HHS Public Access

Author manuscript

Stat Modelling. Author manuscript; available in PMC 2022 February 01.

Published in final edited form as:

Stat Modelling. 2021 February ; 21(1-2): 72–94. doi:10.1177/1471082X20944620.

Joint modelling of longitudinal and survival data in the presence of competing risks with applications to prostate cancer data

Md. Tuhin Sheikh[1], **Joseph G. Ibrahim**[2], **Jonathan A. Gelfond**[3], **Wei Sun**[4], **Ming-Hui Chen**[1]

[1]Department of Statistics, University of Connecticut, Storrs, CT, USA.

[2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[3]Department of Epidemiology and Biostatistics, University of Texas Health San Antonio, San Antonio, TX, USA.

[4]Biostatistics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

Abstract

This research is motivated from the data from a large Selenium and Vitamin E Cancer Prevention Trial (SELECT). The prostate specific antigens (PSAs) were collected longitudinally, and the survival endpoint was the time to low-grade cancer or the time to high-grade cancer (competing risks). In this article, the goal is to model the longitudinal PSA data and the time-to-prostate cancer (PC) due to low- or high-grade. We consider the low-grade and high-grade as two competing causes of developing PC. A joint model for simultaneously analysing longitudinal and time-to-event data in the presence of multiple causes of failure (or competing risk) is proposed within the Bayesian framework. The proposed model allows for handling the missing causes of failure in the SELECT data and implementing an efficient Markov chain Monte Carlo sampling algorithm to sample from the posterior distribution via a novel reparameterization technique. Bayesian criteria, ΔDIC_{Surv}, and $\Delta WAIC_{Surv}$, are introduced to quantify the gain in fit in the survival sub-model due to the inclusion of longitudinal data. A simulation study is conducted to examine the empirical performance of the posterior estimates as well as ΔDIC_{Surv} and $\Delta WAIC_{Surv}$ and a detailed analysis of the SELECT data is also carried out to further demonstrate the proposed methodology.

Keywords

cause-specific competing risks model; DIC; mixed effects model; reparametrization; SELECT data; WAIC

Address for correspondence: Joseph G. Ibrahim, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. ibrahim@bios.unc.edu.

Supplementary materials

Supplementary materials for this article are available at http://www.statmod.org/smij/archive.html.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

1 Introduction

In many clinical trials and medical studies, often both longitudinal (repeated measurements of a response) and time-to-event (time until the occurrence of an event of interest) outcomes are collected along with some other baseline covariates. Studies dealing with such outcomes mainly concern to investigate how the change in longitudinal outcome is associated with different covariates and also to determine the relationship among the longitudinal outcome, the survival outcome and other covariates. The traditional random effects model (Laird and Ware, 1982) and Cox proportional hazards model (Cox, 1972) may be considered for analysing the longitudinal and survival outcomes separately. However, a separate analysis of these two outcomes often leads to biased estimates of the model parameters (Hu et al., 2009). It is due to the fact that the longitudinal measurements are often incomplete (missing) and are subject to measurement error (Tsiatis et al., 1995), which may induce informative censoring mechanism for the survival outcome (Wulfsohn and Tsiatis, 1997). Thus, treating longitudinal outcome as a time-dependent covariate in the survival model fails to capture the endogenity of the longitudinal process and may lead us to a wrong interpretation of the model parameters (Prentice, 1982). In such cases, joint modelling of longitudinal and survival outcomes (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Rizopoulos, 2012) has been widely used in clinical studies.

The standard joint model assumes that the event of interest takes place due to only one cause. However, the event of interest may occur due to one of the many plausible causes under consideration. This study is motivated from the Selenium and Vitamin E Cancer Prevention Trial (SELECT), where the main variable of interest is the time-to-diagnosis of prostate cancer (PC) due to low-grade or high-grade. We consider the two grades of PC as two distinct causes of PC. In the SELECT data, prostate specific antigens (PSAs) of the subjects are collected longitudinally, which is considered to be an important biomarker for the PC. In this study, our goal is to quantify the improvement in the fit of the time-to-PC due to low or high grade after the inclusion of longitudinal PSA in the survival sub-model. We also intend to quantify the cause-specific association between the time-to-PC and longitudinal progression of PSA.

Joint modelling of longitudinal and competing risks survival data has been proposed in the literature. In the joint modelling setting, Elashoff et al. (2007, 2008) extended the cause-specific hazards model and the mixture model by allowing latent variables to adjust for the association between the longitudinal and survival outcomes. However, the development in this area has not been in line under the Bayesian paradigm. A Bayesian extension of the cause-specific hazards models to analyse longitudinal and survival outcomes has been proposed by Huang et al. (2011) and Hu et al. (2009). However, the earlier works do not account for the missing causes of failure, which is common in clinical studies.

In the SELECT data, there are a substantial number of subjects who developed PC without knowing whether PC should be attributed to low-grade or high-grade. This missing cause of PC could happen when subjects reported cancer diagnostics outside the studies home institution. Considering the subjects with unknown causes as missing or ignoring those subjects from the analysis may lead to biased parameter estimates and thus the inference

could be misleading (Lu and Tsiatis, 2001). Particularly, when the missing percentage is substantial (e.g., about 26% for SELECT data), the analysis could be erroneous. Within the competing risks setting, multiple imputation technique has been proposed by Lu and Tsiatis (2001), while Gao and Tsiatis (2005) developed an inverse probability weighted complete case estimator. Within the joint modelling framework, the literature to account for the missing causes under the competing risks setting is still sparse. In this study, we propose a shared parameter joint model within the Bayesian framework. Our proposed joint model imputes the missing causes of PC within the Markov chain Monte Carlo (MCMC) sampling algorithm.

In the SELECT data, the substantially large sample size of 32 261 might also impact the performance of the joint model resulting in weak convergence of the variance covariance matrix of the random effects (Zhang et al., 2019). This can also influence the convergence of other parameters that depend on the random effects in MCMC sampling. To deal with this issue, we use the Cholesky factorization of the variance covariance matrix of the random effects similar to Zhang et al. (2019) and propose a novel reparameterization of the regression coefficients associated with the random effects under the survival sub-model, which leads to a convenient but efficient implementation of the MCMC sampling algorithm under the joint model.

One of the main goals of this study is to quantify the improvement in the fit of the competing risks survival data due to the inclusion of the longitudinal PSA. Several measures, for example, Akaike information criterion (AIC), Bayesian information criterion (BIC), widely applicable or Watanabe–Akaike information criterion (WAIC; Watanabe, 2010), deviance information criterion (DIC; Spiegelhalter et al., 2002), etc., have been routinely used to quantify the overall fit of the model under consideration. Under the joint modelling setting, in order to quantify the overall goodness of fit, it is often desirable to assess the separate contribution of each component of the joint model towards the overall fit. Zhang et al. (2014) developed a decomposition of AIC and BIC under the joint model of longitudinal and survival outcomes. A useful SAS macro has been developed by Zhang et al. (2016) that allows to use various flexible models under the joint modelling framework and computes the decomposition of AIC and BIC. Within the Bayesian framework, a novel decomposition of DIC and logarithm of the pseudo-marginal likelihood (LPML; Ibrahim et al., 2001) has been proposed by Zhang et al. (2017). To investigate the performance of our proposed model, we define a BIC, ΔDIC_{Surv} that quantifies the gain in the performance of joint model due to the inclusion of the longitudinal data. In addition, we extend the idea of Huang et al. (2005) to adjust for the unobserved random effects while calculating DIC under the joint model within the Bayesian framework. We also define $\Delta WAIC_{Surv}$, and the gain in fit of the SELECT data under the proposed model has been assessed by ΔDIC_{Surv} and $\Delta WAIC_{Surv}$. The Bayesian criteria, ΔDIC_{Surv} and $\Delta WAIC_{Surv}$, could be related to the association parameters, which are the regression coefficients associated with the random effects (‘covariates’) under the survival sub-model. To assess the change in ΔDIC_{Surv} and $\Delta WAIC_{Surv}$ due to the change in the association parameters, we carry out a simulation study under the two different scenarios of the association parameters.

The remainder of the article is organized as follows. The description of the SELECT data and some preliminary statistics are presented in Section 2. The development of the proposed methodologies is discussed in Section 3. The Bayesian inference and computation are presented in Section 4. The simulation study design and results are presented in Section 5. A detailed analysis of the SELECT data is carried out in Section 6. We conclude the article with a brief discussion in Section 7. The technical details of MCMC sampling are given in the Supplementary Materials (http://www.statmod.org/smij/archive.html).

2 SELECT data

The dataset for this study is extracted from the SELECT data which was sponsored by National Cancer Institute. Enrolment of the patients into this study started in 2001 and ended in 2004. A total of 35 261 male patients from United States, Puerto Rico and Canada participated in this study. In this study, race, Hispanic status, family history of cancer and smoking status were collected at baseline. Also, height, weight, body mass index (BMI) and PSA were collected longitudinally over the study period. The follow-up process for a patient was terminated when the patient developed PC due to low-grade or high-grade.

