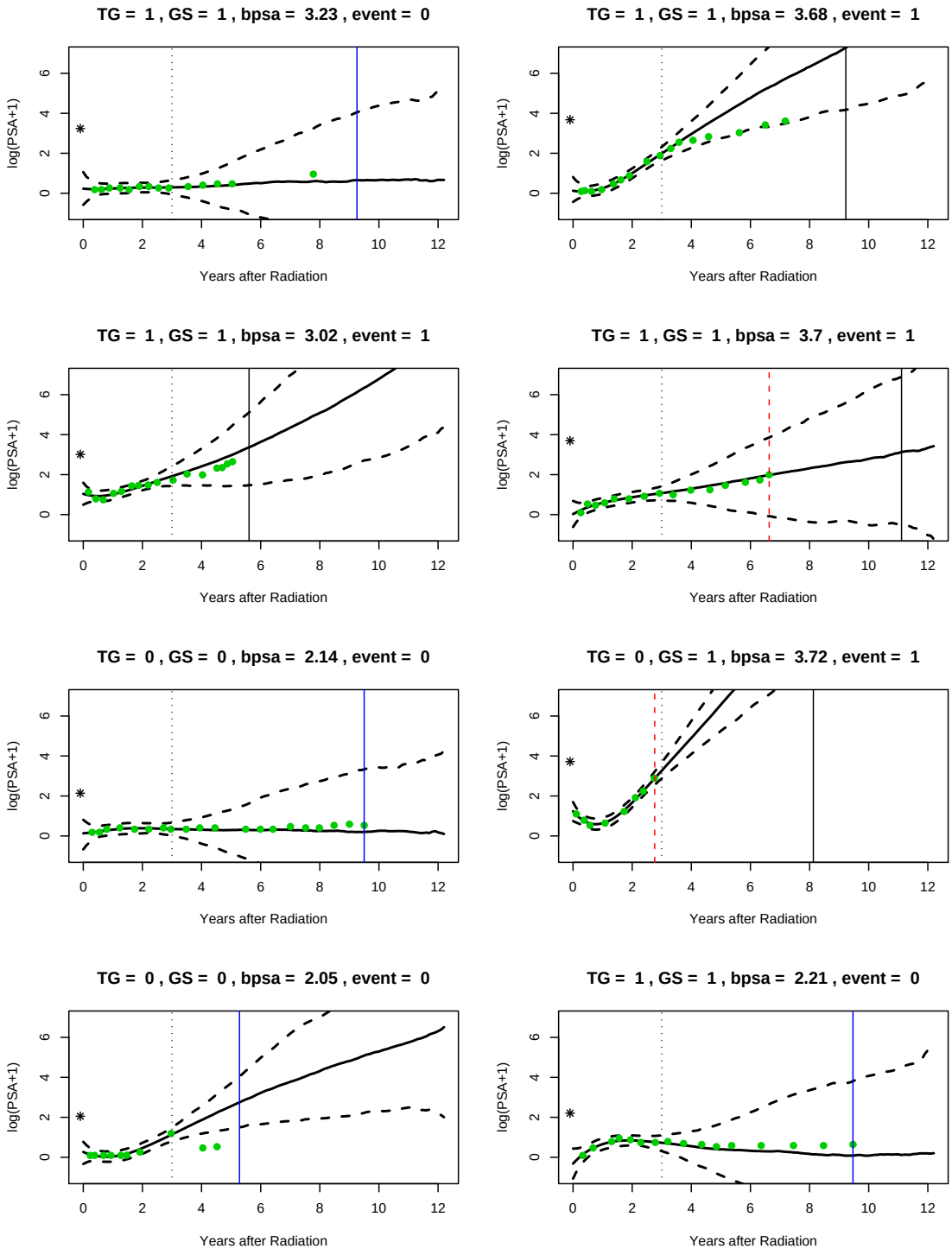

Figure 3.5: Predicted PSA trajectory beyond the censoring year 3

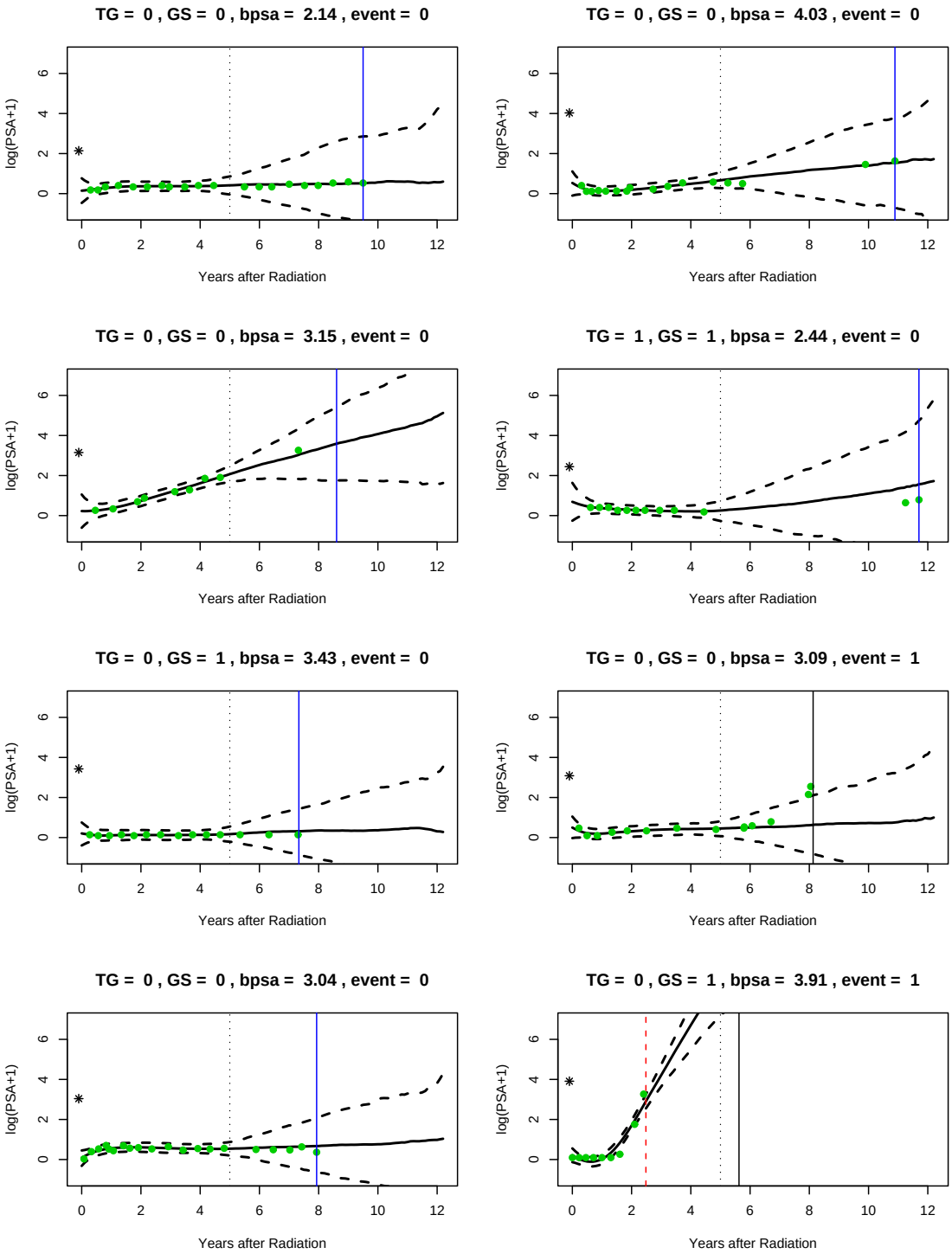

Figure 3.6: Predicted PSA trajectory beyond the censoring year 5

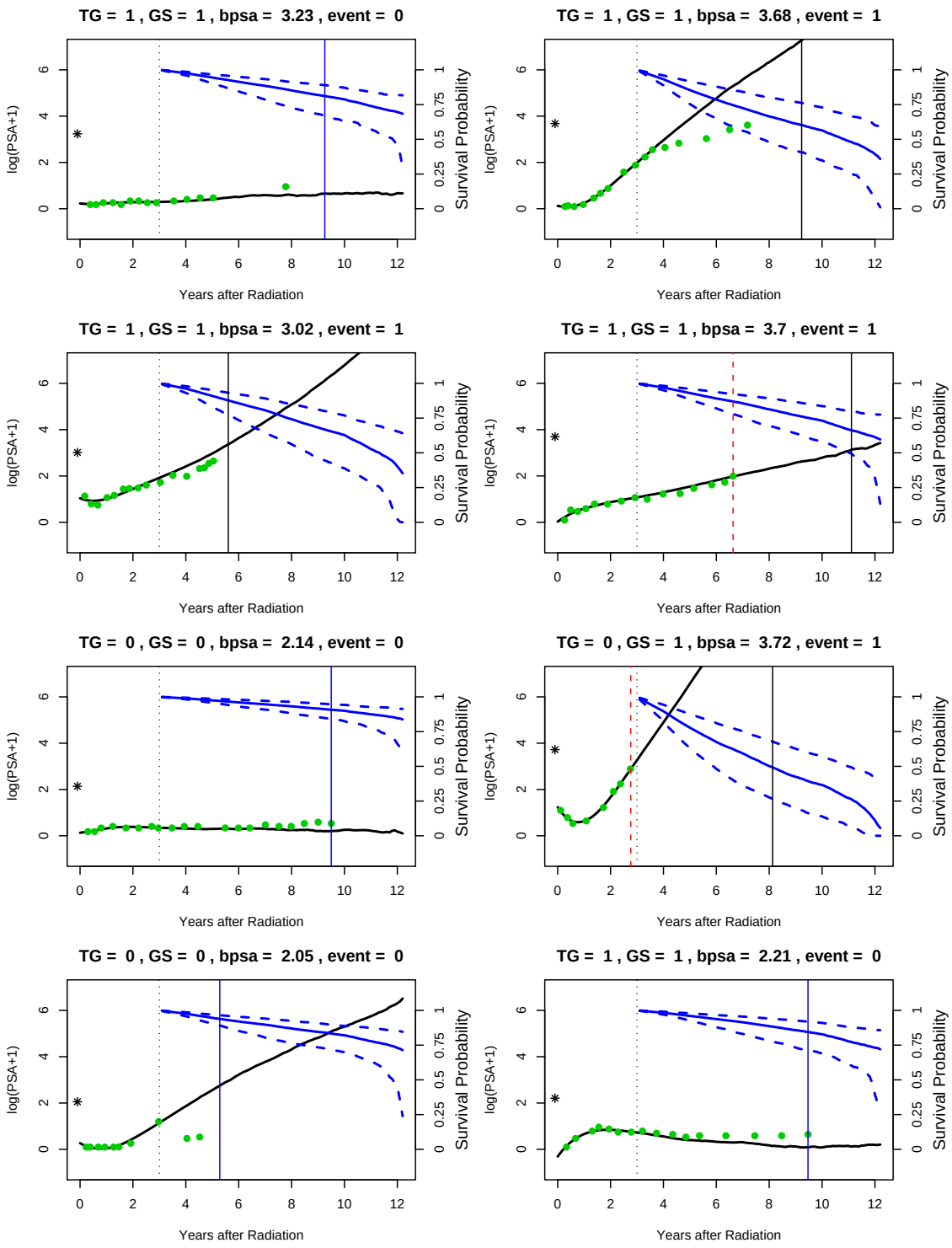

Figure 3.7: Predicted conditional survival probability beyond the censoring year 3

