Bayesian Semiparametric Estimation of Cancer-specific Age-at-onset Penetrance with Application to Li-Fraumeni Syndrome

Seung Jun Shin[1], **Ying Yuan**[2], **Louise C. Strong**[2], **Jasmina Bojadzieva**[2], **Wenyi Wang**[2]

[1]Korea University

[2]The University of Texas MD Anderson Cancer Center

Abstract

Penetrance, which plays a key role in genetic research, is defined as the proportion of individuals with the genetic variants (i.e., genotype) that cause a particular trait and who have clinical symptoms of the trait (i.e., phenotype). We propose a Bayesian semiparametric approach to estimate the cancer-specific age-at-onset penetrance in the presence of the competing risk of multiple cancers. We employ a Bayesian semiparametric competing risk model to model the duration until individuals in a high-risk group develop different cancers, and accommodate family data using family-wise likelihoods. We tackle the ascertainment bias arising when family data are collected through probands in a high-risk population in which disease cases are more likely to be observed. We apply the proposed method to a cohort of 186 families with Li-Fraumeni syndrome identified through probands with sarcoma treated at MD Anderson Cancer Center from 1944 to 1982.

Keywords

cancer specific age-at-onset penetrance; competing risk; gamma frailty model; family-wise likelihood; Li-Fraumeni syndrome

1 Introduction

The Li-Fraumeni syndrome (LFS) is a rare disorder that substantially increases the risk of developing several cancer types, particularly in children and young adults. It is characterized by autosomal dominant mutation inheritance with frequent occurrence of several cancer types: soft tissue/bone sarcoma, breast cancer, lung cancer, and other types of cancer that are grouped together as "other cancers" (Nichols et al.; 2001; Birch et al.; 2001). A majority of LFS is caused by germline mutations in the TP53 tumor suppressor gene (Malkin et al.; 1990; Srivastava et al.; 1990).

The LFS data that motivate our work are family data collected through patients diagnosed with pediatric sarcoma (i.e., probands) who were treated at MD Anderson Cancer Center from 1944 to 1982 and their extended kindred. The data consist of 186 families, with a total of 3686 subjects. The size of the families ranges from 3 to 717, with the median at 7. This dataset is the longest followed-up cohort in the world (followed up for 20–50 years), and among the largest collection of TP53 mutation carriers in all cohorts that are available for LFS. Considering the prevalence of TP53 mutations in a general population is as low as 0.0001 to 0.003, this dataset provides a specially enriched collection of TP53 mutations, which then allow us to characterize its effect on a diverse cancer outcomes. For each subject, the duration until he/she develops cancer is recorded as the primary endpoint. Although it is possible for LFS patients to experience multiple cancers during their lifetime, here, we focus on only the time to the first primary cancer since only a few patients represented in the database experienced multiple primary cancers. Table 1 shows the cancer-specific summaries for the LFS data. Further descriptions of the data are provided by Lustbader et al. (1992), Strong et al. (1992), and Hwang et al. (2003).

The primary objective here is to estimate the cancer-specific age-at-onset penetrance as a measure of the risk of experiencing a specific cancer for a person with a specific genotype(i.e., TP53 mutation status). Penetrance, which plays a crucial role in genetic research, is defined as the proportion of individuals with the genetic variants (i.e., genotype) that cause a particular trait who also have clinical symptoms of the trait (i.e., phenotype). When the clinical traits of interest are age-dependent (e.g., cancers), it is often more desirable to estimate the age-at-onset penetrance, defined as the probability of disease onset by a certain age, while adjusting for additional covariates if necessary. For the LFS study, the age-at-onset penetrance is defined as the conditional probability of having LFS-related cancers by a certain age given a certain TP53 mutation status. Cox proportional hazard regression models (Gauderman and Faucett; 1997; Wu et al.; 2010, among many others) have been most widely used for this task. Other approaches have included nonparametric estimation Wang et al. (2007) and parametric estimation based on logistic regression (Abel et al.; 1990) or a Weibull model (Hashemian et al.; 2009).

Estimating the age-at-onset penetrance for the LFS data is challenging for several reasons. First, LFS involves multiple types of cancer, and subjects have simultaneous competing risks of developing multiple types of cancer. Chatterjee et al. (2003) proposed a penetrance estimation method under a competing risk framework for a kin-cohort design. However, their method is not directly applicable if the pedigree size is large and/or there is additional genetic information from relatives. Gorfine and Hsu (2011) and Gorfine et al. (2014) proposed frailty-based competing risk models for family data, assuming that genotypes are completely observed for all family members, which is not the case for the LFS data.

Second, the genotype (i.e., TP53 mutation status) is not measured for the majority (about 74%) of subjects and the LFS data are clustered in the form of families. Accommodating the missing data and accounting for family or pedigree data structure are statistically and computationally challenging. As shown later, to effciently utilize the observed genotype data nested in the family structure, we need to marginalize the likelihood over (or integrate out)

all possible genotypes for subjects with missing genotype information, and meanwhile take into account the available genotypes in the family under the given pedigree structure.