The main variable of interest is the time-to-PC due to low-grade or high-grade (survival outcome), where the low- and high-grades are considered as two different causes of PC. The main objective of this study is to investigate the association between the longitudinal PSA (longitudinal outcome) and the competing risks survival outcome adjusting for other covariates. We include the cases in the analysis, in which the number of follow-up visits is at least two, leading to a total of 22 792 patients. The follow-up visiting times and event times are recorded in years for each patient. For the longitudinal PSAs, the median follow-up times are 4.02 years and 1.95 years, respectively, for the censored and observed patients with an overall median follow-up time of 3.96 years. The observed minimum, median and maximum survival times are 0.003, 8.071 and 14.249 years, respectively. The median number of follow-up visits for the patients is 6 with the minimum and maximum of 2 and 10, respectively. Figure 1 shows the percentage distribution of the number of repeated measurements for different censoring status. We observe that for low-grade and high-grade, the percentage for the patients with less than five repeated measurements is higher compared to those for censored and missing-grade patients.

In Table 1, the distribution of PC status and the causes are presented. It is observed that about 7% patients developed cancer among whom for more than 26% patients, the cause of PC is missing. In Figure 2, the longitudinal trajectories and corresponding survival times for eight selected patients are presented. The figure shows that with the increasing PSA trajectory, the survival time tends to be lower compared to those for non-increasing PSA trajectory cases. This motivates us to develop the model to investigate the association between longitudinal PSA measures and the time to PC due to two different causes. Also, the missing cause cases should also be taken care of by imputation in MCMC sampling for carrying out an appropriate inference of the SELECT data.

3 Joint modelling of longitudinal and competing risks survival data

The joint model is comprised of a longitudinal sub-model and a survival sub-model. In the following subsections, the notation and the sub-models are defined under the general setting.

3.1 Longitudinal sub-model

Suppose there are n subjects in the study and for the ith subject, m_i longitudinal measurements are collected for $i = 1,\ldots,n$. Let $y_i(a_{ij})$ denote the longitudinal measurement at time a_{ij} for $j = 1,\ldots,m_i$ and $i = 1,\ldots,n$. We assume a mixed effects model (Verbeke and Molenberghs, 2009) for the longitudinal outcome $y_i(a_{ij})$ given by

$$y_i(a_{ij}) = \mathbf{g}(a_{ij})'\boldsymbol{\theta}_i + \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i(a_{ij}), \tag{3.1}$$

where $\mathbf{g}(a_{ij})$ is a $(q+1)$-dimensional vector of functions of time a_{ij}, $\boldsymbol{\theta}_i$ is a $(q+1)$-dimensional vector of the random effects, $\boldsymbol{\gamma}$ is a p-dimensional vector of regression coefficients corresponding to the p-dimensional covariates $\mathbf{x}_i$ and the measurement error $\epsilon_i(a_{ij})$ is assumed to follow $N(0, \sigma^2)$. We assume that $\boldsymbol{\theta}_i = (\boldsymbol{\theta} + \boldsymbol{\theta}_i^*)$, where $\boldsymbol{\theta}_i^* \sim N(0, \Omega)$, Ω is a $(q+1) \times (q+1)$ positive definite variance-covariance matrix, and $\boldsymbol{\theta}$ is the vector of overall effects. We further assume that the subject-specific random effects $\boldsymbol{\theta}_i$ and the measurement error terms $\boldsymbol{\epsilon}_i(a_{ij})$'s are independent. We note that the trajectory function $\mathbf{g}(a_{ij})'\boldsymbol{\theta}_i$ is a linear trajectory if $q = 1$, $\mathbf{g}(a_{ij})' = (1, a_{ij})$ and a quadratic trajectory if $q = 2$, $\mathbf{g}(a_{ij})' = (1, a_{ij}, a_{ij}^2)$. The trajectory function captures the unobserved subject-specific progression of the longitudinal outcome, and it is a critical component in the longitudinal sub-model.

This study is motivated from the PC data-SELECT (discussed in Section 2), where PSA measurements are collected longitudinally. The earlier studies (Ferrer et al., 2016) suggest that the logarithm of PSAs can be explained by a linear mixed effects model with a latent linear trajectory function of observation time. In this study, we consider a linear trajectory for the longitudinal sub-model defined in (3.1). Within the Bayesian framework, the posterior estimation of Ω depends on the data only through the random effects $\boldsymbol{\theta}_i^*$'s which results in slow convergence of MCMC sampling (Zhang et al., 2019). To overcome the convergence issue of the covariance matrix of random effects, we consider the Cholesky decomposition of $\Omega = \Gamma\Gamma'$ and introduce a novel reparameterization

$$\boldsymbol{\theta}_i^* = \Gamma\boldsymbol{\theta}_i^R, \tag{3.2}$$

where $\boldsymbol{\theta}_i^R \sim N(0, I_2)$. We redefine the linear mixed effects model using this reparameterization as

$$\mathbf{y}_i = \mathbf{g}(a_{ij})'(\boldsymbol{\theta} + \Gamma\boldsymbol{\theta}_i^R) + \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i, \tag{3.3}$$

where $\Gamma = \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix}$ with $b_{11} > 0$ and $b_{22} > 0$. This reparameterization will facilitate fast convergence and better mixing of Gibbs sampling from the resulting posterior distribution

for the covariance matrix of random effects. More details regarding the posterior computation and estimation are discussed in Section 4.

3.2 Survival sub-model

Let t_i be the event time or censoring time and also let δ_i denote the censoring indicator for the ith subject. The censoring indicator δ_i takes a value of 0 for a censored event and $\delta_i = k$ indicates that an event occurred for the ith subject due to the kth cause for $k = 1, 2$. In this article, we consider the cause-specific hazards model in which, the random effects are included as covariates along with other fixed covariates. The proposed model is defined on the setting when there are two causes of failure and the dimension of the random effects is two. The cause-specific hazards model for the kth observed cause of failure is defined as

$$h_k(t \mid \mathbf{z}_i) = h_{k0}(t)\exp(\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i^* + \mathbf{z}_i'\boldsymbol{\beta}_k),\ k = 1, 2, \tag{3.4}$$

where $\boldsymbol{\beta}_k$ is the vector of the coefficients corresponding to the fixed effects $\mathbf{z}_i$, $\boldsymbol{\alpha}_k$ is the vector of regression coefficients for the random effects $\boldsymbol{\theta}_i^*$ and $h_{k0}(t)$ is the baseline hazard function for the kth cause of the event at time t. In the survival sub-model, the random effects are linked to the longitudinal sub-model through the association parameter $\boldsymbol{\alpha}_k$. To account for the missing causes of failure for the uncensored subjects, we propose the following additive hazards model

$$h_{12}(t \mid \mathbf{z}_i) = \sum_{k=1}^{2} h_{k0}(t)\exp(\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i^* + \mathbf{z}_i'\boldsymbol{\beta}_k)\,. \tag{3.5}$$

With the reparameterization of $\boldsymbol{\theta}_i^* = \Gamma\boldsymbol{\theta}_i^R$, the random component of the survival sub-model can be expressed as

$$\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i^* = (\alpha_{k1}b_{11} + \alpha_{k2}b_{21})\theta_{i1}^R + (\alpha_{k2}b_{22})\theta_{i2}^R = \boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R, k = 1, 2,$$

where

$$\alpha_{k1}^* = \alpha_{k1}b_{11} + \alpha_{k2}b_{21}, \alpha_{k2}^* = \alpha_{k2}b_{22}\,. \tag{3.6}$$

With this reparameterization, the hazard model (3.4) for the kth cause of event is redefined as

$$h_k(t \mid \mathbf{z}_i) = h_{k0}(t)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k),\ k = 1, 2, \tag{3.7}$$

and the hazard function (3.5) for the unknown cause of failure is redefined as

$$h_{12}(t \mid \mathbf{z}_i) = \sum_{k=1}^{2} h_{k0}(t)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k)\,. \tag{3.8}$$

We further assume that the baseline hazard function due to the kth cause has a piecewise constant form with G_k partitions of the time axis, $0 = s_{k0} < s_{k1} < \cdots < s_{kG_k} = \infty$,

$$h_{k0}(t) = \lambda_{kg}, \quad t \in (s_{k,g-1}, s_{k,g}], \; g = 1, 2, \ldots, G_k, k = 1, 2. \tag{3.9}$$

Let $[\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kG_k}), k = 1, 2\}$ denote the vectors of piecewise constants of the baseline hazard functions defined in (3.9). The likelihood function under the proposed joint model is constructed in the following subsection.