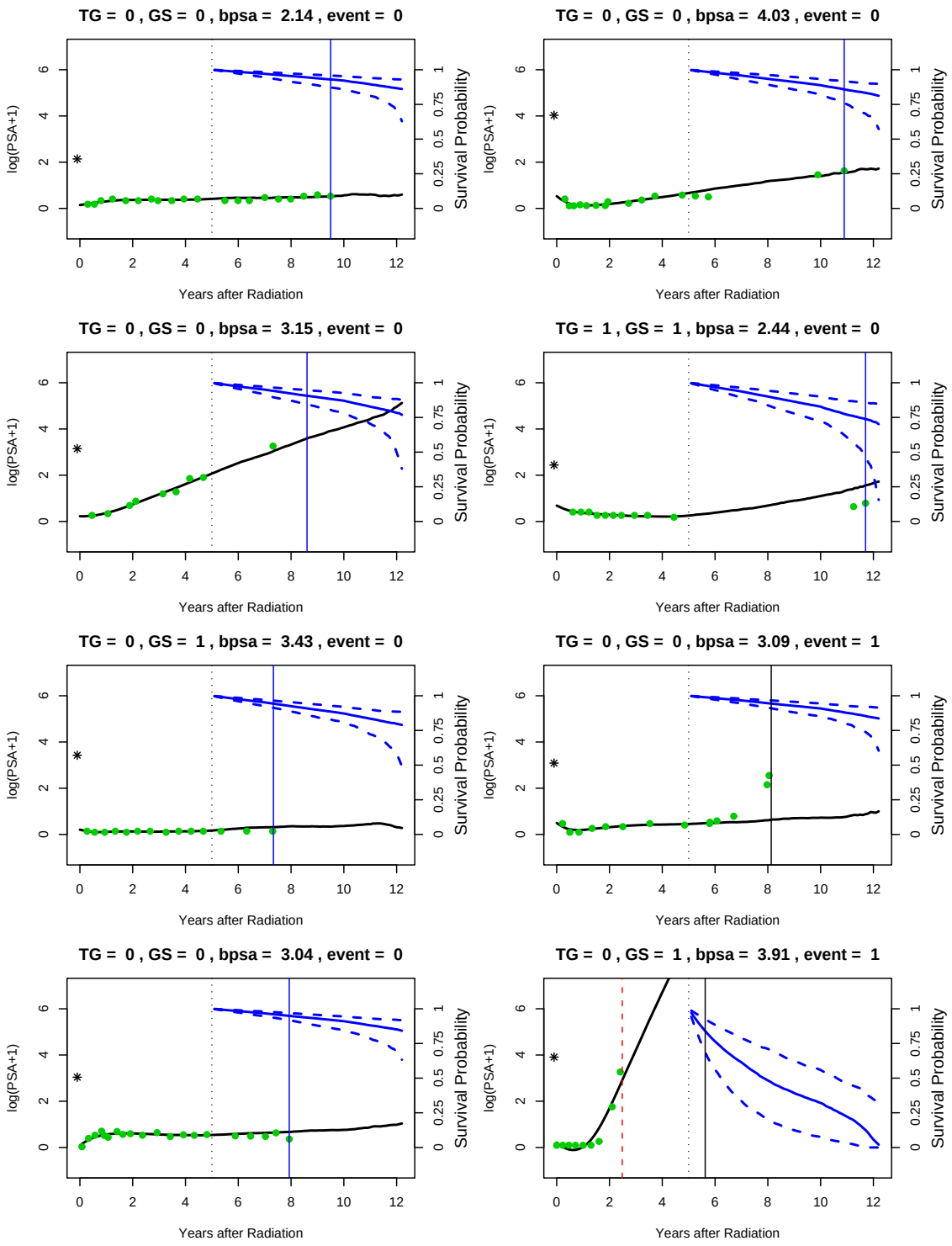

Figure 3.8: Predicted conditional survival probability beyond the censoring year 5

Table 3.5: Number of predicted and observed events in the test sets

	1 year	2 years	3 years
3-year test set			
Predicted number of events	12	26	40
Expected number of events	10	24	42
5-year test set			
Predicted number of events	13	22	34
Expected number of events	13	24	39

that the predicted number events using our model is reasonably close to the expected number of events using Kaplan-Meier estimates.

3.5 Conclusion

In medical studies, we often observe both patients' longitudinal biomarkers and their survival status. The joint modeling of these two data components has been shown to be advantageous compared to two-stage models. In this chapter, we proposed a joint non-parametric model by adapting the penalized B-splines approach proposed by Eiler and Marx to model the longitudinal biomarker process, and linking it with the risk of failure by the Cox proportional hazard model. Compared to parametric assumptions on the longitudinal biomarker process, the non-parametric model does not require the longitudinal trajectories to follow a certain shape or trend, and hence has a much larger degree of flexibility in terms of accommodating the nonlinearity and variation in the longitudinal trajectories. This feature allows our model to be widely applied to various type of medical studies such as cancer and HIV studies.

Brown et. al. (2005) modeled the longitudinal trajectories with cubic B-splines, which has a relatively large number of parameters to estimate for the random effects covariance structure. We used penalized B-splines approach which effectively reduced the number of parameters one has to estimate. The penalized B-splines approach was

based on a number of equally spaced B-splines, and hence there is no issue in terms of determining the location of knots for the B-splines. We developed Bayesian algorithm to estimate the parameters of our joint model, and derived posterior distributions assuming noninformative priors. When the conditional posterior distributions for certain parameters do not have closed form, the Metropolis-Hastings technique was used to estimate these parameters.

We applied our model to prostate cancer data using about three quarters of the data as a training set to build our model and estimate parameters, and about one quarter of the data as a test set to validate our model and evaluate the prediction performance. From this data application, we concluded that the slope of PSA trajectory, T-stage, and Gleason score are all positively associated with the risk of tumor recurrence and death from prostate cancer. Baseline PSA values and salvage hormonal therapy are not significant risk factors. Using the validation data, we demonstrated that the predicted PSA trajectories followed the observed PSA values reasonably closely, the coverage rates of the 95% credible intervals for the predicted PSA trajectories are fairly close to the theoretical value. We also computed the conditional survival probabilities given the information up to the censoring year for the test sets, and the predicted number of events within the first three years after censoring is reasonably close to the observed number of events. However, more simulation studies are needed to better understand our model. In the next chapter, we will present some simulation results to further evaluate the performance of the survival predictions from our model.

CHAPTER IV

Evaluation of Predicted Conditional Survival Estimates Using Absolute Distance Measures

4.1 Introduction

In cancer clinical practice, it is often of interest to predict the survival outcome of a cancer patient. An accurate prediction of disease outcome can help clinicians to tailor therapies for unique patients, to better make difficult clinical decisions, or to select certain patients for clinical trials. Patients can also use this knowledge to improve their quality of life. In recent years, many prognostic models have been proposed to facilitate the prediction of survival outcome. In the previous chapter, we proposed a nonparametric joint model to analyze survival and longitudinal data simultaneously. Specifically, we modeled the longitudinal trajectories with the penalized B-splines, and linked them to the risk of failure using a Cox proportional hazard model. We developed a Bayesian algorithm to estimate the model. In this chapter, we focus on the predictive power of our model by examining the conditional survival probability estimates derived from this joint model.

For any prognostic model, before being applied to any future patients, it is important to evaluate the predictive accuracy of the model. In the previous chapter, we applied the joint model to prostate cancer data where a training set was used to build the model and a test data set was used to evaluate the predictions. We showed

that the predicted number of events within the first three years after censoring was reasonably close to the expected number of events. In this chapter, we conduct a simulation study to evaluate the predicted conditional survival probability estimates using absolute distance measures.

The rest of the chapter is organized as follows. In section 2, we present some background information on the measures developed in the literature to assess the predictive accuracy of a prognostic model. In section 3, we describe the measures that we use in our simulation study to compare three alternative approaches to model the longitudinal and survival data with our joint model from chapter 3. In section 4, we give the details of our simulation study design and results. We also present some results from the data application. In section 5, we close this chapter with some conclusions and discussions.

4.2 Background

In the field of survival data analysis, various measures of predictive accuracy have been proposed to evaluate prognostic models. In this section, we summarize the two main types of measures: the ROC based measure and the proportion-of-explained-variation (R^2) based measure.

4.2.1 The ROC Based Measure

A Receiver Operating Characteristic, or ROC, analysis is a widely used tool in statistics to select possibly optimal models where the outcome variable is typically binary. An ROC curve is a plot of the sensitivity versus (1 - specificity) as the decision threshold is varied. Heagerty and et. al. (2000) applied the ROC curve to the situation where the outcome variable is time-dependent such as survival status. A time-dependent ROC curve based on Kaplan-Meier estimator was proposed to

accomodate the censored survival data. The sensitivity and specificity are therefore defined by

$$P(X > c|D(t) = 1) = \frac{[1 - S(t|X > c)]P(X > c)}{1 - S(t)}$$
$$P(X \leq c|D(t) = 0) = \frac{S(t|X \leq c)P(X \leq c)}{S(t)},$$

where X is baseline covariate, $D(t) = 1$ indicates that the subject had an event before time t, and c is the cut-off point. Here $S(t)$ can be estimated using Kaplan-Meier method.

Heagerty and Zheng (2005) propsed using Cox proportional hazard model to estimate the time-dependent ROC curve, and demonstrated the connection between this ROC method and the concordance measure. Zheng and Heagerty (2007) extended their previous work to the longitudinal biomarker setting, and defined a longitudinal time-dependent ROC curve which can be estimated using a semiparametric model. This method was illustrated using the multicenter AIDS cohort study data.

4.2.2 The Proportion-of-Explained-Variation(R^2) Based Measure

In multiple linear regression analysis, the proportion of explained variation, or R^2, measures the proportion to which the regression model accounts for the variation of a given data set. If the R^2 is large, the regression model is considered to give a good fit of the data, and the significant covariates have a high predictive power. If this R^2 is small, then the regression model is said to be less satisfactory or less powerful. In the context of survival analysis, R^2 based measure has been proposed to assess the predictive accuracy of a Cox proportional hazard model.

Schemper and Henderson (2000) defined the prediction error based on absolute distance measure. Let $D(t)$ be the true underlying survival status of a subject at time t. Clearly, $D(t)$ is a Bernoulli random variable with the marginal probability of

success being $1 - S(t)$ and the conditional probability of success $1 - S(t|X)$, where $S(t)$ is the marginal survival function and $S(t|X)$ is the conditional survival function given baseline covariates X. The marginal prediction error at time t defined by the mean absolute deviation between $S(t)$ and $D(t)$ is

$$\begin{aligned} E(|S(t) - D(t)|) &= [S(t) - 0]Pr(D(t) = 0) + [1 - S(t)]Pr(D(t) = 1) \\ &= S(t)[1 - S(t)] + [1 - S(t)]S(t) \\ &= 2S(t)[1 - S(t)]. \end{aligned}$$

Similarly, the conditional prediction error at time t is given by $2S(t|X)[1 - S(t|X)]$. When interested in a follow-up time period between time 0 to time τ, the prediction error can be defined as

$$PE(\tau) = 2\frac{\int_0^\tau S(t)[1 - S(t)]f(t)dt}{\int_0^\tau f(t)dt}$$
$$PE(\tau; X) = 2\frac{\int_0^\tau E_X S(t|X)[1 - S(t|X)]f(t)dt}{\int_0^\tau f(t)dt},$$

and the relative predictive accuracy of the conditional Cox model versus marginal model is given by

$$R(\tau) = 1 - \frac{PE(\tau; X)}{PE(\tau)}.$$

This measure resembles the R^2 measure in the sense that $R(\tau)$ can be interpreted as the relative gain or the proportion of variation explained by the covariates X.

Henderson, et. al. (2002) extended Schemper and Henderson's work to incorporate the longitudinal biomarker process to Cox proportional hazard model, and focused on the relative gain in predictive accuracy due to the biomarker information. The relative predictive accuracy is therefore defined by

$$R(\tau) = 1 - \frac{PE(\tau; X, Y)}{PE(\tau; X)},$$

where Y is the longitudinal biomarker.

Instead of absolute distance, Graf et. al. (1999) also developed similar R^2 based predictive accuracy measures, but using the quadratic loss function. This measure was shown to be a version of the Brier score, and can be interpreted as the mean square error of prediction. It was applied to standard Cox proportional hazard model without considering longitudinal biomarker information. Gerds and Schumacher (2006) modified the estimator assuming the censoring and event times are conditionally independent given the covariates, and the modified estimator was based on the regression models for the censoring distribution. Schoop et. al. (2008) adapted the quadratic loss function approach to include the longitudinal biomarker as covariates in the Cox proportional hazard model, and re-derived the mean square error of prediction based on the updated information.

So far, there has not been a standard approach to assess the predictive accuracy of a survival model. The ROC based method emphasizes on comparing and discriminating among different diagnostic tests or prognostic models. The R^2 based measure, on the other hand, focuses on evaluating and assessing the predictive accuracy of a prognostic model, and seems to be easier to interpret and is more straightforward to use in model comparison. Proust-Lima and Taylor (2008) compared four different specifications of the R^2-based measure using simulation studies, and concluded that using the absolute loss function in the measure specification yielded systematically smaller bias than the quadratic loss function. In this chapter, our analysis will focus on the R^2 based absolute distance prediction error measure.

4.3 Models and Method

Consider the situation where we have already built a prognostic survival model using a sample of patients. A new patient comes in. We follow this patient until time s when the patient asks about the chance of survival at time $t > s$. If we can assume this patient comes from the same population as the sample of patients we used to build our prognostic model, then we can make prediction on this patient's conditional survival probability using our prognostic model given his or her follow-up information up to time s. The aim of this chapter is to derive this predicted conditional survival probability and assess its accuracy.