Third, the LFS data are not a random sample, but have been collected through probands diagnosed with sarcoma at young ages. That is, the data oversampled sarcoma patients. Such a sampling scheme inevitably creates bias, known as *ascertainment bias*, and should be properly adjusted to obtain unbiased results. Several likelihood-calibrated models have been developed to correct the ascertainment bias, including the retrospective model (Kraft and Thomas; 2000), the conditional-on-ascertainment variable model (Ewens and Shute; 1986; Pfeiffer et al.; 2001), and the ascertainment-corrected joint model (Kraft and Thomas; 2000; Iversen and Chen; 2005), among others.

To address these challenges, in this article, we develop a Bayesian semiparametric approach to estimate the cancer-specific age-at-onset penetrance in the presence of the competing risks of developing multiple cancers. We employ a Bayesian semiparametric competing risk model to model the time to different types of cancer and introduce the family-wise likelihood to minimize information loss from missing genotypes and harness the information contained in the pedigree structure. We employ the peeling algorithm (Elston and Stewart; 1971) to evaluate the family-wise likelihood, and utilize the ascertainment-corrected joint model (Kraft and Thomas, 2000) to correct the ascertainment bias.

The rest of the article is organized as follows. In Section 2, we define the cancer-specific age-at-onset penetrance and describe our Bayesian semiparametric competing risk model including details about the family-wise likelihood and the ascertainment bias correction. In Section 3, we provide an algorithm to fit the models and carry out a simulation study in Section 4. We apply the proposed methodology to the LFS data in Section 5. Discussions follow in Section 6.

2 Model

2.1 Cancer-specific Age-at-onset Penetrance

Let G denote a subject's genotype, and X denote the baseline covariates (e.g., gender). Suppose that K types of cancer are under consideration and compete against each other such that the occurrence of one type of cancer censors the other types of cancer. Let T_k denote the time to the kth type of cancer, $k = 1, \dots, K$, and define $T = \min_{k \in \{1,\dots,K\}} T_k$ and $Y = \min \{T, C\}$, where C is a conditional random censoring time given G and X, i.e., $T \perp C | G, X$. Let D denote the cancer type indicator, with $D = k$ if $T = T_k < C$ (i.e., the kth type of cancer that occurs); otherwise, $D = 0$ (i.e., censored observation). The actual observed time-to-event data are $H = (Y, D)$.

Traditionally, when analyzing subjects at risk of developing a single disease, the age-at-onset penetrance is defined as the probability of having the disease at a certain age given a certain genotype. In order to study LFS, where subjects simultaneously have the (competing) risk of developing multiple types of cancer, this standard definition must be extended. Borrowing ideas from the competing risk literature, we define the kth *cancer-specific age-at-onset penetrance*, denoted by $q_k(t/G, X)$, as the probability of having the kth type of cancer at age t

prior to developing other cancers (competing risks), given a specific genotype G and additional baseline covariates X if necessary, that is,

$$q_k(t|G,X) = \Pr(T \leq t, D = k|G,X), \quad k = 1, \cdots, K. \quad (1)$$

The cancer-specific penetrance $q_k(t|G,X)$ can be estimated as

$$q_k(t|G,X) = \int_0^t \lambda_k(u|G,X)S(u|G,X)du, \quad k = 1, \cdots, K, \quad (2)$$

where

$$\lambda_k(t|G,X) = \lim_{h \downarrow 0} \frac{\Pr(t \leq T < t + h, D = k|T > t, G, X)}{h}, \quad (3)$$

and

$$S(t|G,X) = \exp\left\{ - \sum_{k=1}^{K} \Lambda_k(t|G,X) \right\}$$

With $\Lambda_k(t|G,X) = \int_0^t \lambda_k(u|G,X)du$. In the competing risk literature, $\lambda_k(t|G, X)$ and $\Lambda_k(t|G, X)$ are referred to as the cancer-specific hazard and cancer-specific cumulative hazard, respectively. We note that it may be tempting to define the cancer-specific age-at-onset penetrance function as $\Pr(T_k \leq t|G, X)$, which is analogous to the conventional definition of penetrance for a single disease. However, that quantity is not identifiable in nonparametric models (Tsiatis; 1975).

Besides cancer-specific penetrance, it is often of practical interest to estimate the overall age-at-onset penetrance, defined as

$$q(t|G,X) = \Pr(T \leq t|G,X) \quad (4)$$

which is the probability that a subject has any type of cancer by age t given his/her genotype G and baseline characteristics X, and can be calculated through the cancer-specific penetrance $q_k(t|G, x)$ using $q(t|G,X) = \sum_{k=1}^{K} q_k(t|G,X)$