3.3 Likelihood construction

Let $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2)$, where $\boldsymbol{\varphi}_1 = (\boldsymbol{\theta}^R, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2, \Gamma)$, $\boldsymbol{\varphi}_2 = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\theta}^R = (\boldsymbol{\theta}_1^{R'}, \ldots, \boldsymbol{\theta}_n^{R'})'$. Let $\mathrm{D}_{\mathrm{obs}} = (\mathrm{D}_{\mathrm{Long,obs}}, \mathrm{D}_{\mathrm{Surv,obs}})$ denote the observed data, where $\mathrm{D}_{\mathrm{Long,obs}} = \{(\mathbf{y}_i, \mathbf{x}_i), i = 1,\ldots,n\}$ and $\mathrm{D}_{\mathrm{Surv,obs}} = \{(t_i, \delta_i, \mathbf{z}_i), i = 1,\ldots,n\}$. For the ith subject, the censoring indicator $\delta_i \in \{0, 1, 2\}$, where 0 indicates censoring, 1 and 2 indicate two distinct causes of failure. For instance, in the SELECT data, δ_i takes a value of 0 for no PC, 1 for low-grade cancer, and 2 for high-grade cancer. Let u_i take a value of 1 when the ith subject is uncensored due to an unknown cause and takes a value of 0 when the cause of failure for the ith subject is known. In this article, we consider the unobserved random effects as the latent or unknown parameters, conditional on the random effects, the longitudinal data are independent of the survival data. The joint distribution of $(\mathbf{y}_i, t_i, \boldsymbol{\theta}_i^R)$ is written as

$$f(\mathbf{y}_i, t_i, \boldsymbol{\theta}_i^R \mid \varphi, \delta_i, \mathbf{x}_i, \mathbf{z}_i) = f(\mathbf{y}_i \mid \boldsymbol{\theta}_i^R, \boldsymbol{\varphi}_1, \mathbf{x}_i) f(t_i \mid \delta_i, \boldsymbol{\theta}_i^R, \boldsymbol{\varphi}_2, \mathbf{z}_i) f(\boldsymbol{\theta}_i^R), \tag{3.10}$$

where

$$\begin{aligned}
f(\mathbf{y}_i \mid \boldsymbol{\theta}_i^R, \boldsymbol{\varphi}_1, \mathbf{x}_i) &= \frac{1}{(2\pi\sigma^2)^{m_i/2}} \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}_i - \mathbf{g}(a_{ij})'(\boldsymbol{\theta} + \Gamma\boldsymbol{\theta}_i^R) + \mathbf{x}_i'\boldsymbol{\gamma}\right)'\left(\mathbf{y}_i - \mathbf{g}(a_{ij})'(\boldsymbol{\theta} + \Gamma\boldsymbol{\theta}_i^R) + \mathbf{x}_i'\boldsymbol{\gamma}\right)\right), \\
f(t_i \mid \delta_i, \boldsymbol{\theta}_i^R, \boldsymbol{\varphi}_2, \mathbf{z}_i) &= \left[\prod_{k=1}^{2} \{h_{k0}(t_i)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k)\}^{I(\delta_i = k, u_i = 0)}\right] \\
&\quad \times \exp\left(-\sum_{k=1}^{2} H_{k0}(t_i)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k)\right) \\
&\quad \times \left[\sum_{k=1}^{2} h_{k0}(t_i)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k)\right]^{I(\delta_i > 0, u_i = 1)}, \\
f(\boldsymbol{\theta}_i^R) &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}\boldsymbol{\theta}_i^{R'}\boldsymbol{\theta}_i^R\right).
\end{aligned}$$

In the following Section 4, we present the Bayesian estimation of the parameters under the joint model.

4 Bayesian inference

The Bayesian estimation of the proposed model involves sampling from the full conditional distributions, adaptive rejection sampling (Gilks and Wild, 1992) and Metropolis–Hastings sampling(Metropolis et al., 1953; Hastings, 1970). In the SELECT data, there are a substantial percentage of missing causes of failure. In MCMC sampling from the posterior distribution, we add one additional step to impute the missing causes of failure. Let $\{\delta_i^m, i \in \mathcal{J}_m\}$ be the set of censoring indicators with the missing grade, where $\mathcal{J}_m$ be the set of patients with missing grade. The occurrence of PC is decided due to low-grade with probability

$$\frac{1}{1+\exp(-\{\log\frac{h_{10}(t_i)}{h_{20}(t_i)}+\boldsymbol{\theta}_i^{R'}(\boldsymbol{\alpha}_1^*-\boldsymbol{\alpha}_2^*)+\mathbf{z}_i'\boldsymbol{\beta}_1-\mathbf{z}_i'\boldsymbol{\beta}_2\})}, i\in\mathcal{J}_m. \tag{4.1}$$

The probability for high-grade cancer can be calculated similarly. Let $\{\delta_i^o, i \in \mathcal{J}_o\}$ be the set of observed censoring status, where $\mathcal{J}_o$ denotes the set of subjects with observed causes or censored. We let $\delta_i^* = \delta_i^m$ if $i \in \mathcal{J}_m$ and $\delta_i^* = \delta_i^o$ if $i \in \mathcal{J}_o$, constituting the complete set of censoring status $\{\delta_i^*, i = 1,\ldots,n\}$. With the imputed censoring status, we sample the model parameters from the posterior distribution derived from the augmented likelihood. Particularly, the distribution of the survival time in the augmented likelihood is given by

$$f(t_i \mid \delta_i^*, \boldsymbol{\theta}_i^R, \boldsymbol{\varphi}_2, \mathbf{z}_i) = \left[\prod_{k=1}^{2}\{h_{k0}(t_i)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R+\mathbf{z}_i'\boldsymbol{\beta}_k)\}^{I(\delta_i^*=k)}\right] \times\exp\left(-\sum_{k=1}^{2}H_{k0}(t_i)\exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R+\mathbf{z}_i'\boldsymbol{\beta}_k)\right). \tag{4.2}$$

The other components of the augmented likelihood remain the same as defined in (3.10). In the augmented data likelihood, the longitudinal and survival data are independent to each other conditional on the random effects. Thus, the posterior distributions of the longitudinal and survival sub-model parameters depend on the random effects, which are unobserved. We sample the random effects $\boldsymbol{\theta}^R$ from the posterior distribution and conditional on the sampled random effects, the survival and longitudinal sub-model parameters are updated in MCMC sampling. Full details regarding the prior and MCMC sampling are presented in the Supplementary Materials (http://www.statmod.org/smij/archive.html). We compute and report the posterior means, the posterior standard deviations and the 95% highest posterior density (HPD) intervals of the model parameters. The convergence of MCMC sampling is investigated via the trace plots of MCMC samples.

4.1 Model assessment: DIC_{Surv}, $\Delta\text{DIC}_{\text{Surv}}$, $\text{WAIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$

DIC_{Surv}: The (DIC; Spiegelhalter et al., 2002) has been widely used as a Bayesian model assessment measure. In this article, we are mainly interested in the model assessment for the competing risks data and to investigate how much the performance of the model is improved

by the inclusion of longitudinal data into the survival model. The DIC for the survival sub-model is defined as

$$\mathrm{DIC}_{\mathrm{Surv}} = \mathrm{Dev}_{\mathrm{Surv}}\Big(E(\boldsymbol{\varphi}_2 \mid D_{\mathrm{Surv,obs}}), E(\boldsymbol{\theta}^R \mid D_{\mathrm{Surv,obs}})\Big) + 2P_{D[\mathrm{Surv}]},$$

where $P_{D[\mathrm{Surv}]} = E(\mathrm{Dev}_{\mathrm{Surv}}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R | D_{\mathrm{Surv,obs}})) - \mathrm{Dev}_{\mathrm{Surv}}(E(\boldsymbol{\varphi}_2 | D_{\mathrm{Surv,obs}}), E(\boldsymbol{\theta}^R | D_{\mathrm{Surv,obs}}))$ is the effective number of model parameters and the deviance function is defined as

$$\mathrm{Dev}_{\mathrm{Surv}}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R \mid D_{\mathrm{Surv,obs}}) = -2\log L(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R \mid D_{\mathrm{Surv,obs}}).$$

In this article, the random effects $\boldsymbol{\theta}^R$ are considered as unknown parameters and the DIC could be influenced by the inappropriate posterior estimate of these unknown parameters. Following Huang et al. (2005), we extend the standard DIC by considering a linear combination of the parameters which is defined as

$$\mathrm{DIC}_{\mathrm{Surv}} = \mathrm{Dev}_{\mathrm{Surv}}\Big(E(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\mathrm{Surv,obs}})\Big) + 2P_{D[\mathrm{Surv}]}, \tag{4.3}$$

where $\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\alpha}_1^{*\prime}\boldsymbol{\theta}^R, \boldsymbol{\alpha}_2^{*\prime}\boldsymbol{\theta}^R, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, the effective number of parameters,

$$P_{D[\mathrm{Surv}]} = E\Big(\mathrm{Dev}_{\mathrm{Surv}}\Big(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\mathrm{Surv,obs}}\Big)\Big) - \mathrm{Dev}_{\mathrm{Surv}}\Big(E(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\mathrm{Surv,obs}})\Big),$$

and the deviance function

$$\mathrm{Dev}_{\mathrm{Surv}}\Big(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\mathrm{Surv,obs}}\Big) = -2\log L(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\mathrm{Surv,obs}}).$$