In the literature, the R^2-type measures were proposed with the aim to validate the prognostic model by quantifying the explained variation due to the prognostic model, and hence naturally the same sample subjects that were used to build the prognostic model will be used again in estimating the R^2-measures. On the other hand, what we are interested in is a more realistic situation for clinical practice. We are essentially assuming a split sample design where one sample is used to build the prognostic model and the other sample is used to validate the model and make predictions. We do not have any additional information about the new patients after their censoring time s, and we want to evaluate how accurate our predicted conditional survival probabilities are with the existing prognostic tool.

4.3.1 Predictive Accuracy Measures

Let $S_i(t; s, X_i, Y_i(s), \theta, \delta_i)$ be the true conditional survival function at time $t > s$ given the information up to a certain censoring time s for subject $i, i = 1, ..., n$, where θ is the vector of the true fixed effects parameter values, δ_i is the vector of the true random effects parameter values, X_i is the baseline information, $Y_i(s)$

is the time-dependent biomarker. Let $\hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)$ be the corresponding estimated conditional survival function for subject i, where $\hat{\theta}$ is the vector of the estimated fixed effects parameter values, and $\hat{\delta}_i$ is the vector of the estimated random effects parameter values. We define the prediction error to be

$$\begin{aligned} err(t; s, X, Y(s)) &= \frac{1}{n} \sum_{i=1:n} err_i(t; s, X_i, Y_i(s)) \\ &= \frac{1}{n} \sum_{i=1:n} |S_i(t; s, X_i, Y_i(s), \theta, \delta_i) - \hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)| \\ &= \frac{1}{n} \sum_{i=1:n} |\frac{Pr(u > t|X_i, Y_i(s), \theta, \delta_i)}{Pr(u > s|X_i, Y_i(s), \theta, \delta_i)} - \frac{Pr(u > t|X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)}{Pr(u > s|X_i, Y_(s), \hat{\theta}, \hat{\delta}_i)}|. \end{aligned} \tag{4.1}$$

This prediction error captures the absolute distance between the predicted conditional survival probabilities and the true conditional survival probabilities at a certain time t averaged over all the test subjects. Obviously, at any given censoring time s, this prediction error is a process in time $t \in (s, T)$ where T is the maximum follow-up time of the study. Hence, we can plot the curve of the prediction error against time t, and the model that produces the least prediction errors (the lower prediction error curve) will be preferred to the other models.

Sometimes, one model might yield a smaller prediction error when time t is close to censoring time s, and larger prediction error when time t is farther away from time s. Another model might behave oppositely. It can be misleading to compare these two models solely based on the prediction error at one certain time t. In this case, we can summarize the prediction errors by defining a time-averaged prediction error. That is,

$$AE(t; s, X, Y(s)) = \frac{1}{n} \sum_{i=1:n} \frac{1}{t-s} \int_s^t err_i(u; s, X_i, Y_i(s)) du. \tag{4.2}$$

This time-averaged prediction error calculates the average prediction error between time s and time t conditional on information up to the censoring time s. Depending

on the different study purpose, we can vary the prediction duration $t-s$ to assess the predictive accuracy either in a short time frame (by choosing $t-s$ to be small) or in a long time frame (up to the maximum follow-up time $t=T$). Alternatively, given the same prediction duration $t-s$, the predictive accuracy may differ depending on the time in the follow-up period that we perform the predictive accuracy assessment at (varying s). If the risk of failure is dependent on the accumulated information during time interval $(0,s)$, then we would expect a different predictive accuracy at a different censoring time s. Therefore, we can also vary the censoring time s to evaluate the predictive accuracy during the entire follow-up time period. The time-averaged prediction error $AE(t;s)$ can be plotted in a three-dimensional space for a graphical comparison among different models, and again the smaller average prediction error (lower curve) gives the preferrable model.

Based on the concept of proportion of explained variation, we can define the relative gain of predictive accuracy due to incorporating the biomarker information to be

$$R(t;s) = 1 - \frac{AE(t;s,X,Y(s))}{AE(t;s,X)}, \tag{4.3}$$

where $AE(t;s,X))$ is the time-averaged prediction error assuming the baseline survival model. This relative measure quantifies the improvement in predictive accuracy comparing the survival model with the time-dependent biomarker information as a predictor and the baseline survival model where the time-dependent biomarker information is ignored.

The previous measures are based on the assumption that the true conditional survival distribution is known. When the true survival distribution is unknown, we adapt the prediction error measure proposed by Schemper and Henderson (2000). We want to characterize the distance between the observed survival status $d_i(t)$ and the

predicted conditional survival estimate $\hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)$. The more accurate the prediction is, the smaller distance there is between $d_i(t)$ and $\hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)$. Therefore, for this subject i, the estimated prediction error based on the expected absolute distance measure at time t given all the information up to time s is

$$\begin{aligned}\hat{PE}_i(t; s, X_i, Y_i(s)) &= [1 - (1 - \hat{S}_i(t; s, X_i, , Y_i(s), \hat{\theta}, \hat{\delta}_i))](1 - \hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)) \\ &\quad + [(1 - \hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)) - 0]\hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i) \\ &= 2\hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)(1 - \hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)).\end{aligned}$$

This estimated prediction error is essentially assuming that a subject who is censored at time s will have the same risk as those with a known survival status at time s and the same baseline characteristics. This subject is then classified as either dead or alive by extrapolation based on the corresponding conditional probabilities. Using the test data which have n subjects, the mean estimated prediction error becomes

$$\begin{aligned}\hat{PE}(t; s, X, Y(s)) &= \frac{1}{n}\sum_{i=1:n} \hat{PE}_i(t; s, X_i, Y_i(s)) \\ &= \frac{2}{n}\sum_{i=1:n} \hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)(1 - \hat{S}_i(t; s, X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)) \\ &= \frac{2}{n}\sum_{i=1:n} \frac{Pr_i(u > t|X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)}{Pr_i(u > s|X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)}(1 - \frac{Pr_i(u > t|X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)}{Pr_i(u > s|X_i, Y_i(s), \hat{\theta}, \hat{\delta}_i)}). \qquad (4.4)\end{aligned}$$

In the similar way as equation (4.2), the time-averaged estimated prediction error is given by

$$A\hat{PE}(t; s, X, Y(s)) = \frac{1}{n}\sum_{i=1:n} \frac{1}{t-s}\int_s^t \hat{PE}_i(u; s, X_i, Y_i(s))du. \qquad (4.5)$$

In the discrete-time world, this time-averaged prediction error can be estimated by

$$A\hat{PE}(t; s, X, Y(s)) = \frac{1}{n}\sum_{i=1:n} \frac{1}{J}\sum_{s<t_j<=t} \hat{PE}_i(t_j; s, X_i, Y_i(s)), \qquad (4.6)$$

where $s < t_j \leq t$ are the discrete times where the conditional survival probabilites are estimated. When the interest is in the long-term predictive accuracy, one might

want to take into account the decreasing number of events due to censoring as time passes. A weighted average prediction error can be obtained by applying weights to the estimated prediction errors at different times. That is,

$$W\hat{P}E(t; s, X, Y(s)) = \frac{\sum_{s<t_j<=t} d_{t_j} \hat{G}(t_j)^{-1} \hat{PE}(t_j; s, X, Y(s))}{\sum_{j=1:J} d_{t_j} \hat{G}(t_j)^{-1}}, \tag{4.7}$$

where $s < t_j \leq t : j = 1, ..., J$ are the death times along the prediction region, d_{t_j} are the number of deaths at time t_j, and $\hat{G(t_j)}$ is the Kaplan-Meier estimates of the censoring distribution at time t_j. d_{t_j} and $\hat{G(t_j)}$ can be obtained using the sample that we build the prognostic model with. These weights can compensate the decrease in observed number of deaths as time passes. Here the standard assumption of noninformative censoring applies.

Following the concept of proportion of variation explained, we can define the relative predictive accuracy comparing the survival model with the time-dependent biomarker information as a predictor and the baseline survival model where the time-dependent biomarker information is ignored as

$$\hat{R}(t; s) = 1 - \frac{W\hat{P}E(t; s, X, Y(s))}{W\hat{P}E(t; s, X)}, \tag{4.8}$$

where $W\hat{P}E(t; s, X))$ is the weighted average prediction error assuming the baseline survival model. This relative measure quantifies the improvement in predictive accuracy due to incorporating the biomarker information in the survival model.

4.3.2 Models For Comparison

In this chapter, we conduct a simulation study to to assess the predictive accuracy of the joint nonparametric model that we proposed in chapter 3 using the measures we described in section 4.3.1. In chapter 3, we developed a likelihood based joint model of longitudinal and survival data based on penalized B-splines, wherein the

true longitudinal trajectory was linked to the risk of failure by the Cox model with time-varying covariates. We can then derive the true conditional survival probability as

$$S_i(t;s) = exp[-\int_s^t \lambda(u; Z_i(u), X_i, \theta, \delta_i)du],$$

and the predicted conditional survival probability as

$$\hat{S}_i(t;s) = exp[-\int_s^t \lambda(u; \hat{Z}_i(u), X_i, \hat{\theta}, \hat{\delta}_i)du],$$

where $Z_i(u)$ and $\hat{Z}_i(u)$ are the true and predicted longitudinal trajectory for subject i, X_i is the baseline covariates, θ and $\hat{\theta}$ are the true and estimated fixed effects parameters, and δ_i and $\hat{\delta}_i$ are the true and estimated random effects parameters.

For comparison purposes, we consider three alternative approaches. The first approach is to ignore the longitudinal biomarker information and assume the risk of failure only depends on the baseline information. This is a standard Cox proportional hazard model with only baseline information as risk factors, which we call the "naive" approach. The second approach is to incorporate the extra information from the longitudinal biomarkers by assuming that the risk of failure depends both on the baseline information and the last observed biomarker value before the censoring time. This essentially predicts that the biomarker trajectory will stay flat after the last observation time, and future biomarker values will not change with time. This approach is termed as "last value carry forward", or "LVCF" approach. The third approach is a two-stage model in which a mixed effects model is assumed for the longitudinal biomarkers and then the imputed biomarker values at all event times are used as a time dependent covariate in the Cox proportional hazard model.

More formally, we considered the following four methods to obtain the conditional

survival probability. The naive approach:

$$\lambda(t) = \lambda_0(t)exp(\eta X).$$

The last value carry forward approach:

$$\lambda(t) = \lambda_0(t)exp(\alpha Y_{obs}(t) + \eta X).$$

The two-stage model:

$$\lambda(t) = \lambda_0(t)exp(\alpha \hat{Z}_{TS}(t) + \eta X).$$

The conditional survival estimates from the joint model in Chapter 3:

$$\lambda(t) = \lambda_0(t)exp(\alpha \hat{Z}_{JO}(t) + \eta X).$$

Here, X is the baseline covariate, $Y_{obs}(t)$ is the last observed biomarker value up to time t, $\hat{Z}_{TS}(t)$ is the imputed biomarker value using the two-stage model, and finally $\hat{Z}_{JO}(t)$ is the estimated longitudinal biomarket value based on the joint model developed in Chapter 3. Also, $\lambda(t)$ is the hazard function and $\lambda_0(t)$ is the baseline hazard function.

In our simulation studies, we know the true survival curves that the simulations are based on for each subject . Let the true survival function for subject i be $S_i(t; X_i, Z_i(t), \theta, \delta_i)$. To assess the predictive accuracy of approach D compared to the other three alternatives, we can calculate the prediction error $err(t; s)$ using these true parameter values and the estimated parameter values. Equation (4.1) therefore becomes

$$err(t; s) = \frac{1}{n} \sum_{i=1:n} |\frac{S_i(t; X_i, Z_i(s), \theta, \delta_i)}{S_i(s; X_i, Z_i(s), \theta, \delta_i)} - \frac{\hat{S}_i(t; X_i, \hat{\theta})}{\hat{S}_i(s; X_i, \hat{\theta})}|$$

for the naive approach,

$$err(t; s) = \frac{1}{n} \sum_{i=1:n} |\frac{S_i(t; X_i, Z_i(s), \theta, \delta_i)}{S_i(s; X_i, Z_i(s), \theta, \delta_i)} - \frac{\hat{S}_i(t; X_i, Y_{obs,i}(s), \hat{\theta})}{\hat{S}_i(s; X_i, Y_{obs,i}(s), \hat{\theta})}|$$

for the last value carry forward approach,

$$err(t;s) = \frac{1}{n}\sum_{i=1:n} |\frac{S_i(t;X_i,Z_i(s),\theta,\delta_i)}{S_i(s;X_i,Z_i(s),\theta,\delta_i)} - \frac{\hat{S}_i(t;X_i,\hat{Z}_{TS,i}(s),\hat{\theta},\hat{\delta}_i)}{\hat{S}_i(s;X_i,\hat{Z}_{TS,i}(s),\hat{\theta},\hat{\delta}_i)}|$$

for the two stage model, and

$$err(t;s) = \frac{1}{n}\sum_{i=1:n} |\frac{S_i(t;X_i,Z_i(s),\theta,\delta_i)}{S_i(s;X_i,Z_i(s),\theta,\delta_i)} - \frac{\hat{S}_i(t;X_i,\hat{Z}_{JO,i}(s),\hat{\theta},\hat{\delta}_i)}{\hat{S}_i(s;X_i,\hat{Z}_{JO,i}(s),\hat{\theta},\hat{\delta}_i)}|$$

for the joint model. The time-averaged prediction error (AE) using equation 4.2 can then be subsequently obtained. We use the naive approach as the benchmark model in our comparisons. That is, all the other three models will be compared to the naive approach, and $R(t;s)$ in equation 4.3 will be be calculated to evaluate the relative gain in terms of explained variation due to incorporating the longitudinal biomarker $Y(s)$ into the survival model.