2.2 Competing Risk Model

Let $\mathbf{Z} = (G, X, G \times X)^T$, with $G \times X$ denoting the interaction between G and X. For ease of exposition, hereafter we focus on the LFS data with X denoting gender, coded as 1 for the male and 0 for the female, and G denoting the TP53 mutation status. As LFS is autosomal dominant, we use $G = 1$ to denote genotype Aa or AA, and $G = 0$ to denote genotype aa,

where A and a denote the (minor) mutated and wildtype alleles in the TP53 tumor suppressor gene, respectively. We model the hazard for the kth type of cancer, say $\lambda_k(t|\mathbf{Z}, \xi_{i,k})$, using a frailty model as follows:

$$\lambda_k(t|\mathbf{Z}, \xi_{i,k}) = \lambda_{0,k}(t)\xi_{i,k}\exp\left\{\boldsymbol{\beta}_k^T\mathbf{Z}\right\} \quad k = 1, \cdots, K, \quad (5)$$

where $\boldsymbol{\beta}_k$ denotes the regression coeffcient parameter vector; $\lambda_{0,k}(t)$ is a baseline hazard function; and $\xi_{i,k}$ is the ith family-specific frailty (or random effect) used to account for the within-family correlation induced by non-genetic factors that are not included in X. The pedigree information (or genetic relationship) within a family will be incorporated through the family-wise likelihood described in Section 2.3. We assume that $\xi_{i,k}$ follows a gamma distribution, $\xi_{1,k}, \cdots, \xi_{I,k} \sim Gamma(v_k, v_k)$. Such a gamma frailty has been widely used in frailty models (Duchateau and Janssen; 2007).

Under this model, the cancer-specific age-at-onset penetrance can be expressed as

$$q_k(t|\mathbf{Z}) = \int_0^t \int_{\boldsymbol{\xi} \in [0,\infty)^K} \lambda_k(u|\mathbf{Z}, \xi_k)S(u|\mathbf{Z}, \boldsymbol{\xi})f(\boldsymbol{\xi}|v)d\boldsymbol{\xi}du \quad (6)$$

$$= \int_0^t \frac{v_k}{\left(v_k - \log\left\{S_k^*(u|\mathbf{Z})\right\}\right)}\lambda_{0,k}(u)\exp\left\{\boldsymbol{\beta}_k^T\mathbf{Z}\right\} S(u|\mathbf{Z})du,$$

where $S_k^*(t|\mathbf{Z}) = \exp\left\{-\int_0^t \lambda_{k,0}(u)\exp\left\{\boldsymbol{\beta}_k^T\mathbf{Z}\right\}du\right\}$ and $S(t|\mathbf{Z}) = \prod_{k=1}^K S_k(t|\mathbf{Z})$ with

$$S_k(t|\mathbf{Z}) = \int_0^\infty \exp\left\{-\int_0^t \lambda_k(u|\mathbf{Z}, \xi_k)du\right\}f(\xi_k|v_k)d\xi_k$$

$$= \left[\frac{v_k}{v_k - \log\left\{S_k^*(t|\mathbf{Z})\right\}}\right]^{v_k}.$$

Because the penetrance depends on the survival function, it is imperative to specify the baseline hazard $\lambda_{0,k}(t)$, which appears in (5). To this end, we propose to approximate the cumulative baseline hazard $\Lambda_{0,k}(t) = \int_0^t \lambda_{0,k}(s)ds$ via Bernstein polynomials (Lorentz; 1953) since $\Lambda_{0,k}(t)$ is monotone increasing. Bernstein polynomials are popular in Bayesian nonparametric function estimation, with shape restrictions due to desired properties such as the optimal shape restriction property (Carnicer and Peña; 1993) and the convergence property of their derivatives (Lorentz; 1953). Without loss of generality, we assume t has been rescaled, e.g., by the largest observed time, such that $t \in [0, 1]$. Now, we have $\Lambda_{0,k}(t)$ approximated by Bernstein polynomials of degree M as follows (Chang et al.; 2005).

$$\Lambda_{0,k}(t) \approx \sum_{l=1}^M \omega_{l,k}\binom{M}{l}t^l(1-t)^{M-l} \quad (7)$$

where $\omega_{l,k} = \Lambda_{0,k}(l/M)$ and $\omega_{1,k} \quad \cdots \quad \omega_{M,k}$ to ensure that $\Lambda_{0,k}(t)$ is monotone increasing. Notice that l is running from 1 because of $\Lambda_{0,k}(0) = 0$. Applying the re-parameterization of $\gamma_{l,k} = \omega_{l,k} - \omega_{l-1,k}$ with $\omega_{0,k} = 0$ and $\gamma_{l,k} \quad 0$, $l = 1, \cdots, M$, the right-hand side of (7) can be equivalently rewritten as

$$\sum_{m=1}^{M} \gamma_{m,k} \int_0^t \frac{u^m(1-u)^{M}-m}{Beta(m, M-m+1)} du = \gamma_k^T \mathbf{B}_M(t), \quad (8)$$

where $\mathbf{B}_M(t) = (B_M(t, 1), \cdots, B_M(t, M))^T$, with $B_M(t, m)$ being the distribution function of the beta distribution evaluated at the value of t with parameters m and $M - m + 1$, and $\gamma_k = (\gamma_{l,k}, \cdots, \gamma_{M,k})^T$ (Curtis and Ghosh; 2011). Therefore, it follows that

$$\lambda_{0,k}(t) \approx \gamma_k^T \mathbf{b}_M(t) \quad (9)$$

Where $\mathbf{b}_M(t) = (b_M(t, 1), \cdots, b_M(t, M))^T$ and $b_M(t, m) = B_M(t, m)/ t$ (i.e., associated beta density). Finally, the frailty model (5) can be written as

$$\lambda_k(t|\mathbf{Z}; \beta_k, \gamma_k, \xi_{i,k}) = \left\{ \gamma_k^T \mathbf{b}_M(t) \right\} \xi_{i,k} \exp\left\{ \beta_k^T \mathbf{Z} \right\}. \quad (10)$$