$\Delta\mathrm{DIC}_{\mathrm{Surv}}$: Another important aspect of the model assessment is to measure the gain in fit due to the inclusion of longitudinal data into the survival model. For this, we fit the survival data alone, that is, considering $\boldsymbol{\alpha}_1^* = \boldsymbol{\alpha}_2^* = 0$ and the cause-specific hazard function takes the form

$$h_k(t \mid h_{k0}, \boldsymbol{\alpha}_k^* = 0, \boldsymbol{\beta}_k, \mathbf{z}_i) = h_{k0}(t)\exp(\mathbf{z}_i'\boldsymbol{\beta}_k).$$

The DIC for the survival sub-model with the above consideration

$$\mathrm{DIC}_{\mathrm{Surv},0} = \mathrm{Dev}_{\mathrm{Surv},0}\big(E(\boldsymbol{\varphi}_2^* \mid D_{\mathrm{Surv,obs}})\big) + 2P_{D[\mathrm{Surv},0]}, \tag{4.4}$$

where $\boldsymbol{\varphi}_2^* = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $\mathrm{Dev}_{\mathrm{Surv},0}(\boldsymbol{\varphi}_2^* \mid D_{\mathrm{Surv,obs}}) = -2\log L(\boldsymbol{\varphi}_2^* \mid D_{\mathrm{Surv,obs}})$ and

$$P_{D[\mathrm{Surv},0]} = E\big(\mathrm{Dev}_{\mathrm{Surv},0}(\boldsymbol{\varphi}_2^* \mid D_{\mathrm{Surv,obs}})\big) - \mathrm{Dev}_{\mathrm{Surv},0}\big(E(\boldsymbol{\varphi}_2^* \mid D_{\mathrm{Surv,obs}})\big).$$

The following assessment criterion is defined

$$\Delta\text{DIC}_{\text{Surv}} = \text{DIC}_{\text{Surv},0} - \text{DIC}_{\text{Surv}}, \tag{4.5}$$

which quantifies the gain in the fit in the competing risks survival sub-model due to the inclusion of longitudinal data penalizing the additional parameters in the competing survival sub-model.

$\text{WAIC}_{\text{Surv}}$: The WAIC (Watanabe, 2010) is another known Bayesian criterion for model assessment. The WAIC computes the logarithm of the pointwise posterior predictive density (LPPD) and adds the penalty term to adjust for overfitting (Gelman et al., 2014). LPPD for the survival sub-model is defined as

$$\text{LPPD}_{\text{Surv}} = \log E\Big[L(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\text{Surv,obs}})\Big].$$

The effective number of parameters is evaluated as

$$P_{\text{WAIC[Surv]}} = 2\Big(\log E\Big[L(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\text{Surv,obs}})\Big] - E\Big[\log L(\mathbf{g}(\boldsymbol{\varphi}_2, \boldsymbol{\theta}^R) \mid D_{\text{Surv,obs}})\Big]\Big).$$

Finally, the $\text{WAIC}_{\text{Surv}}$ is defined as

$$\text{WAIC}_{\text{Surv}} = -2(\text{LPPD}_{\text{Surv}} - P_{\text{WAIC[Surv]}}). \tag{4.6}$$

$\Delta\text{WAIC}_{\text{Surv}}$: Similar to the formulation of $\Delta\text{DIC}_{\text{Surv}}$, we first define $\text{WAIC}_{\text{Surv},0}$, which is defined on the parameter vector $\boldsymbol{\varphi}_2^* = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and setting $\boldsymbol{\alpha}_1^* = \boldsymbol{\alpha}_2^* = 0$. The $\text{WAIC}_{\text{Surv},0}$ is given by

$$\text{WAIC}_{\text{Surv},0} = -2(\text{LPPD}_{\text{Surv},0} - P_{\text{WAIC[Surv},0]}), \tag{4.7}$$

where $\text{LPPD}_{\text{Surv},0} = \log E\big[L(\boldsymbol{\varphi}_2^* \mid D_{\text{Surv,obs}})\big]$ and the effective number of parameter, $P_{\text{WAIC[Surv},0]} = 2\big(\log E\big[L(\boldsymbol{\varphi}_2^* \mid D_{\text{Surv,obs}})\big] - E\big[\log L(\boldsymbol{\varphi}_2^* \mid D_{\text{Surv,obs}})\big]\big)$. Finally, the $\Delta\text{WAIC}_{\text{Surv}}$ is given by

$$\Delta\text{WAIC}_{\text{Surv}} = \text{WAIC}_{\text{Surv},0} - \text{WAIC}_{\text{Surv}}. \tag{4.8}$$

The interpretation of the $\Delta\text{WAIC}_{\text{Surv}}$ is similar to $\Delta\text{DIC}_{\text{Surv}}$, which again quantifies the gain in fit due to the inclusion of longitudinal outcome in the survival sub-model.

5 A simulation study

5.1 Simulation design

We carry out a simulation study to examine the empirical performance of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ under different values of the association parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. We define the sub-models and specify the design values of the parameters using the same notations

introduced in Sections 3 and 4. The longitudinal measurements are generated from the following sub-model

$$y_i(a_{ij}) = \mathbf{g}(a_{ij})'(\boldsymbol{\theta} + \Gamma\boldsymbol{\theta}_i^R) + \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i(a_{ij}), \tag{5.1}$$

where $\mathbf{g}(a_{ij})' = (1, a_{ij})$, $\mathbf{x}_i = (x_{1i}, x_{2i})$ with $x_{1i} \sim \mathcal{N}(0, 1)$ and $x_{2i}|x_{1i} \sim \mathcal{N}(0.2x_{1i}, 1)$, $\epsilon_i(a_{ij}) \sim \mathcal{N}(0, \sigma^2)$, and $\boldsymbol{\theta}_i^R = (\theta_{0i}^R, \theta_{1i}^R) \sim N(0, \mathbf{I}_2)$. The design values of the parameters are given as $\boldsymbol{\theta} = (\theta_0, \theta_1)' = (0.5, 1)'$, $\Gamma = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.8 \end{pmatrix}$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)' = (0.15, 0.3)'$ and $\sigma^2 = 0.5$.

For the survival sub-model, the same set of covariates is used as for the longitudinal sub-model and set $\mathbf{z}_i = \mathbf{x}_i$. The exponential distribution for the baseline hazard functions is considered due to two different causes

$$h_{k0}(t) = \lambda_{k0},\ k = 1, 2, t \in \mathbf{R}^+.$$

Thus the survival sub-model for the kth cause of failure becomes

$$h_k(t \mid \lambda_{k0}, \boldsymbol{\alpha}_k^*, \boldsymbol{\theta}_i^R, \boldsymbol{\beta}_k, \mathbf{z}_i) = \lambda_{k0} \exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k),\ k = 1, 2\,. \tag{5.2}$$

We generate the true event time due to the kth cause of failure as follows

$$t_{ik} = \left[-\lambda_{k0} \exp(\boldsymbol{\alpha}_k^{*'}\boldsymbol{\theta}_i^R + \mathbf{z}_i'\boldsymbol{\beta}_k)\right]^{-1} \log(1 - U),\ k = 1, 2,$$

where $U \sim \mathcal{U}(0, 1)$. The censoring times c_i are generated from an exponential distribution with mean 25. Then the observed event time for the ith subject is computed as $t_i = \min(t_{i1}, t_{i2}, c_i)$. The censoring status for the ith subject is evaluated as $\delta_i = k$ if $I(t_{ik} \leq c_i)$ for $k = 1, 2$ and 0 otherwise. To simulate the unknown causes of failure, for each uncensored case, we first generate $v \sim \mathcal{U}(0, 1)$, and the cause of failure is then set to be missing if $v < 0.1$.

The design values of the survival sub-model parameters are given as $\boldsymbol{\beta}_1 = (0.5, 0.6)'$, $\boldsymbol{\beta}_2 = (0.2, -0.5)'$, $\lambda_{10} = 0.1$ and $\lambda_{20} = 0.08$. For the longitudinal data, we consider a balanced study design with $m_i = 20$ for $i = 1, \ldots, n$. The observation times of the longitudinal measurements for the ith subject are evaluated at $a_{ij} = 21\ (j - 1)/365$, $j = 1, \ldots, m_i$. We generate the datasets for the following two scenarios with all other parameter values specified as above:

1. True-SPM: In this scenario, we generate the data from the proposed share parameter joint model. In particular, we generate the longitudinal data from the model (5.1) and the survival data from the model (5.2) with $\alpha_{11}^* = 0.2$, $\alpha_{12}^* = 0.8$, $\alpha_{21}^* = 0.5$ and $\alpha_{22}^* = 1$.