Assuming we do not know the true survival function of our simulation data, we can also use equation 4.6 and 4.7 to obtain the estimated average prediction error ($\hat{APE}(t;s)$) and the estimated weighted average prediction error ($\hat{WPE}(t;s)$). Again, using the naive approach as the reference model, we can compute $\hat{R}(t;s)$ and compare these four approaches in terms of their increased power in explaining the variation in the survival data through biomarker information.

4.4 Simulation Studies and Data Application

4.4.1 Design of the Simulation Study

We carry out a simulation study to evaluate the predictive accuracy of the proposed model in section 3.3 in comparison with the three other alternatives described in section 4.3.2. Similar to the strategy in chapter 3 for model validation, we adopt a split sample design where training data are generated to build the model and test data are used to validate the model. Each training data set is associated with six

test data sets corresponding to six different censoring times, and the subjects in the same test data set have the same censoring time.

For simplicity, we assume both the longitudinal biomarker and survival probability depend on the same baseline covariate. That is, the longitudinal biomarker is simulated based on the following model:

$$\begin{aligned} Y_i(t) &= Z_i(t) + e_i(t), \\ Z_i(t) &= -0.4 + 0.35X_i + 0.2t + 0.45X_it - 0.15t^2 + 0.02t^3 + \beta_{1i} + \beta_{2i}t + \beta_{3i}t^2 + \beta_{4i}t^3, \end{aligned}$$

where $X_i \sim Bernoulli(\frac{1}{2})$ for subject i. The random effects parameters are distributed as $\beta_{1i} \sim N(0, 0.3)$, $\beta_{2i} \sim N(0, 0.01)$, $\beta_{3i} \sim N(0, 0.01)$, and $\beta_{4i} \sim N(0, 0.0001)$. The measurement errors are assumed to be $e_i(t) \sim N(0, 0.1)$. The maximum follow-up time is chosen to be 10.

For the survival data, we assume:

$$\lambda_i(t) = \lambda_0(t)exp(Z_i(t) + 2X_i),$$

where the baseline hazard is a constant over the entire follow-up period, that is, $\lambda_0(t) = 0.01$. For the training data set, censoring is included using a uniform distribution over $(0, 10)$, and this yields about 50% censoring rate on average. For the six corresponding test data sets, we censor the survival times at times 1, 2, 3, 4, 5, and 6. That is, the subjects in the first 200 test sets are all censored at time 1; the subjects in the second 200 test sets are all censored at time 2; and so on. So no additional censoring needs to be introduced to the test data.

In our simulation, two hundred training data sets are generated, and each training data set has 300 subjects. Correspondingly, 1200 data sets are generated as test data sets, and each test data set has 100 subjects. We sample $Y(t)$ at 25 random time points along the longitudinal trajectories as our observed longitudinal measurements.

These 25 $Y(t)$ samples are further censored by the event or censoring times obtained from the survival model. We force each subject to have at least one longitudinal measurement, and no more than 25 longitudinal measurements.

Figure 4.1 shows some sample longitudinal trajectories. We can see a very large shape variation in the longitudinal profiles. All the longitudinal trajectories display a general upward slope trend. But some subjects have their longitudinal biomarkers dropping low in the first few years and then picking back up high, whereas some other subjects show an upward trend right away. Figure 4.2 shows the Kaplan-Meier survival plot from one random training data set.

4.4.2 Fitting the Simulated Data

We use the four approaches described in Section 4.3.2 to fit the training data sets, and make predictions on the test data sets by calculating the corresponding predicted conditional survival probabilities. Let superscript D represent training data, and superscript C represent test data. Specifically, let X_i^D be the baseline information of subject i from the training data, and $X_{i'}^C$ be the baseline information of subject i' from the test data. For the naive approach, we run SAS PROC PHREG with baseline information X_i^D as covariates on every training dataset, and obtain the cumulative baseline hazard estimates $\hat{\Lambda}_0(t)$ and the parameter estimate $\hat{\eta}$. We assume the baseline hazard $\lambda_0(t)$ to be a step function, and is estimated by the difference of the cumulative baseline hazard estimates of two consecutive event times. We then can calculate the predicted conditional survival probabilities using the test datasets conditional on each censoring time. That is,

$$\hat{S}_{i'}(t;s) = exp(-\int_s^t \hat{\lambda}_0(u)exp(\hat{\eta}X_{i'}^C)du),$$

where the censoring time s takes integer values from 1 to 6, and $t > s$.

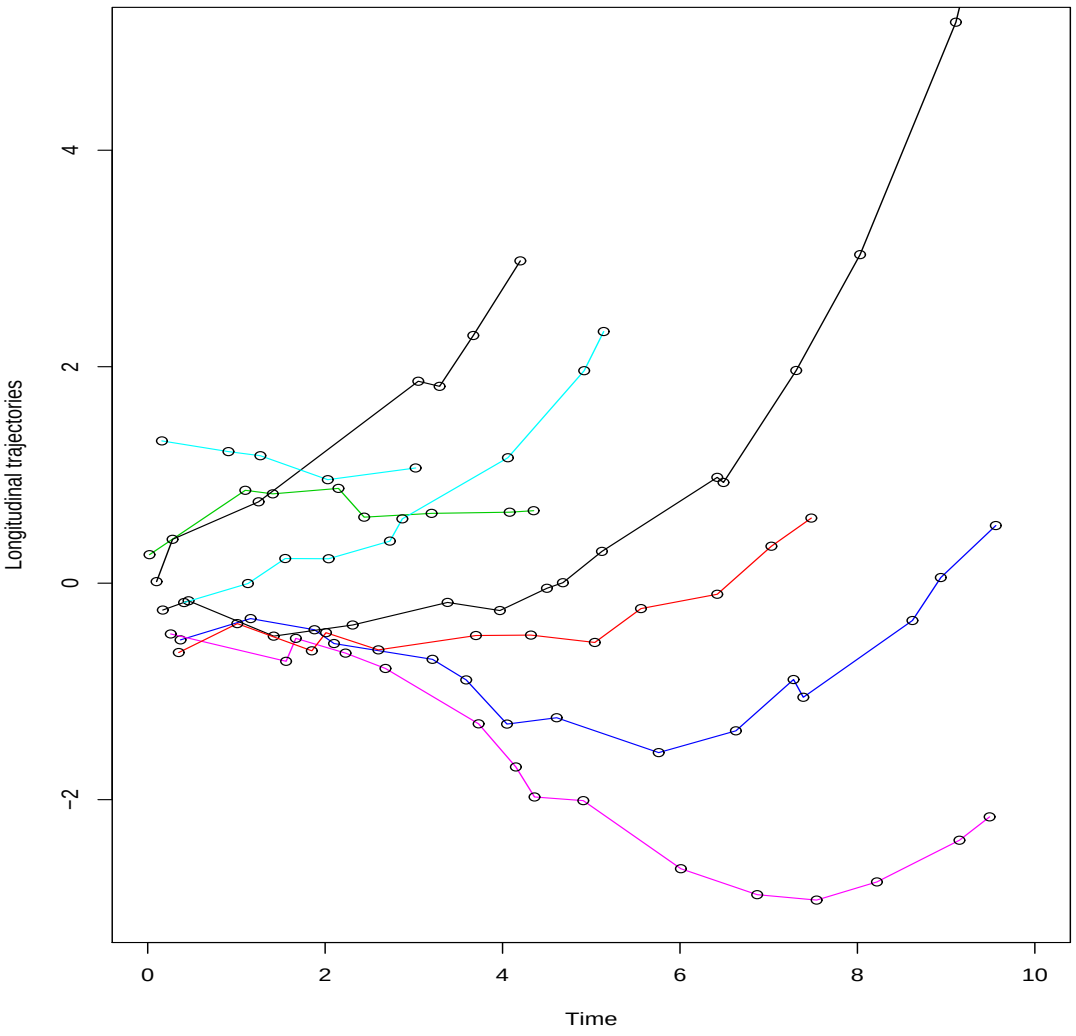

Figure 4.1: Sample longitudinal trajectories

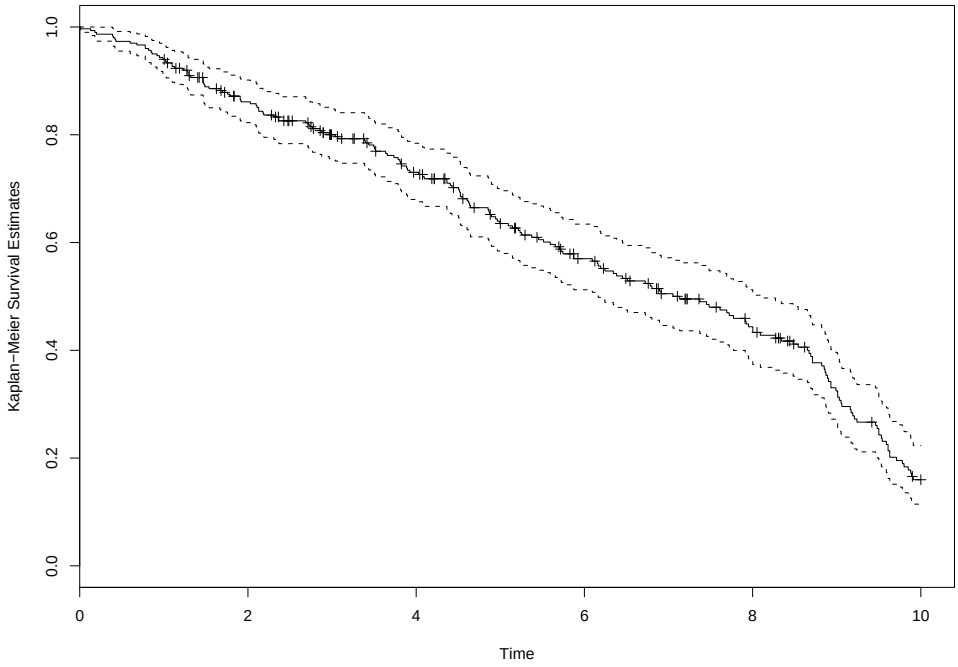

Figure 4.2: Sample Kaplan-Meier survival plot

For the "last value carry forward" approach, we again run SAS PROC PHREG on every training set with both the last observed longitudinal values $Y_i^D(t)$ and baseline information X_i^C as covariates, and obtain the cumulative baseline hazard estimates $\hat{\Lambda}_0(t)$ and the parameter estimates $\hat{\alpha}$ and $\hat{\eta}$. The baseline hazard function can be estimated in the same way as before. Using these parameter estimates, the predicted conditional survival probability becomes

$$\hat{S}_{i'}(t;s) = exp(-\int_s^t \hat{\lambda}_0(u) exp[\hat{\alpha} Y_{i'}^C(s) + \hat{\eta} X_{i'}^C)]du,$$

where $Y_{i'}^C(s)$ is the last observed longitudinal measurement before time s for subject i'. Obviously, $Y_{i'}^C(s)$ does not change with time, and this model is again a standard Cox proportional hazard model.