The proposed nonparametric baseline hazard model (9) is more flexible than parametric models, such as exponential and Weibull models, without imposing a restrictive parametric structure on the shape of the baseline hazard. Compared to the piecewise constant hazard model, our approach produces a smooth estimate of hazard and also avoids selection of knots, which is often subjective. The numerical comparison of different baseline models is provided in Section 5.5 and Supplementary Materials Section C.

2.3   Family-wise Likelihood

Let i index the family and j index the individual within the family, where $i = 1, \cdots, I$, and $j = 1, \cdots, n_i$. For the ith family, let $\mathbf{H}_i = (H_{i1}, \cdots, H_{in_i})$ with $H_{ij} = (Y_{ij}, D_{ij})$ and $\mathbf{X}_i = (X_{i1}, \cdots, X_{in_i})$. Let $\mathbf{G}_{i,obs}$ and $\mathbf{G}_{i,mis}$ respectively denote the observed and missing parts of genotype data, i.e., $\mathbf{G}_i = (\mathbf{G}_{i,obs}, \mathbf{G}_{i,mis})$. Conditional on frailty $\xi_i = (\xi_{i,1}, \cdots, \xi_{i,K})$, the likelihood of $\mathbf{H}_i$ for the ith family is $\Pr(\mathbf{H}_i|\mathbf{G}_{i,obs}, \mathbf{X}_i, \theta, \xi_i)$ which we call the family-wise likelihood, where $\theta = \left\{ \left(\beta_k^T, \gamma_k^T \right) : k = 1, \cdots, K \right\}$ denotes a vector of model parameters except the frailty.

Evaluation of the family-wise likelihood $\Pr(\mathbf{H}_i|\mathbf{G}_{i,obs}, \mathbf{X}_i, \theta, \xi_i)$ is not trivial because the individual disease histories $H_{i1}, \cdots, H_{in_i}$ are not conditionally independent given $\mathbf{G}_{i,obs}$ and ξ_i, due to the dependency through $\mathbf{G}_{i,mis}$. Note that $H_{i1}, \cdots, H_{in_i}$ will be conditionally independent when conditional on complete genotype data $\mathbf{G}_i$ and ξ_i. In this article, we use Elston-Stewart's peeling algorithm (Elston and Stewart; 1971; Lange and Elston; 1975; Fernando et al.; 1993) to compute the family-wise likelihood, described as follows. We

assume that there is no loop in the pedigree, which is generally true in practice, and suppress the family subscript i and the conditional arguments except G_{obs} for notational brevity.

A pedigree without loop can be partitioned into two disjoint groups, known as anterior and posterior, that are connected only through an arbitrary pivot member, say j. The anterior are the member in the pedigree who are connected to the pivot member through his/her parents, and the posterior are the member in the pedigree who are connected to the pivot member through his/her spouse and o springs, see Figure 1 for an example. In our implementation, we use the proband as the pivot member of each family. Let $\mathbf{H}_j^-$, and $\mathbf{H}_j^+$ denote the phenotypes of anterior and posterior, respectively. We partition $\mathbf{H} = \left(\mathbf{H}_j^-, H_j \mathbf{H}_j^+\right)$. Because anterior and posterior are connected only through the pivot member j, $\mathbf{H}_j^-$ and $\mathbf{H}_j^+$ are conditionally independent given pivot member's genotype G_j.

If G_j is unobserved, the family-wise likelihood $P(\mathbf{H}/G_{obs})$ can be written as

$$
\begin{aligned}
\Pr(\mathbf{H}|\mathbf{G}_{obs}) &= \sum_{G_j} \left\{ \Pr\left(G_j \middle| \mathbf{G}_{obs}\right) \Pr\left(\mathbf{H}_j^-, H_j, \mathbf{H}_j^+ \middle| G_j, \mathbf{G}_{obs}\right) \right\} \qquad (11) \\
&= \sum_{G_j} \left\{ \Pr\left(G_j \middle| \mathbf{G}_{obs}\right) \Pr\left(\mathbf{H}_j^- \middle| G_j, \mathbf{G}_{obs}\right) \Pr\left(H_j \middle| G_j, \mathbf{G}_{obs}\right) \Pr\left(\mathbf{H}_j^+ \middle| G_j, \mathbf{G}_{obs}\right) \right\} \\
&= \sum_{G_j} \left\{ \mathbb{A}_j\left(G_j \middle| \mathbf{G}_{obs}\right) \Pr\left(H_j \middle| G_j\right) \mathbb{P}_j\left(G_j \middle| \mathbf{G}_{obs}\right) \right\},
\end{aligned}
$$