2. True-Surv: In this scenario, the data are generated from the survival data only model by specifying $\alpha^*_{k1} = \alpha^*_{k2} = 0$ for $k = 1, 2$. Thus, the survival sub-model defined in (5.2) becomes

$$h_k(t \mid \lambda_{k0}, \boldsymbol{\beta}_k, \mathbf{z}_i) = \lambda_{k0} \exp(\mathbf{z}_i' \boldsymbol{\beta}_k), \ k = 1, 2 . \tag{5.3}$$

5.2 Simulation results

We generate two hundred simulated datasets independently with a sample size (n) of 6 000 under each of the scenarios discussed above. The censoring percentages of True-SPM and True-Surv scenarios are about 20% and 17%, respectively. For each of the two scenarios, we fit the data using the proposed joint model and the survival data only model to evaluate $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$. We take 2 000 burn-in iteration and 5 000 Gibbs samples with five thinned steps for each simulated dataset. For the True-SPM scenario, the simulation results under the proposed joint model and the survival data only model are presented in Tables 2 and 3, respectively. Table 2 shows that the joint model analysis of the simulated data under the True-SPM scenario leads to unbiased and efficient parameter estimates with high coverage probabilities. We also analyse the same simulated data under the True-SPM scenario by the survival data only model and the results are presented in Table 3. Table 3 shows that the biases of the model parameters are higher compared to the joint model analysis and the coverage probabilities are substantially lower. Particularly, the coverage probabilities for λ_1 and λ_2 are zeros. The Bayesian criteria $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ are evaluated for each simulated data under the True-SPM scenario.

We repeat a similar analysis for the simulated data under the True-Surv scenario. To preserve the space, we present the simulation results under the proposed joint model and the survival data only model in Tables S.1 and S.2, respectively, in the Supplementary Materials (http://www.statmod.org/smij/archive.html). We also evaluate the $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ for each simulated dataset under the True-Surv scenario to compare with the True-SPM scenario. Table 4 presents the summary statistics of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ for the True-Surv and True-SPM scenarios. Under the True-Surv scenario, the median values of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ are −4.771 and −4.669, respectively, while under True-SPM scenario, the median values of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ are 3793.524 and 3811.183, respectively.

The boxplots of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ are shown in Figures 3 and 4, respectively, for the True-Surv and the True-SPM. Figures 3 and 4 show that when there exists no association between the longitudinal and survival data, the gain in fit due to the inclusion of longitudinal data in the survival sub-model is not substantial. In the True-Surv scenario, the inter quartile ranges (IQRs) of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ in Table 4 include the value of −4. This indicates that DIC_{Surv} (4.3) and $\text{WAIC}_{\text{Surv}}$ (4.6) penalize more due to the inclusion of four association parameters ($\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$) compared to $\text{DIC}_{\text{Surv},0}$ (4.4) and $\text{WAIC}_{\text{Surv},0}$ (4.7). However, the boxplots of $\Delta\text{DIC}_{\text{Surv}}$ and $\Delta\text{WAIC}_{\text{Surv}}$ in Figures 3 and 4, respectively, under the True-SPM scenario show a substantial gain in fit. These figures clearly illustrate that our proposed joint

model performs better than the survival data only model when there exists an association between the longitudinal and survival data.

6 Analysis of SELECT data

We carry out an analysis of the SELECT data. The response variable is the time to diagnosis of PC due to low-grade or high-grade. In this analysis, we consider seven covariates, including patients age (in years), BMI in kg/m^2, Hispanic status ('Yes' = 1, 'No' = 0), Race ('African American' = 1, 'Others' = 0), Family history of cancer ('Yes' = 1, 'No' = 0), current smoking status ('Yes' = 1, 'No' = 0) and former smoking status ('Yes' = 1, 'No' = 0). Table 5 shows the summary statistics of the continuous variables. These summary statistics show that the average age of high-grade patients is higher compared to the other groups. The average BMI of high-grade patients is also found to be higher. Table 6 shows the frequency distribution of the patients with respect to the censoring status. Among the patients who developed cancer, missing-grade patients show a higher percentage of Hispanic group compared to the others. From the percentage distribution, we observe that there is no indication of a positive association between smoking status and cancer status. For all three groups of PC, the percentage of family history of cancer was found to be higher compared to the censored group. A similar trend is observed among the African American group indicating a positive association between race and cancer status.

The main goal of this study is to assess the association between the longitudinal PSA and the time-to-PC due to low-grade or high-grade after adjusting for the covariates under consideration. For the longitudinal sub-model, we consider the natural logarithm of PSA as the response variable. For both the longitudinal and survival data, the same seven covariates are used. We consider the shared parameter joint model, in which the random effects are assumed for the longitudinal and survival data. We consider the piecewise constant forms for the baseline hazards functions due to two different causes. To construct the appropriate partitions $\{s_{kl}, l = 1,\ldots, G_k, k = 1, 2\}$, we use $\mathrm{DIC}_{\mathrm{Surv},0}$ defined in (4.4) to determine the best combination of (G_1, G_2). Table 7 shows the results of $\mathrm{DIC}_{\mathrm{Surv},0}$ for different combinations of (G_1, G_2) suggesting the piecewise constant hazard function with $(G_1 = 70, G_2 = 70)$ fits the SELECT data best. For the analysis, we used 5 000 burn-ins and 10 000 MCMC samples.

The posterior estimates of $\boldsymbol{\gamma}$ along with the 95% HPD intervals are presented in Figure 5. The figure suggests that except Hispanic status and race of the patients, all other covariates have significant effects on the longitudinal PSA. In particular, we see that BMI is negatively associated with the PSA, which is consistent with other studies (Hekal and Ibrahiem, 2010). Some studies (Algotar et al., 2011) found there is a positive association between smoking status and PSA. However, the sample size for that study was very small (140 subjects). In our analysis of the SELECT data, we find that smoking status is negatively associated with the PSA. The estimates and the 95% HPD intervals of the association parameters $\{\boldsymbol{\alpha}_k, k = 1, 2\}$ are presented in Table 8. The results show that both intercept and slope parameters of the longitudinal PSA in the shared parameter joint model have significant effects on both the low-grade cancer and the high-grade cancer. The positive values of the association parameters indicate that there is a positive association between the PSA and the risk of developing low-grade or high-grade cancer. Comparing the risks of the low-grade and high-

grade patients, both the intercept parameter and the slope parameter for high-grade are higher compared to those low-grade patients. This indicates that with the increase in the PSA, the risk of developing high-grade cancer is slightly higher compared to the risk of low-grade cancer.

The survival sub-model parameter estimates under the joint model and the survival data only model are presented in Table 9. The results show that age is positively associated with the risk of high-grade cancer and not significantly associated with low-grade cancer, which is found to be consistent under both the joint model and the survival data only model. Both joint model and he survival data only model show that BMI is not associated with the risk of low-grade cancer. Although BMI is significantly positively associated with high-grade cancer under the survival data only model, BMI is not significantly associated with the risk of high-grade cancer under the joint model. Hispanic patients were found to have a lower risk of developing low-grade cancer compared to non-Hispanic patients and Hispanic status is found to be not associated with the risk of developing high-grade cancer. Although race of the patients is a significant factor for the cancer development under the survival data only model, race is not a significant factor associated with the risk of developing low-grade and high-grade cancer under the joint model. We also present the estimates and the 95% HPD intervals of $\alpha_{21} - \alpha_{11}$ and $\alpha_{22} - \alpha_{12}$ in Table 8. The results in Table 8 indicate that the risk of developing PC due to low grade is higher than high grade at the baseline. The difference in the slope parameters indicates that the risk of developing high grade cancer increases at a faster rate compared to low grade cancer due to increase in PSA.

The family history of cancer has been an important risk factor associated with the development of low-grade and high-grade cancer. The patients who have family history of cancer tend to have a higher risk of developing low-grade and high-grade cancer compared to the patients who do not have any family history of cancer. This association is observed under both the joint model and the survival data only model. The patients who currently smoke have a lower risk of developing both low-grade and high-grade cancer under the joint model, although under the survival data only model, currently smoking status is not associated with low-grade cancer. The patients who used to smoke before have a lower risk of developing low-grade cancer, however, they do not have a higher risk of high-grade cancer. Finally, we present the Bayesian DIC and WAIC criteria to assess the fit of the models. DIC and WAIC under the survival data only model are $DIC_{surv,0} = 19082.667$ and $WAIC_{Surv,0} = 19118.526$. The $DIC_{Surv} = 14915.091$ and $WAIC_{Surv} = 15207.596$ under the joint model, which are much lower than the $DIC_{Surv,0}$ and $WAIC_{Surv,0}$ under the survival data only model. To quantify the gain in fit with the inclusion of longitudinal data in the survival sub-model, we also defined ΔDIC_{Surv} in (4.5). We found $\Delta DIC_{Surv} = 4167.5768$, which indicates that we have about 21% gain in fit through the inclusion of the longitudinal PSA in the survival sub-model. In addition to ΔDIC_{Surv}, we also present the results of $\Delta WAIC_{Surv}$ in Table 9. The $\Delta WAIC_{Surv} = 3910.930$ indicates that through the shared parameter joint model analysis, we get approximately 20% gain in fit due to the inclusion of longitudinal data compared to the survival data only model. The convergence of the model parameters is checked via the trace plots of MCMC samples of the model parameters. These figures are given in the Supplementary Materials (http://www.statmod.org/smij/archive.html).