For the two-stage approach, we first run SAS PROC MIXED on the training longitudinal data based on model 3.7. Let F be the fixed effects design matrix, and

R be the random effects design matrix. In terms of model 3.7, we choose the same B-spline basis for $B_1(t)$ and $B_2(t)$, that is, $B_1(t) = B_2(t) = B(t)$. Thus, for the training subjects, $F^D(t) = B(t)\Phi^{UN}$, and $R^D(t) = B(t)(\Phi^P + \Psi^{UN} + \Psi^P)$. Then the mixed model can be written as

$$\begin{aligned} Y_i^D(t) &= Z_i^D(t) + e_i \\ &= F_i^D(t)\gamma + R^D(t)\theta_i^D + e_i, \end{aligned}$$

where γ is the fixed effects parameter corresponding to the γ in (3.7), and θ_i^D are the random effects parameters from the training data including β_P, δ_i^{UN}, and δ_i^P. Let G be the random effects covariance matrix, and E be the residual covariance matrix. The mixed model on training data produce estimates of $\hat{G}$, $\hat{E}$, $\hat{\alpha}$, and $\hat{\theta_i^D}$. We plug these parameter estimates back into the $Y_i^D(t)$ model to impute the missing longitudinal biomarkers at each event times. That is, $\hat{Z}_i^D(t) = F_i^D(t)\hat{\gamma} + R^D(t)\hat{\theta}_i^D$. We then implement the time-dependent Cox proportional hazard model with the imputed longitudinal biomarkers $\hat{Z}_i^D(t)$ as the time-dependent covariate, and X_i^D as baseline covariate. This yields the survival parameter estimates of $\hat{\alpha}$ and $\hat{\eta}$. To make survival predictions on the test data, we need to impute the longitudinal profile for the test subjects too. Let $F^C(t)$ and $R^C(t)$ be the fixed effects and random effects design matrix for the test subjects. The fixed effects parameter estimates $\hat{\gamma}$ has been obtained from the mixed model fit. The random effects parameters of the test subjects can be estimated using $\hat{\theta}^C = \hat{G}R^C(t)'\hat{V}^{-1}(Y^C(t) - F^C(t)\hat{\gamma})$, where $\hat{V} = R^C(t)\hat{G}R^C(t)' + \hat{E}$. Then, we can plug these parameter estimates back to the model to obtain the imputed longitudinal profile as $\hat{Z}_{i'}^C(t) = F_{i'}^C(t)\hat{\gamma} + R^C(t)\hat{\theta}_{i'}^C$. Thus, the predicted conditional survival probabities can be obtained by

$$\hat{S}_{i'}(t;s) = exp(-\int_s^t \hat{\lambda}_0(u)exp(\hat{\alpha}\hat{Z}_{i'}^C(u) + \hat{\eta}X_{i'}^C)du).$$

Lastly, our Bayesian joint model as described in chapter 3 is used to fit the training data. Again, we assumed the same $B(t)$ in model 3.7, and we choose six equally-spaced knots (between 0 and 10) to construct $B(t)$. Similar to chapter 3, we apply noninformative priors to the parameters and derive the posterior distributions accordingly. MCMC chains are generated using the Metropolis-Hastings algorithm, and 1000 samples of the parameter estimates are kept after running 20000 iterations and taking the first 10000 as burn-in period. To make predictions on the test subjects, we need first to predict their longitudinal trajectories, which requires us to estimate their random effects parameters. We run another short MCMC chain using the test data and the 1000 samples of fixed effects parameters including $(\hat{\gamma}, \hat{\beta^P}, \hat{\sigma}^2_{\beta^P}, \hat{\alpha}, \hat{\eta})$ obtained from fitting the training data set. Conditional on each sample of the fixed effects parameters, we run a 50 iteration MCMC chain of the random effects parameters based on the conditional posterior distribution derived from our joint model, and we keep the last iteration from this small MCMC chain. Thus we end up with 1000 samples of the fixed effects parameters obtained using the training data and 1000 samples of the random effects parameters estimated using the test data. We then can predict the longitudinal trajectories by

$$\hat{Z}^C_{i',JO}(t) = B(t)\Phi^{UN}X^C_{i'}\hat{\gamma} + B(t)\Phi^P\hat{\beta^P} + B(t)\Psi^{UN}\hat{\beta}^{UN}_{i'} + B(t)\Psi^P\hat{\beta}^P_{i'},$$

and the conditional survival probabilities can be predicted by

$$\hat{S}_{i'}(t;s) = exp(-\int_s^t \hat{\lambda}_0(u)exp[\hat{\alpha}\hat{Z}^C_{i',JO}(u) + \hat{\eta}X^C_{i'})]du.$$

4.4.3 Simulation Results

We use the above strategy to calculate the predicted survival probabilities conditional on the information up to the censoring time for each subject in each of the

Table 4.1: True versus predicted number of events

Censoring Time	Model	Prediction Time								
		2	3	4	5	6	7	8	9	10
1	Truth	7	14	22	31	39	49	51	61	77
	Naive	11	20	30	35	44	55	58	67	84
	LVCF	10	18	25	35	43	50	57	66	82
	Two-stage	9	15	20	30	36	46	55	65	80
	Joint	8	13	20	30	41	51	54	65	73
2	Truth		7	16	25	33	40	47	57	75
	Naive		10	19	29	38	48	55	64	85
	LVCF		10	18	29	38	46	54	63	82
	Two-stage		8	14	23	36	44	42	52	80
	Joint		9	14	27	35	44	51	61	72
3	Truth			10	20	29	36	44	55	73
	Naive			14	25	33	44	50	61	79
	LVCF			14	23	27	41	51	60	79
	Two-stage			9	22	30	32	46	50	74
	Joint			9	18	29	38	44	52	72
4	Truth				12	22	30	38	50	70
	Naive				14	25	34	44	45	77
	LVCF				14	24	36	41	59	76
	Two-stage				12	24	32	35	45	71
	Joint				11	21	32	36	47	70
5	Truth					11	20	30	43	67
	Naive					15	24	35	50	61
	LVCF					13	24	31	38	73
	Two-stage					12	23	26	46	71
	Joint					11	21	32	46	64
6	Truth						10	21	36	62
	Naive						15	29	30	52
	LVCF						14	25	42	70
	Two-stage						12	23	31	66
	Joint						10	22	34	60

1200 simulated test sets. Table 4.1 compares the predicted number of events (averaged over 200 simulated test data sets) from these four approaches versus the true number of events in the simulated test data. We can see that the predicted number of events using the joint model is closer to the true number of events than using the other three methods.

To compare the predictive accuracy of these four approaches, we calculate the time-averaged predition errors (AE) and the estimated weighted average prediction errors ($W\hat{P}E$)for each test data set at the six different censoring times. We then

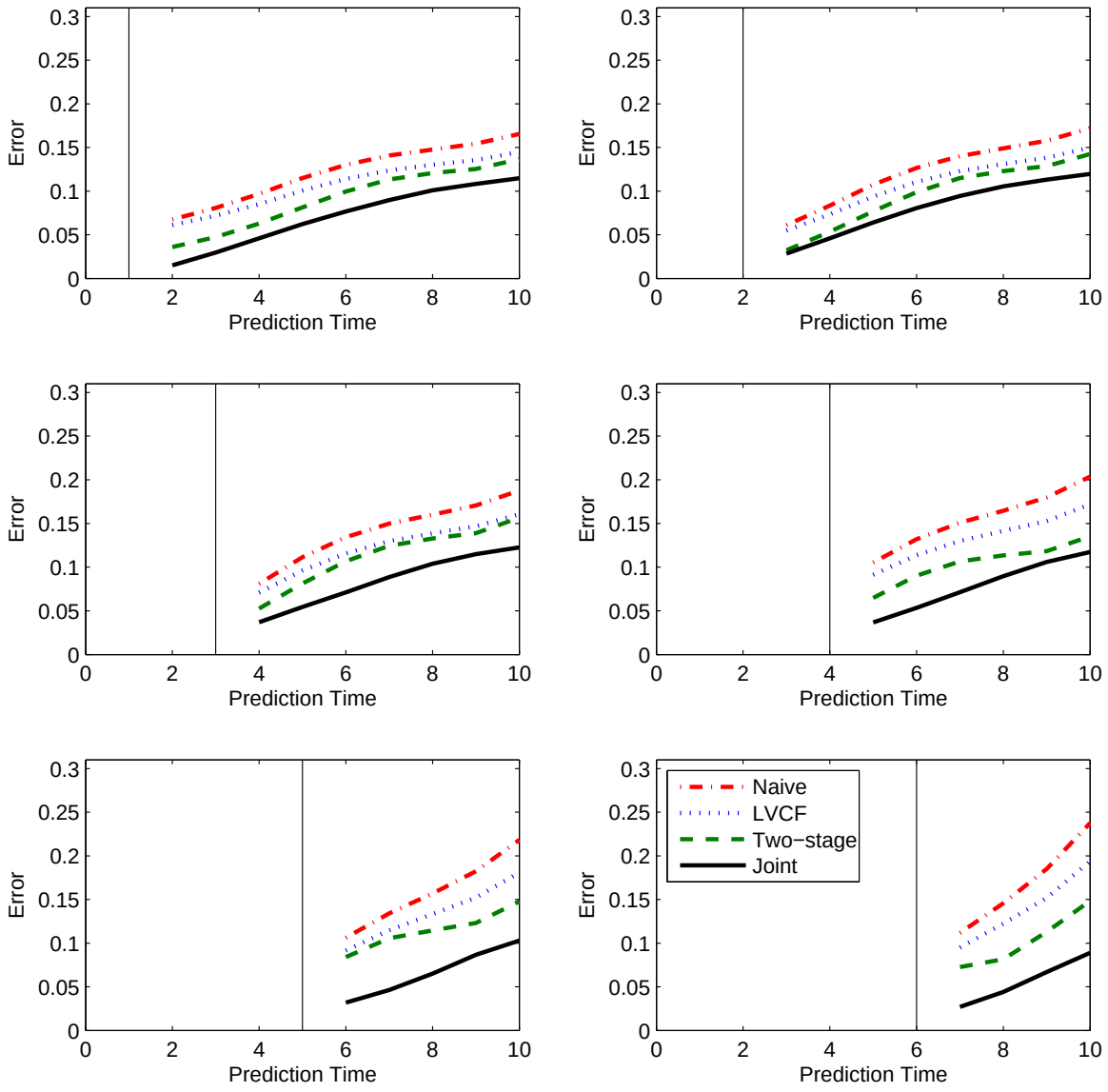

Figure 4.3: Average prediction error ($AE(s,t)$)

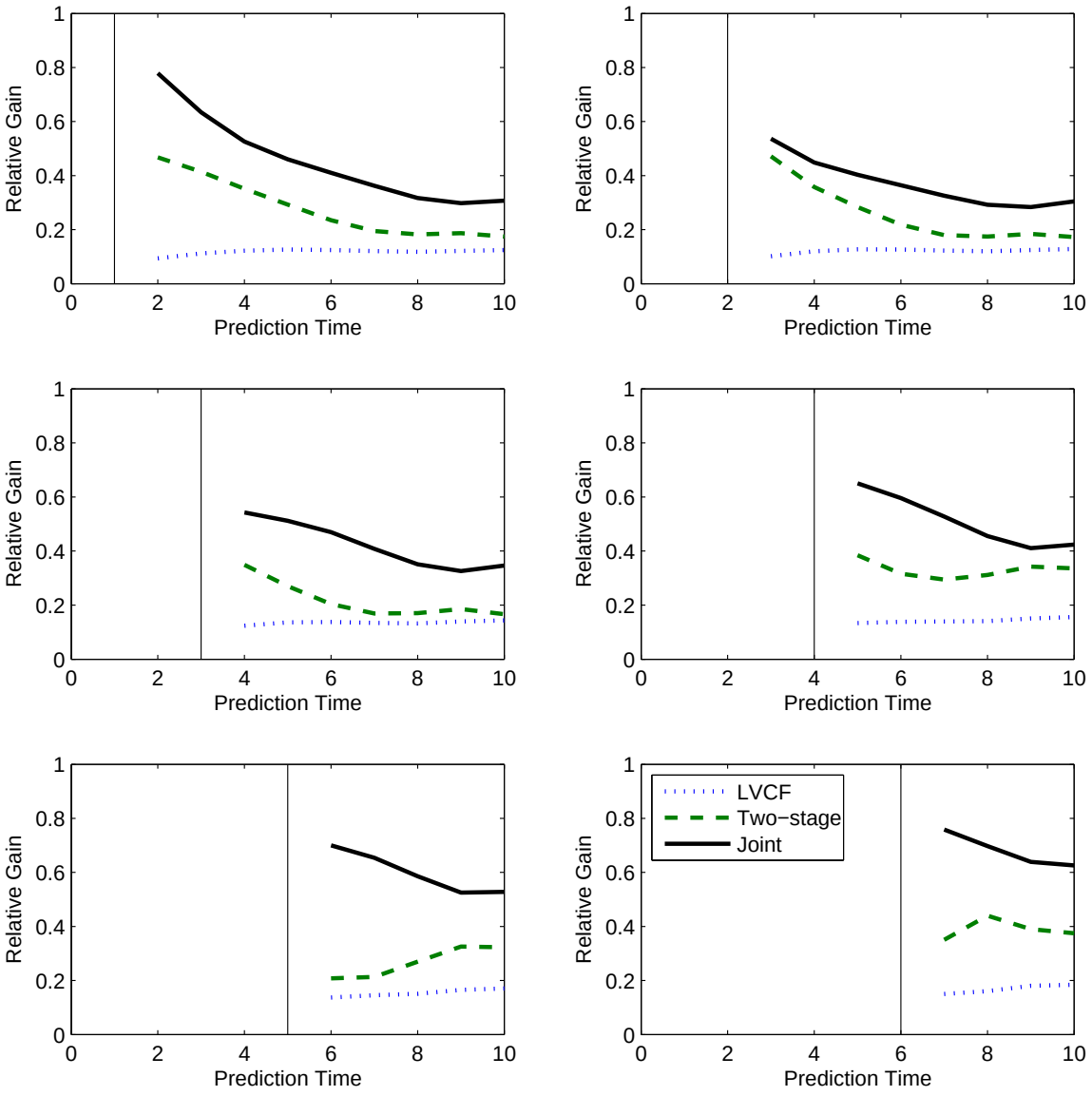

Figure 4.4: Relative predictive accuracy($R(s,t)$)