Where

$$
\begin{array}{llll}
(\text{Anterior probability of } j) & \mathbb{A}_j\left(G_j \middle| \mathbf{G}_{obs}\right) &=& \Pr\left(\mathbf{H}_j^-, G_j \middle| \mathbf{G}_{obs}\right), \\
(\text{Posterior probability of } j) & \mathbb{P}_j\left(G_j \middle| \mathbf{G}_{obs}\right) &=& \Pr\left(\mathbf{H}_j^+ \middle| G_j, \mathbf{G}_{obs}\right),
\end{array}
$$

and the individual likelihood $\Pr(H_j/G_j)$ is computed from the proposed model as

$$
\Pr\left(H_j \middle| G_j\right) \; := \; \Pr\left(H_{ij} \middle| G_{ij}, X_{ij}, \boldsymbol{\theta}, \xi_i\right) \; \propto \; \prod_{k=1}^{K} \left\{ \lambda_k\left(Y_{ij} \middle| \mathbf{Z}_{ij}, \boldsymbol{\theta}, \xi_i\right) \right\}^{\Delta_{ijk}} \quad \exp\left\{ - \Lambda_k\left(Y_{ij} \middle| \mathbf{Z}_{ij}, \boldsymbol{\theta}, \xi\right) \right\}
$$

with $\Delta_{ijk} = 1$ if $D_{ij} = k$ and 0 otherwise (Prentice et al.; 1978; Maller and Zhou; 2002). In the case that G_j is observed, the summation in (11) is not needed and the family-wise likelihood is reduced to

$$
\Pr(\mathbf{H}|\mathbf{G}_{obs}) = \mathbb{A}_j\left(G_j \middle| \mathbf{G}_{obs}\right) \Pr\left(H_j \middle| G_j\right) \mathbb{P}_j\left(G_j \middle| \mathbf{G}_{obs}\right). \quad (12)
$$

To calculate $\mathbb{A}_j\left(G_j \middle| \mathbf{G}_{obs}\right)$ and $\mathbb{P}_j\left(G_j \middle| \mathbf{G}_{obs}\right)$, $\mathbf{H}_j^-$ and $\mathbf{H}_j^+$ can be further partitioned into anterior and posterior in a similar way as above. Thus, the family-wise likelihood $\Pr(\mathbf{H}/G_{obs})$ can be evaluated in a recursive way. An illustrative example of using the peeling algorithm to

evaluate the family-wise likelihood is provided in Supplementary Materials Section A. Fernando et al. (1993) provides the details on the recursive formulation of the algorithm.

2.4 Ascertainment Bias Correction

For studies of rare diseases, such as LFS, ascertainment bias is inevitable when family data are collected through probands in high-risk populations in which disease cases are more likely to be observed. We employ the ascertainment-corrected joint (ACJ) likelihood (Kraft and Thomas; 2000; Iversen and Chen; 2005) to correct the ascertainment bias. Iversen and Chen (2005) provides excellent description on the general methodology of the ACJ approach. We here focus on the application of that approach to the LFS data, and refer the readers to Iversen and Chen (2005) for more details. Let $\mathscr{A}_i$ denote the ascertainment indicator variable, such that $\mathscr{A}_i = 1$ if the ith family is ascertained and 0 otherwise. In the LFS data, a family is ascertained and included in the sample only if the proband is diagnosed with sarcoma. Following Iversen and Chen (2005), the ACJ likelihood for the LFS data is given by

$$\Pr\left(\mathbf{H}_i, \mathbf{G}_{i,obs} \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i, \mathscr{A}_i = 1\right) = \tag{13}$$

$$\frac{\Pr\left(\mathscr{A}_i = 1 \middle| \mathbf{H}_i, \mathbf{G}_{i,obs}, \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right) \Pr\left(\mathbf{H}_i \middle| \mathbf{G}_{i,obs}, \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right) \Pr\left(\mathbf{G}_{i,obs} \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)}{\Pr\left(\mathscr{A}_i = 1 \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)}.$$

Because the ascertainment decision is made on the basis of H_{i1} (i.e., phenotype of the proband) in a deterministic way, the first term in the numerator of equation (13), i.e., $\Pr\left(\mathscr{A}_i = 1 \middle| \mathbf{H}_i, \mathbf{G}_{i,obs}, \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)$, is independent of the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}_i$. We assume that the distribution of genotype $\Pr\left(\mathbf{G}_{i,obs} \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)$, i.e., the third term in the numerator of equation (13), is also independent of both $\boldsymbol{\theta}$ and $\boldsymbol{\xi}_i$, the parameters of the penetrance model. This is a reasonable assumption that generally holds in practice. As a result, we have

$$\Pr\left(\mathbf{H}_i, \mathbf{G}_{i,obs}, \mathbf{X}_i \middle| \boldsymbol{\theta}, \boldsymbol{\xi}_i, \mathscr{A}_i = 1\right) \propto \frac{\Pr\left(\mathbf{H}_i \middle| \mathbf{X}_i, \mathbf{G}_{i,obs}, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)}{\Pr\left(\mathscr{A}_i = 1 \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)}. \tag{14}$$