7 Discussion

In this article, a joint model for the longitudinal and competing risks survival data is proposed with an applications to SELECT data. Our proposed model is flexible to account for the uncensored subjects with missing causes of failure. We develop a novel reparameterization to facilitate an efficient and convenient implementation of the MCMC sampling algorithm to sample from the posterior distribution. The proposed model is applied to SELECT data and to assess the model fit, we introduce four model assessment criteria, DIC_{Surv}, ΔDIC_{Surv}, $WAIC_{Surv}$ and $\Delta WAIC_{Surv}$. The Bayesian criteria ΔDIC_{Surv} and $\Delta WAIC_{Surv}$ measure the gain in fit of the survival sub-model due to the inclusion of longitudinal data. A simulation study is conducted to examine the change in ΔDIC_{Surv} and $\Delta WAIC_{Surv}$ for different values of the association parameters. Our simulation study shows that when there exists an association between the longitudinal and survival data, our proposed joint model fits the data better compared to the survival data only model. The analysis of the SELECT data shows that there is a significant positive association with PSA and the risk of developing low-grade and high-grade cancer. The performance of the joint model is assessed by ΔDIC_{Surv} and ΔWIC_{Surv} that show about the 21% and 20% gain in fit compared to the model with survival data only, respectively. Computer code was written for the FORTRAN 95 compiler, and we used IMSL subroutines with double precision accuracy. The FORTRAN code is available from the authors upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors wish to thank the editors and two referees for the insightful comments and suggestions, which helped to improve the article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by US NIH grant #GM 70335.

References

Algotar AM, Stratton SP, Ranger-Moore J, Stratton MS, Hsu C-H, Ahmann FR, Nagle RB and Thompson PA (2011) Association of obesity and smoking with PSA and PSA velocity in men with prostate cancer. American Journal of Men's Health, 5, 272–78.

Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 34, 187–202.

Elashoff RM, Li G and Li N (2007) An approach to joint analysis of longitudinal measurements and competing risks failure time data. Statistics in Medicine, 6, 2813–35.

Elashoff RM, Li G and Li N (2008) A joint model for longitudinal measurements and survival data in the presence of multiple failure types. Biometrics, 64, 762–71. doi:10.1111/j.1541-0420.2007.00952.x. [PubMed: 18162112]

Faucett CL and Thomas DC (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. Statistics in Medicine, 15, 1663–85. [PubMed: 8858789]

Ferrer L, Rondeau V, Dignam J, Pickles T, Jacqmin-Gadda H and Proust-Lima C (2016). Joint modelling of longitudinal and multi-state processes: Application to clinical progressions in prostate cancer. Statistics in Medicine, 35, 3933–48. [PubMed: 27090611]

Gao G and Tsiatis AA (2005) Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. Biometrika, 92, 875–91.

Gelman A, Hwang J and Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24, 997–1016.

Gilks WR and Wild P (1992) Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41, 337–48.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97–109.

Hekal I and Ibrahiem E (2010). Obesity-PSA relationship: A new formula. Prostate Cancer and Prostatic Diseases, 13, 186. [PubMed: 20029402]

Henderson R, Diggle P and Dobson A (2000). Joint modelling of longitudinal measurements and event time data. Biostatistics, 1, 465–80. [PubMed: 12933568]

Hu W, Li G and Li N (2009) A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. Statistics in Medicine, 28, 1601–19. [PubMed: 19308919]

Huang L, Chen M-H and Ibrahim JG (2005) Bayesian analysis for generalized linear models with nonignorably missing covariates. Biometrics, 61, 767–80. [PubMed: 16135028]

Huang X, Li G, Elasho RM and Pan J (2011) A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. Lifetime Data Analysis, 17, 80–100. [PubMed: 20549344]

Ibrahim JG, Chen M-H and Sinha D (2001). Bayesian Survival Analysis. NewYork, NY: Springer.

Laird NM and Ware JH (1982) Random-effects models for longitudinal data. Biometrics, 38, 963–74. [PubMed: 7168798]

Lu K and Tsiatis AA (2001) Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics, 57, 1191–97. [PubMed: 11764260]

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21, 1087–92.

Prentice RL (1982) Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika, 69, 331–342.

Rizopoulos D (2012) Joint models for longitudinal and time-to-event data: With applications in R. Boca Raton, FL: Chapman and Hall/CRC.

Spiegelhalter DJ, Best NG, Carlin BP and Van Der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64, 583–639.

Tsiatis A, Degruttola V and Wulfsohn M (1995) 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.

Verbeke G and Molenberghs G (2009). Linear Mixed Models for Longitudinal Data. Berlin: Springer Science & Business Media.

Watanabe S (2010). Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11, 3571–94.

Wulfsohn MS and Tsiatis AA (1997) A joint model for survival and longitudinal data measured with error. Biometrics, 53, 330–39. [PubMed: 9147598]

Zhang D, Chen M-H, Ibrahim JG, Boye ME and Shen W (2016) JMFit: A SAS macro for joint models of longitudinal and survival data. Journal of Statistical Software, 71. doi: 10.18637/jss.v071.i03.

——— (2017) Bayesian model assessment in joint modeling of longitudinal and survival data with applications to cancer clinical trials. Journal of Computational and Graphical Statistics, 26, 121–33. [PubMed: 28239247]

Zhang D, Chen M-H, Ibrahim JG, Boye ME, Wang P and Shen W (2014) Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. Statistics in Medicine, 33, 4715–33. [PubMed: 25044061]

Zhang F, Chen MH, Cong X and Chen Q (2019) A Bayesian joint modeling approach of longitudinal and survival data with semicompeting risks (Technical report). Department of Statistics, University of Connecticut, Mansfield, Connecticut.

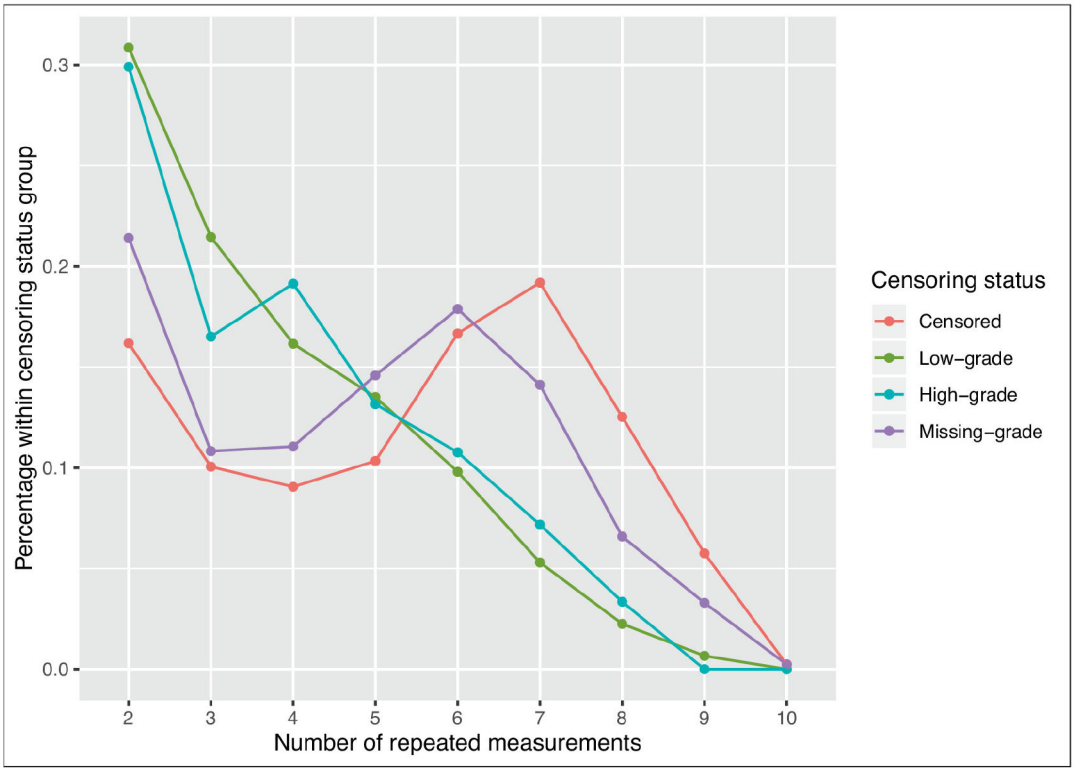

Figure 1:
Percentage distribution of PSA repeated measurements for different censoring status