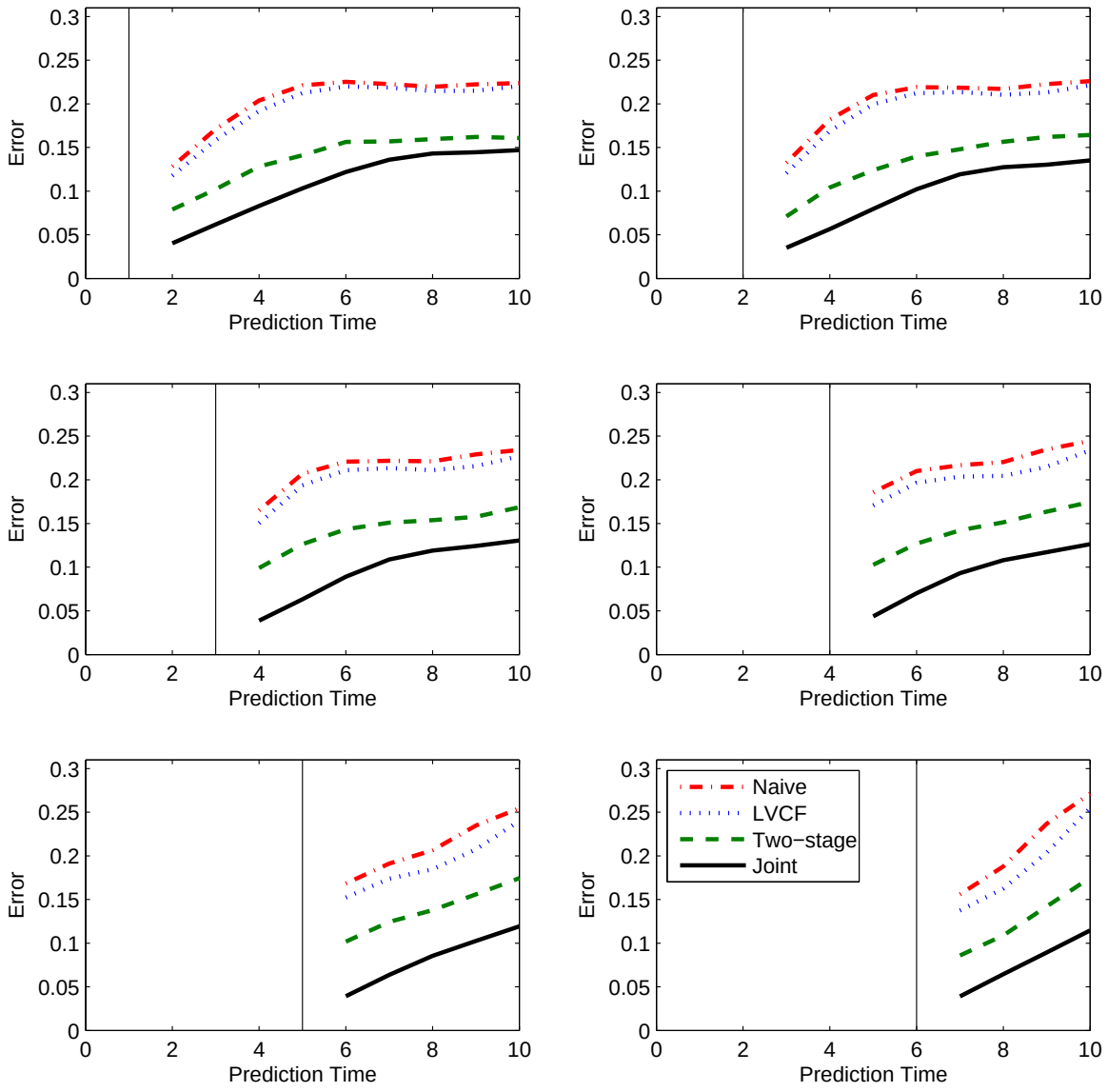

Figure 4.5: Estimated average prediction error ($\hat{WPE}(s,t)$)

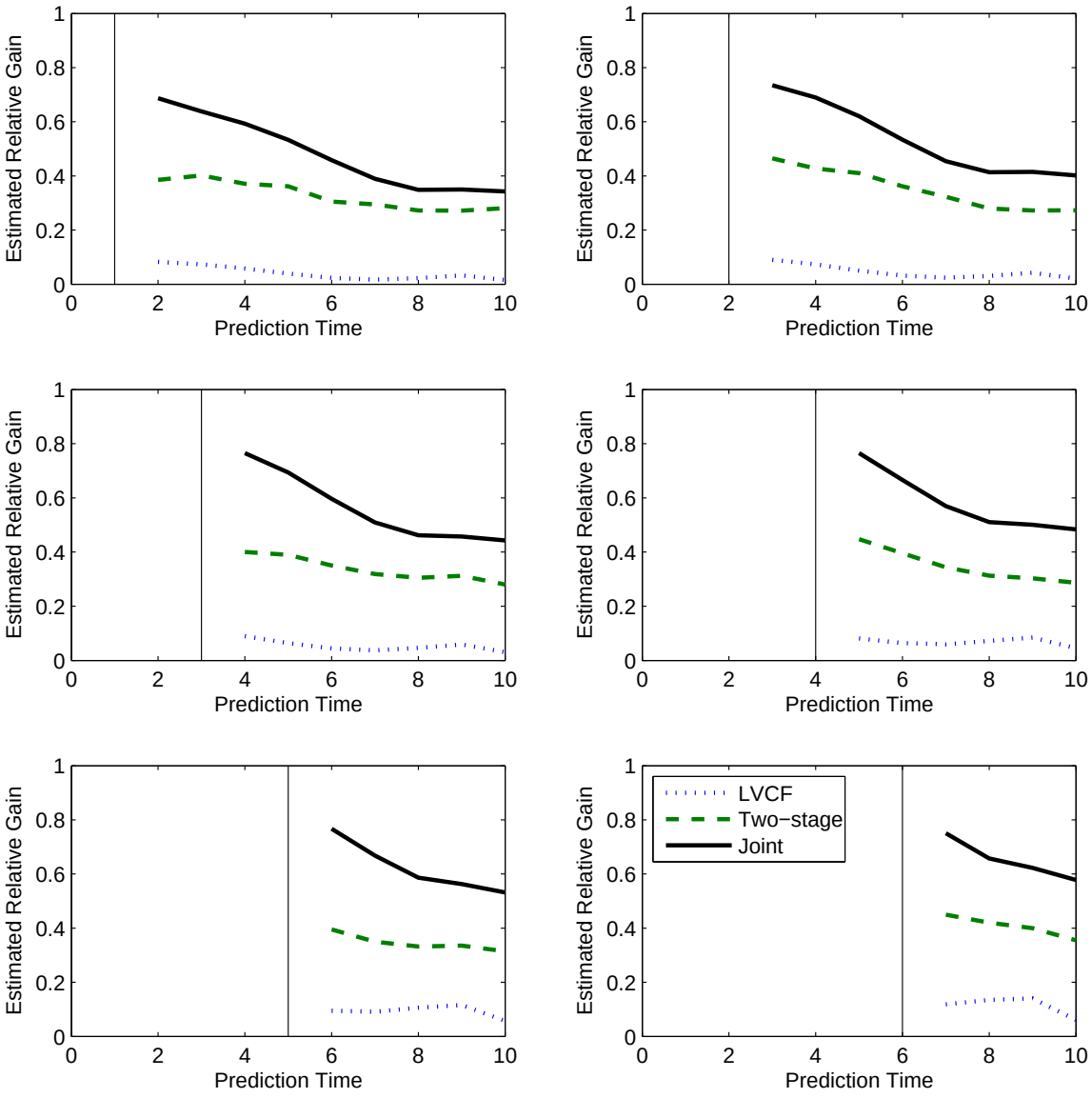

Figure 4.6: Estimated relative predictive accuracy($\hat{R}(s,t)$)

compute the mean of AE and $W\hat{P}E$ across the 200 test sets. The mean time-averaged predition errors (AE) are plotted in figure 4.3. The solid vertical line is the censoring time, which varies from 1 to 6, and the prediction region varies from 2 to 10. The graphs suggest the further away the prediction time is, the higher average prediction errors are for all four models. The naive approach and the last value carry forward approach both yield very similar and relatively high average prediction errors with the last value carry forward approach performing slightly better. Comparing the two-stage model with our joint model, our joint model yields consistently lower average prediction errors across all six different censoring times. Also the later the censoring time s is, the bigger difference in the average prediction error there is between our approach and the two-stage approach. We believe this is because the longer follow-up time in the test set allows more accurate estimates of the longitudinal trajectories, and hence relatively more accurate survival prediction.

Figure 4.4 shows the relative predictive accuracy ($R(t;s)$) comparing the LVCF, two-stage, and the joint model approaches to the naive approach. The joint model approach gives the largest relative gain in terms of the explained variation compared to the other two models. Also the closer the prediction time to the censoring time, the higher the relative predictive accuracy. As the prediction time moves further away from the censoring time, the relative predictive accuracy decreases and then flattens out. The LVCF approach gives a very minimal increase in relative predictive accuracy than the naive approach, and this suggests the last observed longitudinal value is not a very good predictor for the risk of failure in this survival model.

We observe similar results in figure 4.5 which presents the mean estimated weighted average prediction errors ($W\hat{P}E$) at different censoring times and figure 4.6 which gives the estimated relative predictive accuracy ($\hat{R}(t,s)$). That is, our joint model

approach yields larger predictive accuracy compared to the other three alternatives, and therefore is a better prognostic tool in predicting the risk of failure.

4.4.4 Data Application

In chapter 3, we applied our joint model to a prostate cancer data set where we used part of the data as a training set to build our model and estimate parameters, and the remaining data as a test set to validate our model. Now we apply the concepts of predictive errors and relative predictive accuracy to this prostate cancer data, and we evaluate the performance of the joint model in terms of predictive accuracy by using the same strategy in our simulation studies, that is, by comparing it with the three alternative approaches mentioned in section 4.3.2. Specifically, these four approaches are the following. The naive approach:

$$\lambda(t) = \lambda_0(t)exp(\eta_1 log(BaselinePSA_i + 1) + \eta_2 Tstage_i + \eta_3 Gleason_i).$$

The last value carry forward approach:

$$\begin{aligned}\lambda(t) &= \lambda_0(t)exp(\alpha log(PSA_{obs}(t) + 1) \\ &\quad +\eta_1 log(BaselinePSA_i + 1) + \eta_2 Tstage_i + \eta_3 Gleason_i).\end{aligned}$$

The two-stage model:

$$\begin{aligned}\lambda(t) &= \lambda_0(t)exp(\alpha log(\hat{PSA_{TS}}(t) + 1) \\ &\quad +\eta_1 log(BaselinePSA_i + 1) + \eta_2 Tstage_i + \eta_3 Gleason_i + \eta_4 HT_i(t)).\end{aligned}$$

The conditional survival estimates from the joint model in Chapter 3:

$$\begin{aligned}\lambda(t) &= \lambda_0(t)exp(\alpha log(\hat{PSA_{JO}}(t) + 1) + \eta_1 log(BaselinePSA_i + 1) \\ &\quad +\eta_2 Tstage_i + \eta_3 Gleason_i + \eta_4 HT_i(t)).\end{aligned}$$

The estimation method has been described in detail in section 4.4.2, and will not be repeated here.