This means that the ascertainment bias can be corrected by inverse-probability weighting the likelihood by the corresponding ascertainment probability, which is given by

$$\Pr\left(\mathscr{A}_i = 1 \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right) = \sum_{H_{i1}} \Pr\left(\mathscr{A}_i = 1 \middle| H_{i1}\right) \Pr\left(H_{i1} \middle| \mathbf{X}_{i1}, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right). \tag{15}$$

In the LFS data, a family is ascertained if the proband is diagnosed with sarcoma (coded as $D = 2$). Based on our experience, there is little evidence indicating that cancer patients treated at MD Anderson Cancer Center are systematically different from the population of

cancer patients in US. Therefore, our sample can be approximately viewed as a random sample from the US population of cancer patients that are ascertained under the same procedure. This assumption is also supported by the comparison of our penetrance estimates for the non-carriers to those based on the US population (see Section 5.4). Therefore, it follows

$$\Pr\left(\mathcal{A}_i = 1 \middle| Y_{i1}, D_{i1} = 2\right) = 1 \text{ and } \Pr\left(\mathcal{A}_i = 1 \middle| Y_{i1}, D_{i1} \neq 2\right) = 0.$$

In the case that the sampling population (e.g., cancer patients visiting MD Anderson Cancer Center) is different from the target population (e.g., US population of cancer patients), we should restrict the results and inference on the sampling population only.

Recalling $H_{i1} = (Y_{i1}, D_{i1})$, the ascertainment probability (15) is given by

$$\Pr\left(\mathcal{A}_i = 1 \middle| \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right) = \Pr\left(Y_{i1}, D_{i1} = 2 \middle| X_{i1}, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)$$

$$= \sum_G \Pr\left(Y_{i1}, D_{i1} = 2 \middle| G, X_{i1}, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right) \Pr\left(G \middle| X_{i1}\right)$$

$$= \sum_G \left[\lambda_2\left(Y_{i1} \middle| G, X_{i1}, G \times X_{i1}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2, \xi_{i,2}\right) \times \exp\left\{ - \sum_{k=1}^k \Lambda_k\left(Y_{i1} \middle| G, X_{i1}, G \times X_{i1}, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \xi_{i,k}\right) \right\} \Pr\left(G \middle| X_{i1}\right) \right],$$

$$(16)$$

where $\Pr(G/X_{i1})$ is the covariate specific prevalence of a genotype G, which is often assumed to be given (Iversen and Chen; 2005). In our application, the TP53 mutation prevalence is independent of $X =$ gender i.e., $\Pr(G/X) = \Pr(G)$, and can be calculated on the basis of the mutated allele frequency ϕ_A, i.e., $\Pr(G = 0) = (1 - \phi_A)^2$ and $\Pr(G = 1) = 1 - (1 - \phi_A)^2$. The prevalence of a germline TP53 mutation in the Western population is estimated as $\phi_A = 0.0006$ (Lalloo et al.; 2003).

As shown above, the key is that we assume that the mutated allele frequency ϕ_A is known or can be reliably estimated from external data sources. Given a known mutated allele frequency ϕ_A, the frequency of each genotype G can be determined using the Mendelian laws of inheritance. Thus, coupling with the penetrance model, the sampling probability can be estimated, e.g., equation (16), and used to inversely weight the observed data likelihood to make inference for the target population. For many genetic studies, it is often reasonable to assume that the mutated allele frequency ϕ_A is known or can be reliably estimated from external data sources.

The ACJ likelihood of the entire data of I mutually independent families is given by the product of (14)

$$\Pr\left(\mathbf{H}, \mathbf{G}_{obs} | \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathscr{A}\right) \propto \prod_{i=1}^{I} \frac{\Pr\left(\mathbf{H}_i | \mathbf{G}_{i,obs}, \mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)}{\Pr\left(\mathscr{A}_i = 1 | \boldsymbol{\theta}, \boldsymbol{\xi}_i\right)},$$

where $\mathbf{H} = (\mathbf{H}_i, \cdots, \mathbf{H}_I)$, $\mathbf{G} = (\mathbf{G}_{1,obs}, \cdots, \mathbf{G}_{I,obs})$ and $\mathscr{A} = \left(\mathscr{A}_1, \cdots, \mathscr{A}_I\right)$.