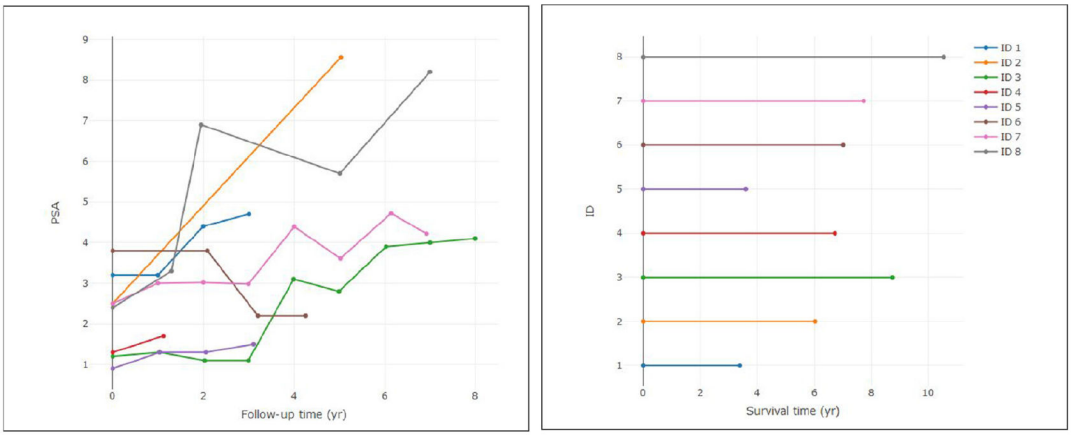

Figure 2:
PSA trajectories and survival times for eight random patients

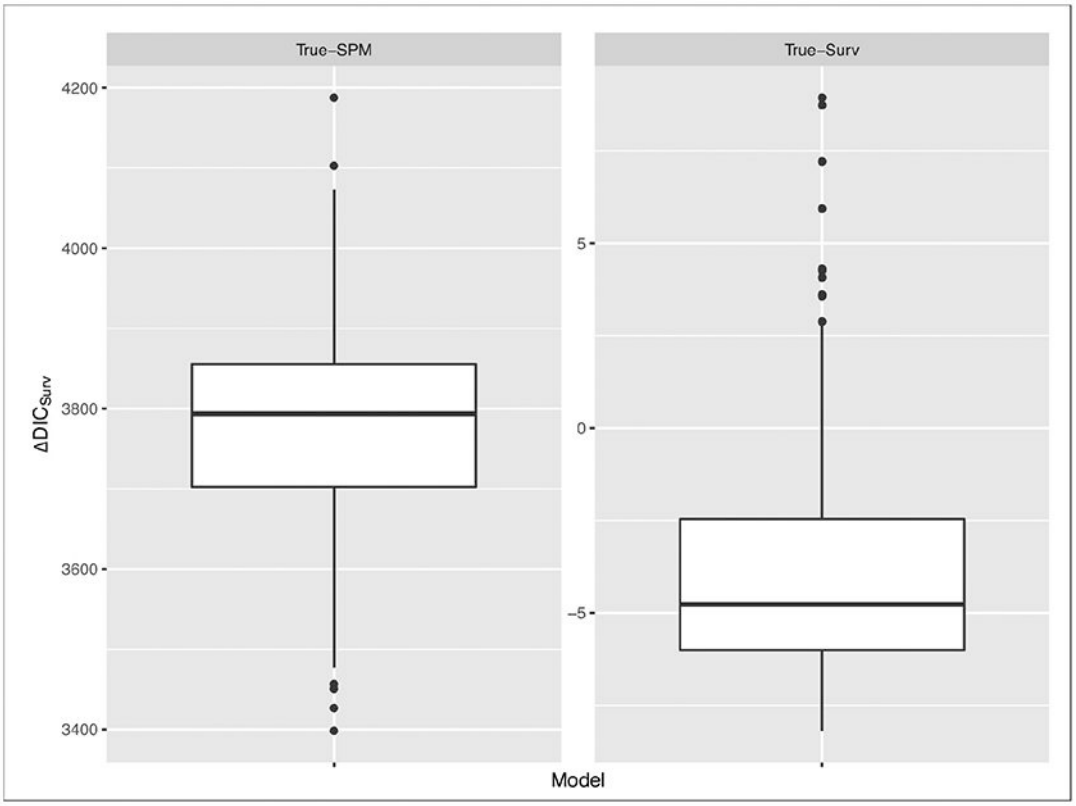

Figure 3:
Boxplots of DIC under True-SPM and True-Surv model

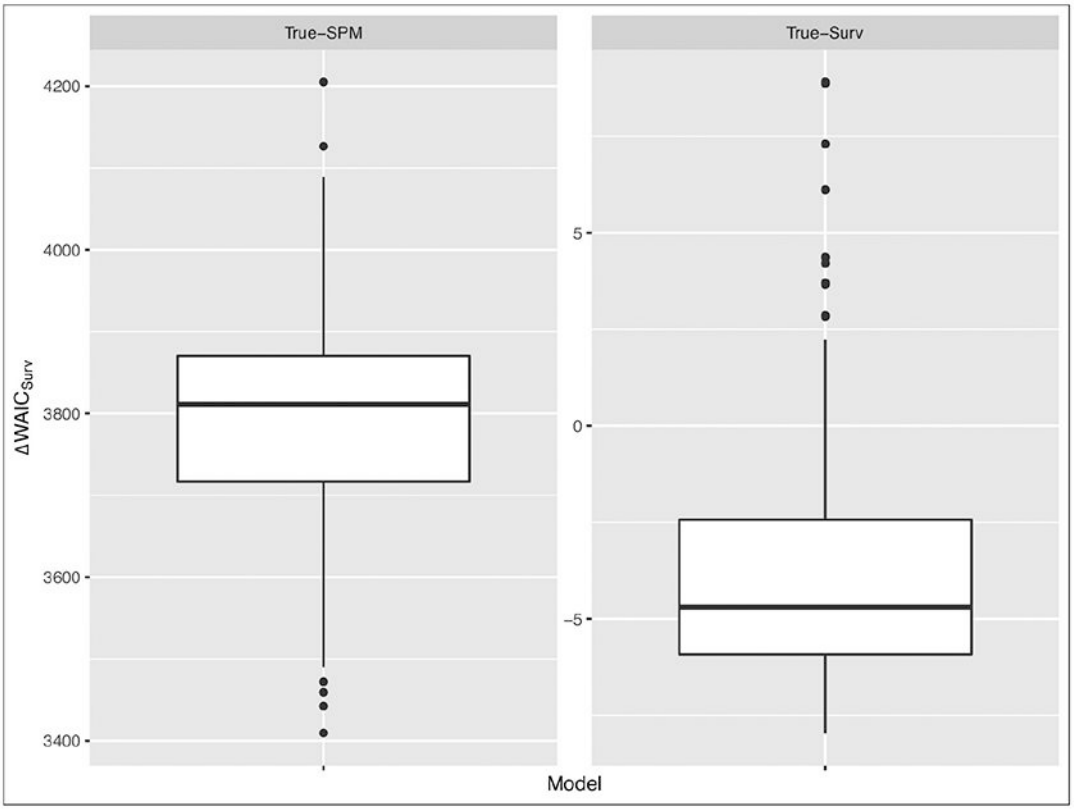

Figure 4:
Boxplots of WAIC under True-SPM and True-Surv model

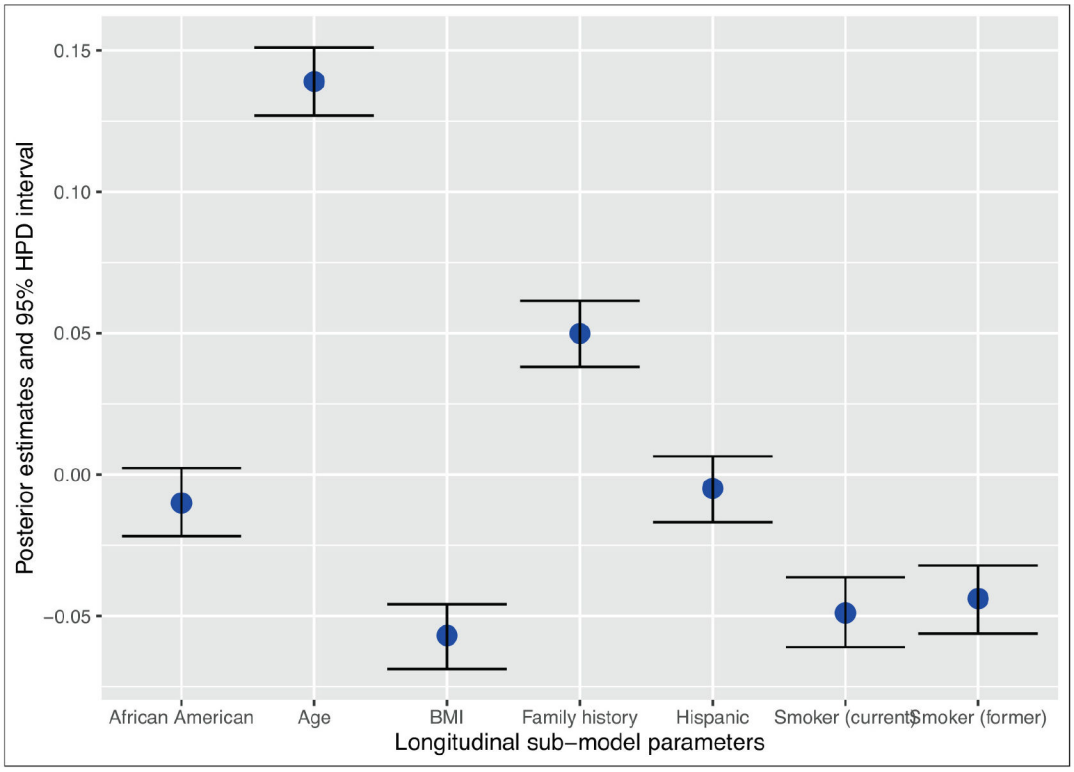

Figure 5.
95% HPD interval of the $\boldsymbol{\gamma}$

Table 1:

Distribution of cancer status by cancer grade

PC	Frequency (%)
No	21 194 (92.99)
Yes	1 598 (7.01)
Low-grade	755 (47.25)
High-grade	418 (26.16)
Missing-grade	425 (26.6)

Table 2:

Posterior estimates under the joint model for True-SPM in Section 5.1

Parameter	True	Est.	Bias	SE	SD	Coverage
γ_1	0.15	0.150	−0.000	0.011	0.010	0.970
γ_2	0.30	0.302	−0.002	0.011	0.012	0.930
σ^2	0.50	0.500	0.000	0.002	0.002	0.910
b_{11}	1.00	1.000	0.000	0.009	0.009	0.970
b_{21}	0.50	0.500	0.000	0.012	0.012	0.925
b_{22}	0.80	0.799	0.001	0.008	0.008	0.950
θ_1	0.50	0.496	0.004	0.013	0.014	0.905
θ_2	1.00	0.990	0.010	0.012	0.013	0.875
α^*_{11}	0.20	0.202	−0.002	0.023	0.022	0.950
α^*_{12}	0.80	0.800	0.000	0.023	0.022	0.950
α^*_{21}	0.50	0.494	0.006	0.027	0.029	0.925
α^*_{22}	1.00	0.997	0.003	0.026	0.023	0.975
β_{11}	0.50	0.498	0.002	0.021	0.021	0.940
β_{12}	0.60	0.587	0.013	0.023	0.024	0.890
β_{21}	0.20	0.204	−0.004	0.023	0.022	0.960
β_{22}	−0.50	−0.484	−0.016	0.025	0.022	0.950
λ_1	0.10	0.099	0.001	0.002	0.002	0.940
λ_2	0.08	0.080	−0.000	0.002	0.002	0.955

Table 3:

Posterior estimates under the survival data only model for True-SPM in Section 5.1

Parameter	True	Est.	Bias	SE	SD	Coverage
β_{11}	0.50	0.468	0.032	0.021	0.026	0.650
β_{12}	0.60	0.594	0.006	0.022	0.029	0.845
β_{21}	0.20	0.172	0.028	0.023	0.028	0.705
β_{22}	−0.50	−0.436	−0.064	0.024	0.029	0.310
λ_1	0.10	0.079	0.021	0.002	0.002	0.000
λ_2	0.08	0.067	0.013	0.002	0.002	0.000

Table 4

Summary of ΔDIC_{Surv} and $\Delta WAIC_{Surv}$ for True-SPM and True-Surv

	True-Surv		True-SPM	
Criterion	**Median**	**IQR=(Q_1, Q_3)**	**Median**	**IQR=(Q_1, Q_3)**
ΔDIC_{Surv}	−4.771	(−6.004, −2.460)	3 793.524	(3 702.444, 3 855.218)
$\Delta WAIC_{Surv}$	−4.699	(−5.923, −2.432)	3 811.183	(3 716.918, 3 870.227)

Table 5:

Summary statistics of continuous covariates at baseline by PC

	Mean (SD)			
	Censored	**Low-grade**	**High-grade**	**Missing-grade**
Age	62.94 (6.72)	62.95 (6.03)	64.58 (6.24)	62.59 (7.39)
Height	176.64 (7.44)	176.84 (7.29)	176.83 (7.17)	177.09 (7.39)
Weight	89.19 (15.57)	88.11 (14.85)	90.92 (16.23)	89.41 (14.45)
BMI	28.56 (4.55)	28.12 (4.06)	29.08 (4.91)	28.48 (4.15)

Table 6:

Summary statistics of categorical covariates by PC

	Frequency (%)			
	Censored	**Low grade**	**High grade**	**Missing grade**
Hispanic				
No	20 124 (94.95)	738 (97.75)	405 (96.89)	400 (94.12)
Yes	1 070 (5.05)	17 (2.25)	13 (3.11)	25 (5.88)
Smoking				
Current	1 681 (7.93)	51 (6.75)	19 (4.55)	23 (5.41)
Former	10 499 (49.54)	344 (45.56)	193 (46.17)	203 (47.76)
Never	9 014 (42.53)	360 (47.68)	206 (49.28)	199 (46.82)
Family history				
No	17 805 (84.01)	533 (70.60)	304 (72.73)	316 (74.35)
Yes	3 389 (15.99)	222 (29.40)	114 (27.27)	109 (25.65)
Race				
African American	2 887 (13.62)	108 (14.30)	61 (14.59)	76 (17.88)
Others	18 307 (86.38)	647 (85.70)	357 (85.41)	349 (82.12)

Table 7

$DIC_{Surv,0}$ and dimension penalty for different combinations of G_1 and G_2

G1	G2	$DIC_{Surv,0}$	$P_{D[Surv,0]}$
60	60	19 137.72	136.06
60	70	19 112.95	146.68
60	80	19 123.86	157.33
60	90	19 137.66	167.99
70	60	19 099.99	146.37
70	70	19 082.67	156.77
70	80	19 094.20	167.72
70	90	19 107.96	178.36
80	60	19 110.30	156.70
80	70	19 093.18	167.19
80	80	19 104.48	178.05
80	90	19 118.93	189.04

Table 8

Posterior estimates and 95% HPD intervals of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ under the joint model with G_1=70 and G_2 = 70

	Est.	SD	95% HPD interval
α_{11}	1.467	0.057	(1.454, 1.460)
α_{12}	4.333	0.295	(4.291, 4.326)
α_{21}	1.387	0.075	(1.371, 1.380)
α_{22}	6.169	0.349	(6.153, 6.192)
$\alpha_{21} - \alpha_{11}$	−0.080	0.096	(−0.108, −0.097)
$\alpha_{22} - \alpha_{12}$	1.836	0.457	(1.845, 1.900)

Table 9

Posterior estimates, 95% HPD intervals, DIC_{Surv}, $DIC_{Surv,0}$, and ΔDIC_{Surv} under the joint model and the survival data only model with G_1=70 and G_2 = 70

	Joint model			Survival data only model		
Variable	**Est.**	**SD**	**95% HPD int.**	**Est.**	**SD**	**95% HPD int.**
Age	0.028	0.039	(−0.05, 0.103)	0.003	0.034	(−0.063, 0.071)
	0.254	0.05	(0.162, 0.359)	0.234	0.044	(0.15, 0.321)
BMI	−0.153	0.039	(−0.228, −0.075)	−0.096	0.035	(−0.162, −0.025)
	0.044	0.049	(−0.054, 0.138)	0.117	0.043	(0.03, 0.2)
Hispanic	−0.119	0.049	(−0.22, −0.028)	−0.085	0.045	(−0.173, 0)
	−0.073	0.061	(−0.194, 0.041)	−0.048	0.057	(−0.162, 0.058)
African	−0.043	0.039	(−0.12, 0.033)	0.07	0.034	(0.005, 0.138)
American	0.011	0.051	(−0.09, 0.11)	0.139	0.045	(0.052, 0.229)
Family	0.274	0.032	(0.211, 0.335)	0.259	0.027	(0.204, 0.312)
History	0.244	0.044	(0.157, 0.328)	0.231	0.038	(0.156, 0.306)
Currently	−0.086	0.042	(−0.165, −0.003)	−0.066	0.037	(−0.138, 0.008)
Smoker	−0.173	0.066	(−0.304, −0.049)	−0.16	0.063	(−0.286, −0.04)
Former	−0.081	0.039	(−0.155, −0.003)	−0.068	0.035	(−0.138, −0.002)
Smoker	−0.102	0.052	(−0.201, 0)	−0.103	0.047	(−0.196, −0.013)
DIC_{Surv}			14 915.091			
$DIC_{Surv,0}$						19 082.668
ΔDIC_{Surv}			4 167.577			
$WAIC_{Surv}$			15 207.596			
$WAIC_{Surv,0}$						19 118.526
$\Delta WAIC_{Surv}$			3 910.930			