Our test data were censored at two different time points, 3 year time point and 5 year time point. Figure 4.7 shows the estimated average prediction error ($W\hat{P}E(s,t)$) using the test data censored at year 3. We observe an increasing estimated prediction error as the prediction time passes for all four approaches, and the joint model consistently yields the least amount of estimated prediction error followed by the two-stage method and the LVCF method. The naive method produces the largest estimated prediction error.

Figure 4.8 presents the relative predictive accuracy using the naive method as the reference model, and comparing all the other three methods to it. We can see that within the one to two year time period after the censoring time, the joint method and the two-stage method give very high relative gain in terms of predictive accuracy. This relative gain decreases as the prediction time passes. The joint model yields about 10% more relative gain in variation explained than the two-stage model.

For the test data with censoring time at year 5, we observe very similar results (figure 4.9 and figure 4.10) to those from the test data censored at year 3. The estimated prediction error is about the same when the test data were censored at year 5 as when censored at year 3 for all four approaches. The relative predictive accuracy curves behave similarly as before, but the relative gain is slightly smaller when the test data are censored at year 5 than when censored at year 3.

4.5 Conclusions

In this chapter, we focused on predicting the survival outcome of a disease and evaluating the accuracy of that prediction. We presented absolute distance measures

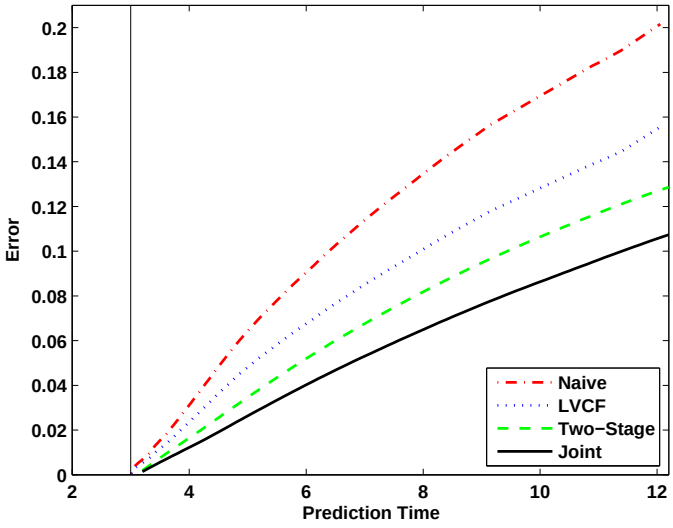

Figure 4.7: Estimated average prediction error ($W\hat{P}E(s,t)$) with censoring time = 3

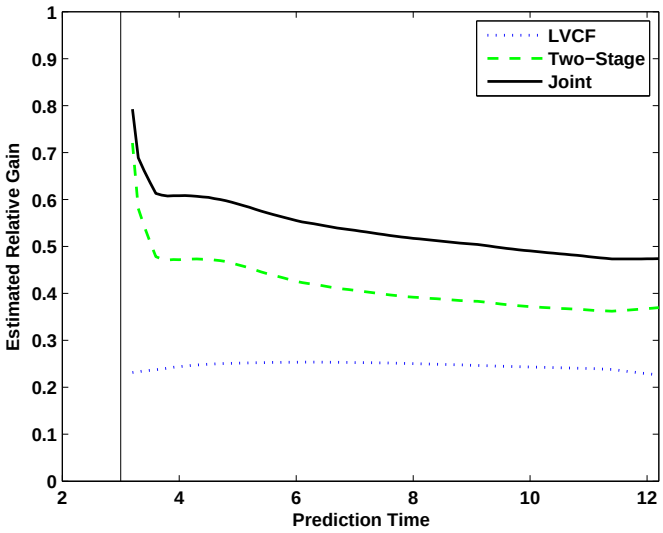

Figure 4.8: Estimated relative predictive accuracy($\hat{R}(s,t)$) with censoring time = 3

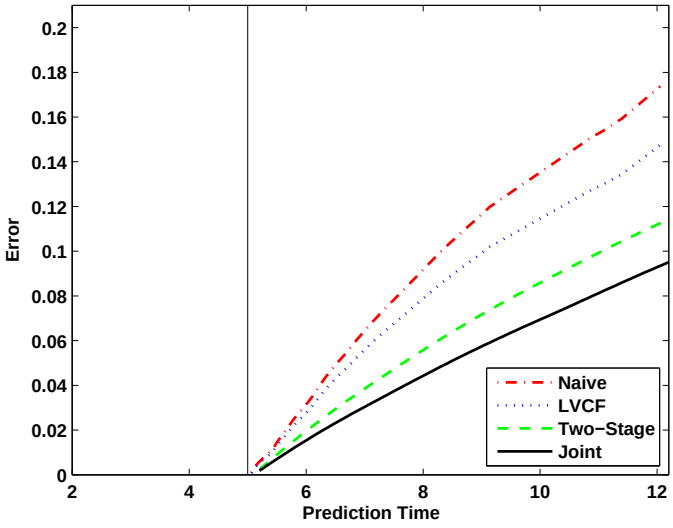

Figure 4.9: Estimated average prediction error ($W\hat{P}E(s,t)$) with censoring time = 5

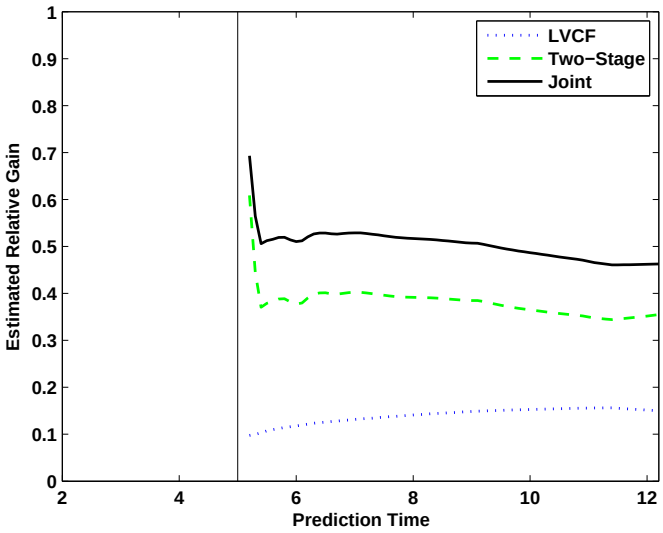

Figure 4.10: Estimated relative predictive accuracy($\hat{R}(s,t)$) with censoring time = 5

to quantify the predictive accuracy of a survival prognostic model. We conducted simulation studies to evaluate the predictive accuracy of the joint model we developed in chapter 3 by comparing it with three alternative approaches: the naive approach, the last value carry forward approach, and the two-stage approach. We divided the simulated data into training data and test data. We built the prognostic models based on the training data sets, and made the predictions of survival probaility on the test data sets with six different censoring times. We formulated true prediction errors based on absolute distance measures, and derived a formula to estimate prediction errors when the true underlying survival probabities are unavailable. We compared these prediction errors obtained from these four approaches using the simulated data, and observed that our model performed consistently better in terms of the prediction accuracy than the three alternative approaches. We believe this is due to the relatively unbiased parameter estimates generated from our joint model whereas the three alternative approaches all introduced different amounts of bias into the parameter estimates as suggested in the literature.

We also applied our predictive accuracy measures to the prostate cancer data we introduced in chapter 3. The results from this data application gave similar conclusions as from the simulated data. We demonstrated that our model can yield fairly high short-term predictive accuracy, which suggested that our model can be a useful prognostic tool for clinicians and patients.

The proposed predictive accuracy measures can incorporate time-dependent covariates in them, and hence allow us to compare different models using frequently updated information. This is especially useful in clinical practice where patients are followed over a long time period and the biomarkers are measured frequently along the way. We can also generate additional R^2-type measures to quantify the gain in

predictive accuracy by taking the ratio of the prediction errors from our model versus the prediction error from a beachmark model such as a marginal survival model or a meaningful reference model. Due to the finite support of the B-splines, we are not able to extrapolate the longitudinal trajectories beyond the last follow-up time, and hence the survival probability prediction is also limited to this time range. The three alternative approaches will still be able to make predictions beyond the last follow-up time, although the predictive accuracy may be decreased even further.

CHAPTER V

Conclusions and Discussions

In the field of survival analysis, we often encounter the situation where a fraction of the study subjects will never experience an event. Cure models have been formulated to address this issue. We developed a family of cure models, indexed by a Box-Cox type transformation parameter, such that different formulations of cure models can be obtained by varying the index parameter. A profile likelihood approach was used for the parameter estimates. Simulation studies were conduted to show unbiasedness. This model was applied to bone marrow transplant data and tonsil cancer data.

Along with survival information, medical studies also often collect longitudinal biomarkers. In recent years, joint models have been proposed to analyze these data simultaneously. We developed a non-parametric joint model of longitudinal biomarker and survival data where the longitudinal trajectories are modeled based on penalized B-splines and linked with the risk of failure by the Cox proportional hazard model. This model can accomodate nonlinearity in the longitudinal trajectories with a great degree of flexibility. A Bayesian approach was used for parameter estimates, and the Metropolis-Hastings algorithm was implemented to construct the MCMC chains. This model was applied to a prostate cancer data, and a validation set was fit to evaluate the model performance.

Furthermore, we evaluated our joint model in terms of its prognostic power by focusing on the predicted conditional survival probabilities. We derived the predicted conditional survival function using our joint model. We proposed absolute distance based measures to assess the predictive accuracy. We carried out simulation studies to evaluate the predictive accuracy of our joint model by comparing it with three alternative approaches: the naive approach, the last value carry forward approach, and a two-stage model. The simulation results showed that our joint model yielded consistently lower average prediction errors, and hence out-performed the other three approaches in terms of its prognostic power.

For future studies, in terms of the family of cure models, we made certain parametric assumptions of the survival function. Future research can explore the possibilities of semiparametric or nonparametric assumptions for the survival function. In our model we used the maximum likelihood approach to obtain the parameter estimates. The Bayesian algorithm can also be used to estimate the posterior distributions of the parameters.

In terms of the joint models of longitudinal and survival data, we applied our model to a prostate cancer data, and we were able to obtain good prediction results. It will be interesting to see this model applied to other medical fields such as in HIV studies. Although we used the estimated longitudinal trajectories and their slopes as time-dependent covariates in the Cox proportional hazard model, further investigation can try to incorporate the entire longitudinal trajectory into the hazard model. In our joint model, we only considered the situation of univariate longitudinal measurements. We believe our model can be easily extended to multivariate repeated measurements. We considered a Cox proportional hazard model for the risk of failure, but the assumption can be relaxed to allow nonproportional hazard situation. We can

furthermore use the proposed absolute distance measures to evaluate the predictive accuracy of these new models and compare with our joint model. Even though we focused on absolute distance measures when we evaluated the predictive accuracy, a quadratic loss function in the form of mean squared error could alternatively be considered.

APPENDIX

APPENDIX A

First and Second Derivatives of the Likelihood Function

Model

$$\begin{aligned}
\frac{S^\lambda - 1}{\lambda} &= \frac{p^\lambda - 1}{\lambda}(1 - S_0) \\
S_0 &= exp(-\tau t^\gamma exp(\beta_1 X_1 + \beta_2 X_2)) \\
h_0 &= \tau\gamma t^{\gamma-1} exp(\beta_1 X_1 + \beta_2 X_2) \\
p &= exp(-exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2))
\end{aligned}$$

Therefore,

$$\begin{aligned}
S &= [p^\lambda + (1 - p^\lambda)S_0]^{1/\lambda} \\
f &= \frac{1 - p^\lambda}{\lambda} S_0 h_0 S^{1-\lambda} \\
L_i &= f_i^\delta S_i^{1-\delta}
\end{aligned}$$

Plugging in, we get

$$L_i = (\frac{(1 - p^\lambda)S_0 h_0}{\lambda})^\delta S^{1-\lambda\delta}$$

Taking the log of both sides, we get

$$logL_i = \delta[log(1 - p^\lambda) - log\lambda + logS_0 + logh_0] + (1 - \lambda\delta)logS$$