3 Prior and Posterior Sampling

We use an independent normal prior for $\boldsymbol{\beta}_k$, i.e., $\boldsymbol{\beta}_k \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, where $\mathbf{0}$ and $\mathbf{I}$ denote a zero vector and an identity matrix, respectively, and we set a large value of σ for vague priors. For the nonnegative parameter $\gamma_{m,k}$, $m = 1, \cdots, M$, $k = 1, \cdots, K$ for the baseline hazard, we use the noninformative flat prior. We assign v_k, $k = 1, \cdots, K$, the independent vague gamma prior $Gamma(0.01, 0.01)$. See Section 5.6 for the results of the sensitivity analysis of $\gamma_{m,k}$ and v_k. For the choice of M, a large value provides more flexibility to model the shape of the baseline hazard, but at the cost of increasing the computational burden. Gelfand and Mallick (1995) suggest that a small value of M works well for most applications. We set $M = 5$ in the analysis.

Let $\Pr(\boldsymbol{\theta})$ and $\Pr(\boldsymbol{v})$ denote the prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{v}$, respectively. The joint posterior distribution of $\boldsymbol{v}$, $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ is given by

$$\Pr\left(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{v} | \mathbf{H}, \mathbf{G}_{obs}, \mathbf{X}, \mathscr{A}\right) \propto \Pr\left(\mathbf{H}, \mathbf{G}_{obs} | \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathscr{A}\right) \cdot \Pr(\boldsymbol{\theta}) \cdot \Pr(\boldsymbol{\xi}|\boldsymbol{v}) \cdot \Pr(\boldsymbol{v}).$$

We employ the random walk Metropolis-Hastings algorithm within Gibbs sampler to sample the posterior distribution. We generate 100,000 posterior samples in total and take every fifth sample for thinning after discarding the first 10,000 samples for burn-in. We implement the Markov chain Monte Carlo (MCMC) algorithm in R, which takes about three seconds per single MCMC iteration. We observe that the physical computing time is approximately linear, corresponding to the number of families, I, regardless of the family size n_i.

4 Simulation

We conduct a simulation study to evaluate the performance of the proposed method. Suppose that there are two competing cancers, indicated by $D = 1$ and 2, respectively. We simulate 200 families of three generations with 30 members (see Figure 1) that are collected through probands indexed by $\{1\}$ in Figure 1 with the second type of cancers (i.e., $D = 2$), as follows:

1. We first simulate a genotype $G \sim Bernoulli(0.0001)$ for the proband. Given G, we then simulate his/her true time to cancer, T_k, $k = 1, 2$, from the following cancer-specific frailty model:

$$\lambda_k(t|G) = \lambda_{0,k}(t)\xi_k \exp\left(\beta_k G\right), \quad k = 1, 2, \quad (17)$$

with $\beta_1 = 4$, $\beta_2 = 10$, $\lambda_{0,1}(t) = 0.1$, $\lambda_{0,2}(t) = 0.0005$, , and $\xi_1, \xi_2 \overset{iid}{\sim} Gamma(0.25, 0.25)$. We choose these simulation parameters such that the second type of cancer (i.e., $D = 2$) is rare with the prevalence of about 0.0003, while that of the first type of cancer is about 0.05. Random censoring time C is simulated from *Exponential*(2). To mimic the ascertainment process of the LFS data, only probands with $D = 2$ are selected and included in the sample as probands. We repeat the above procedure until 200 probands are collected.

2. Given probands' data, we generate genotypes of their family members as follows. If proband {1} is a non-carrier ($G = 0$), all family members are set as non-carriers; other wise, we randomly select one of proband's parents {3, 4} as a carrier and set the another as a non-carrier. O springs and siblings of the proband, including {7, 8, 9, 10, 11, 12}, are set as carriers with probability 0.5. If {11} is carrier, his o springs, including {19, 20, 21}, are set as carries with probability 0.5, otherwise set as noncarriers. Geno-types of {22, 23, 24} are generated similarly based on the genotype of their mother {12}. Assuming that the mutation is extremely rare, the family members who are not genetically related with the proband, including {2, 5, 6, 13, 14, 15, 16, 17, 18, 25, 26, 27, 28, 29, 30}, are set as non-carriers.

3. Given the genotypes, the time to cancer of the family members are generated from model (17).

4. Lastly, we randomly delete genotypes for a half of subjects who are not a proband.

We set $M = 3$ for the Bernstein model for the baseline hazard functions, $\lambda_{0,k}(t)$, $k = 1, 2$. For estimation, we generate 10,000 posterior samples after discarding the first 1,000 samples as burn-in. Trace plots suggest that the posterior sampling converges well.

The proposed method has three main components: the family-wise likelihood to handle missing genotypes, the ACJ likelihood to correct the ascertainment bias, and the frailty to capture the family-specific random effects. To evaluate the effects of these three components, we compare our approach with alternative approaches, under which there is (1) no missing genotype, (2) no ascertainment bias correction, and (3) no frailty.

Table 2 shows absolute biases and standard deviations of estimates under different approaches. For the baseline hazard $\lambda_{0,k}(t)$, bias and standard deviation are numerically integrated over t. We can see that the estimates without ascertainment bias correction are severely biased, especially for β_2 and $\lambda_{0,2}(t)$, showing the importance of performing the ascertainment bias correction. In addition, the estimates with frailty tend to have smaller biases than those assuming no frailty. Lastly, the effciency loss due to missing genotypes is generally small, suggesting that the family-wise likelihood effciently utilizes the observed data.