First Derivatives:

$$
\begin{aligned}
S(\alpha_0) &= (\frac{\lambda\delta}{p^\lambda - 1} + \frac{(1-\lambda\delta)(1-S_0)}{S^\lambda})p^\lambda logp \\
S(\alpha_1) &= S(\alpha_0)X_1 \\
S(\alpha_2) &= S(\alpha_0)X_2 \\
S(\beta_1) &= (-\frac{\delta t}{\gamma} + \frac{\delta}{h_0} - \frac{(1-\lambda\delta)(1-p^\lambda)S_0 t}{\lambda\gamma S^\lambda})h_0 X_1 \\
S(\beta_2) &= (-\frac{\delta t}{\gamma} + \frac{\delta}{h_0} - \frac{(1-\lambda\delta)(1-p^\lambda)S_0 t}{\lambda\gamma S^\lambda})h_0 X_2 \\
S(\gamma) &= -\frac{\delta h_0 t logt}{\gamma} + \frac{\delta}{\gamma} + \delta logt - \frac{(1-\lambda\delta)(1-p^\lambda)S_0 h_0 t logt}{\lambda\gamma S^\lambda} \\
S(\tau) &= (-\frac{\delta t}{\gamma} + \frac{\delta}{h_0} - \frac{(1-\lambda\delta)(1-p^\lambda)S_0 t}{\lambda\gamma S^\lambda})\frac{h_0}{\tau} \\
S(\lambda) &= \frac{\delta p^\lambda logp}{p^\lambda - 1} - \frac{\delta}{\gamma} - \frac{logS}{\lambda} - \frac{(1-\lambda\delta)(1-S_0)p^\lambda logp}{\lambda S^\lambda}
\end{aligned}
$$

Second Derivatives:

$$
\begin{aligned}
I(\alpha_0,\alpha_0) &= p^{\lambda}[\frac{\delta\lambda(p^{\lambda}-1-\lambda logp)}{(p^{\lambda}-1)^2}+\frac{(1-\delta\lambda)(1-S_0)(\lambda S_0 logp+S^{\lambda})}{S^{2\lambda}}]logp\\
I(\alpha_0,\alpha_1) &= I(\alpha_0,\alpha_0)X_1\\
I(\alpha_0,\alpha_2) &= I(\alpha_0,\alpha_0)X_2\\
I(\alpha_0,\beta_1) &= \frac{(1-\lambda\delta)h_0S_0X_1tp^{\lambda}logp}{\gamma S^{2\lambda}}\\
I(\alpha_0,\beta_2) &= \frac{(1-\lambda\delta)h_0S_0X_2tp^{\lambda}logp}{\gamma S^{2\lambda}}\\
I(\alpha_0,\gamma) &= \frac{(1-\lambda\delta)h_0S_0tp^{\lambda}logtlogp}{\gamma S^{2\lambda}}\\
I(\alpha_0,\tau) &= \frac{(1-\lambda\delta)h_0S_0tp^{\lambda}logp}{\gamma\tau S^{2\lambda}}\\
I(\alpha_0,\lambda) &= [\frac{\delta(p^{\lambda}-1-\lambda logp)}{(p^{\lambda}-1)^2}+\frac{(1-S_0)[-\delta S^{\lambda}+(1-\delta\lambda)S_0logp]}{S^{2\lambda}}]p^{\lambda}logp\\
I(\alpha_1,\alpha_1) &= I(\alpha_0,\alpha_0)X_1^2\\
I(\alpha_1,\alpha_2) &= I(\alpha_0,\alpha_0)X_1X_2\\
I(\alpha_1,\beta_1) &= I(\alpha_0,\beta_1)X_1\\
I(\alpha_1,\beta_2) &= I(\alpha_0,\beta_2)X_1\\
I(\alpha_1,\gamma) &= I(\alpha_0,\gamma)X_1\\
I(\alpha_1,\tau) &= I(\alpha_0,\tau)X_1\\
I(\alpha_1,\lambda) &= I(\alpha_0,\lambda)X_1\\
I(\alpha_2,\alpha_2) &= I(\alpha_0,\alpha_0)X_2^2\\
I(\alpha_2,\beta_1) &= I(\alpha_0,\beta_1)X_2\\
I(\alpha_2,\beta_2) &= I(\alpha_0,\beta_2)X_2\\
I(\alpha_2,\gamma) &= I(\alpha_0,\gamma)X_2\\
I(\alpha_2,\tau) &= I(\alpha_0,\tau)X_2\\
I(\alpha_2,\lambda) &= I(\alpha_0,\lambda)X_2
\end{aligned}
$$

$$
\begin{aligned}
I(\beta_1,\beta_1) &= -\frac{\delta t X_1^2 h_0}{\gamma} - \frac{h_0 S_0 X_1^2 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda)}{\lambda\gamma^2 S^{2\lambda}} \\
I(\beta_1,\beta_2) &= -\frac{\delta t X_1 X_2 h_0}{\gamma} - \frac{h_0 S_0 X_1 X_2 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda)}{\lambda\gamma^2 S^{2\lambda}} \\
I(\beta_1,\gamma) &= -\frac{\delta t X_1 h_0 logt}{\gamma} - \frac{h_0 S_0 X_1 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda) logt}{\lambda\gamma^2 S^{2\lambda}} \\
I(\beta_1,\tau) &= -\frac{\delta t X_1 h_0}{\gamma\tau} - \frac{h_0 S_0 X_1 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda)}{\lambda\gamma^2\tau S^{2\lambda}} \\
I(\beta_1,\lambda) &= \frac{S_0 t h_0 X_1[(1-p^\lambda)S^\lambda + (1-\delta\lambda)p^\lambda\lambda logp]}{\lambda^2\gamma S^{2\lambda}} \\
I(\beta_2,\beta_2) &= -\frac{\delta t X_2^2 h_0}{\gamma} - \frac{h_0 S_0 X_2^2 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda)}{\lambda\gamma^2 S^{2\lambda}} \\
I(\beta_2,\gamma) &= -\frac{\delta t X_2 h_0 logt}{\gamma} - \frac{h_0 S_0 X_2 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda) logt}{\lambda\gamma^2 S^{2\lambda}} \\
I(\beta_2,\tau) &= -\frac{\delta t X_2 h_0}{\gamma\tau} - \frac{h_0 S_0 X_2 t(1-\delta\lambda)(1-p^\lambda)(\gamma S^\lambda - h_0 t p^\lambda)}{\lambda\gamma^2\tau S^{2\lambda}} \\
I(\beta_2,\lambda) &= \frac{S_0 t h_0 X_2[(1-p^\lambda)S^\lambda + (1-\delta\lambda)p^\lambda\lambda logp]}{\lambda^2\gamma S^{2\lambda}} \\
I(\gamma,\gamma) &= -\frac{\delta t h_0 (logt)^2}{\gamma} - \frac{\delta}{\gamma^2} - \frac{(1-\lambda\delta)(1-p^\lambda)S_0 h_0 t(\gamma S^\lambda - h_0 t p^\lambda)(logt)^2}{\lambda\gamma^2 S^{2\lambda}} \\
I(\gamma,\tau) &= -\frac{\delta t h_0 logt}{\gamma\tau} - \frac{(1-\lambda\delta)(1-p^\lambda)S_0 h_0 t(\gamma S^\lambda - h_0 t p^\lambda) logt}{\lambda\gamma^2\tau S^{2\lambda}} \\
I(\gamma,\lambda) &= \frac{S_0 t h_0 logt[(1-p^\lambda)S^\lambda + (1-\delta\lambda)p^\lambda\lambda logp]}{\lambda^2\gamma S^{2\lambda}} \\
I(\tau,\tau) &= -\frac{\delta}{\tau^2} - \frac{(1-\lambda\delta)(1-p^\lambda)t^2 h_0^2 S_0 p^\lambda}{\lambda\gamma^2\tau^2 S^{2\lambda}} \\
I(\tau,\lambda) &= \frac{S_0 t h_0[(1-p^\lambda)S^\lambda + (1-\delta\lambda)p^\lambda\lambda logp]}{\lambda^2\gamma\tau S^{2\lambda}} \\
I(\lambda,\lambda) &= -\frac{\delta p^\lambda (logp)^2}{(1-p^\lambda)^2} + \frac{\delta}{\lambda^2} - \frac{(1-S_0)p^\lambda logp}{\lambda^2 S^\lambda} + \frac{2logS}{\lambda^2} + \\
&\quad \frac{(1-S_0)[-p^\lambda S^\lambda + (1-\delta\lambda)\lambda p^\lambda S_0 logp]logp}{\lambda^2 S^{2\lambda}}
\end{aligned}
$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. Berkson and R.P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515, 1952.

[2] E. Brown, J.G. Ibrahim, and V. DeGruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61:64–73, 2005.

[3] P. Bycott and J.M.G. Taylor. A comparison of smoothing techniques for cd4 data measured with error in a time-dependent cox proportion hazards model. *Statistics in Medicine*, 17:2061–2077, 1999.

[4] M.H. Chen, J.G. Ibrahim, , and S.R. Lipsitz. Bayesian methods for missing covariates in cure rate models. *Lifetime Data Analysis*, 8:117–146, 2002.

[5] M.H. Chen and J.G. Ibrahim. Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, 57:43–52, 2001.

[6] M.H. Chen, J.G. Ibrahim, and D. Sinha. A new bayesian model for survival data with a survival fraction. *Journal of the American Statistical Association*, 94:909–919, 1999.

[7] D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–200, 1972.

[8] U.G. Dafni and A.A. Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54:1445–1462, 1998.

[9] V. DeGruttola and X.M. Tu. Modeling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

[10] P. Eilers and B. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121, 1996.

[11] P.H.C. Eilers and B.D. Marx. Splines, knots, and penalties. *Technical report*, page Available at http://www.stat.lsu.edu/bmarx/, 2004.

[12] L. Fahrmeir, T. Kneib, and S. Lang. Penalized structured additive regression for space-time data: a bayesian perspective. *Statistical Sinica*, 14:715–745, 2004.

[13] V.T. Farewell. A model for a binary variable with time censored observations. *Biometrika*, 64:43–46, 1977.

[14] V.T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

[15] C.L. Faucett and D.C. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in Medicine*, 15:1663–1685, 1999.

[16] T. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48:1029–1040, 2006.

[17] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo In Practice*. Chapman and Hall, London, 1996.

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

[19] P. Heagerty, T. Lumley, and M. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–344, 2000.

[20] P. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61:92–105, 2005.

[21] R. Henderson, P. Diggle, and A. Dobson. Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3:33–50, 2002.

[22] C.S. Li and J.M.G. Taylor. A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21:3235–3247, 2002.

[23] D.K. Pauler and D.M. Finkelstein. Predicting time to prostate cancer recurrence based on joint models for nonlinear longitudinal biomarkers and event time outcomes. *Statistics in Medicine*, 21:3897–3911, 2002.

[24] Y. Peng and K.B.G. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243, 2000.

[25] Y. Peng, K.G.B. Dear, and K.C. Carriere. Testing for the presence of cured patients: a simulation study. *Statistics in Medicine*, 20:1783–1796, 2001.

[26] S. Ratcliffe, W. Guo, and T.R. Ten Have. Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, 60:892–899, 2004.

[27] M. Schemper and R. Henderson. Predictive accuracy and explained variation in cox regression. *Biometrics*, 56:249–255, 2000.

[28] R. Schoop, E. Graf, and M. Schumacher. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64:603–610, 2008.

[29] J.P. Sy and J.M.G. Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56:227–236, 2000.

[30] J.M.G. Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907, 1995.

[31] J.M.G. Taylor, M. Yu, and H.M. Sandler. Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology*, 23:816–825, 2004.

[32] A.A. Tsiatis and M. Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88:447–458, 2001.

[33] A.A. Tsiatis and M. Davidian. An overview of joint modeling of longitudinal and time-to-event data. *Statistical Sinica*, 14:793–818, 2004.

[34] A.A. Tsiatis, V. DeGruttola, and M.S. 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*, 90:27–37, 1995.

[35] A. Tsodikov. A proportional hazard model taking account of long-term survivors. *Biometrics*, 54:1508–1516, 1998.

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

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

[38] M.S. Wulfsohn and A.A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

[39] J. Xu and S.L. Zeger. The evaluation of multiple surrogate endpoints. *Biometrics*, 57:81–87, 2001.

[40] J. Xu and S.L. Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, 50:375–387, 2001.

[41] A.Y. Yakovlev and A.D. Tsodikov. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, River Edge, New Jersey, 1996.

[42] G. Yin and J.G. Ibrahim. Cure rate models: a unified approach. *Canadian Journal of Statistics*, 33:559–570, 2005.

[43] G. Yin and J.G. Ibrahim. A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics*, 61:403–412, 2005.

[44] M. Yu, N.J. Law, J.M.G. Taylor, and H.M. Sandler. Joint longitudinal-survival cure models and their application to prostate cancer. *Statistical Sinica*, 14:835–862, 2004.

[45] D. Zeng, G. Yin, and J.G. Ibrahim. Inference for a class of transformed hazards models. *Journal of the American Statistical Association*, 100:1000–1008, 2005.

[46] D. Zeng, G. Yin, and J.G. Ibrahim. Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101:670–684, 2006.

[47] D. Zhang, X. Lin, J. Raz, and M. Sowers. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Assocation*, 93:710–719, 1998.

[48] Y. Zheng and P. Heagerty. Prospective accuracy for longitudinal markers. *Biometrics*, 63:332–341, 2007.