5 Application

We apply the proposed methodology to analyze the LFS data. We consider three types of LFS-related cancers ($K = 3$): breast cancer ($k = 1$), sarcoma ($k = 2$), and other cancers ($k = 3$). Because the individuals with breast cancer in the LFS data are all female (Table 1), we impose the following constraint on the hazard of developing breast cancer:

$$\lambda_1(t|G, X) = \begin{cases} 0, & \text{for } X = 0 \text{(male)}, \\ \lambda_{0,1}(t)\xi_1\exp\{\beta_{G,1}G\}, & \text{for } X = 1 \text{(female)}, \end{cases} \quad (18)$$

while other types of cancer ($k = 2, 3$) are assumed to follow the model of the form (5). There is only one baseline covariate available in the LFS database (i.e., gender), however our method can readily accommodate more covariates. We ignore all cancers that occurred after 75 years of age and treat them as censored at age 75, since cancers diagnosed after 75 years of age are clinically irrelevant for estimating the penetrance of LFS.

5.1 Model Parameter Estimates

Posterior estimates for the regression coefficients β_k and the inverse of the frailty variances ν_k, $k = 1, 2, 3$ are reported in Table 3. Genotype has a strong effect on the incidence of all cancer types, with TP53 mutation carriers being more likely to have cancers. Gender also plays a significant role in sarcoma and other cancers. The regression coeffcient of gender is negative, suggesting that males in this population are more likely to develop sarcoma and other cancers than females.

The estimates of ν_ks are quite large, which suggests that after accounting for the pedigree structure through the family-wise likelihood, within-family correlations are not very strong in this particular dataset. To check this, we compared the penetrance estimates obtained from our model to those from the model that does not include frailty and found them to be quite similar (see Supplementary Materials Section D.3). Although the model without frailty may be preferred in practice due to its parsimony, we present the results of the frailty model to emphasize that our approach allows for further flexibility; the results are nearly identical in terms of the penetrance estimates.

Figure 2 depicts the posterior estimates of the cumulative baseline hazard. Age has stronger effects on breast and other cancers than on sarcoma. The cumulative baseline hazards of breast and other cancers increase exponentially with age, while that of sarcoma increases approximately linearly with age. We observe that the uncertainty of the sarcoma baseline hazard estimate is much larger than those of the others. This is because the ascertainment bias is generated from the probands with sarcoma, which makes the ascertainmentbias-corrected likelihood (14) more sensitive to the parameters directly related to sarcoma.

5.2 Age-at-onset Penetrance

The first three panels (a)–(c) of Figure 3 depict the estimated age-at-onset penetrances, $q_k(t/G, X)$, $k = 1, \cdots, 3$, respectively, for breast cancer, sarcoma, and other cancers. It is not surprising that the TP53 mutation carriers ($G = 1$) have higher risk of developing cancer than

the non-carriers ($G = 0$), regardless of cancer type. The patterns of cancer-specific penetrance are quite different across cancer types, which justifies the proposed cancer-specific approach. It is of clinical interest that there is a sizable chance that the female TP53 mutation carrier will develop breast cancer before 20 years of age, which is rarely seen in females with BRCA1 and BRCA2 mutations (two well-known susceptibility gene mutations for breast cancer) (Berry et al.; 2002). This suggests that early-onset breast cancer is an important feature of TP53 mutation. We also find that non-carriers have very low probability of developing sarcoma, although the data contain many cases of sarcoma in non-carriers due to the use of individuals with sarcoma as probands for collecting the samples (see Table 1). In contrast, ignoring the ascertainment bias leads to substantially biased estimates, see Supplementary Materials Section D.2 for the comparison between our estimates and the estimates without performing ascertainment bias correction.

Figure 3, panel (d) shows the overall age-at-onset penetrance obtained by stacking three cancer-specific penetrances, i.e., $q(t|G, X) = \Sigma_{k=1}^{3} q_k(t|G, X)$. The overall age-at-onset penetrance quantifies the probability of having any type of cancer by a certain age for carriers of TP53 mutations. Among the non-carriers, females have lower cancer risk than males; whereas the female mutation carrier has higher risk than the male mutation carrier due to the excessively high risk of the female carrier developing breast cancer. Overall, TP53 mutation carriers have very high lifetime risk of developing cancer, demonstrating the importance of the accurate detection of TP53 germline mutations.

5.3 Personalized Risk Prediction

An important application of our analysis results and estimate of age-at-onset penetrance $q_k(t|G, X)$ is to provide a personalized risk prediction for future subjects who are at risk of developing LFS-related cancers. Our prediction method has two important advantages. First, it allows us to incorporate the subject's family cancer history to make more accurate risk prediction. Second, it is capable to make risk prediction for a subject without knowing his/her genotype. This is desirable because in practice, genetic test is often of a great financial and psychological burden for patients. Making risk prediction without performing a genetic test allows us to quickly detect individuals with high risk of LFS and provide prompt and proper clinical treatments during an early stage of disease, which is particularly important in the management of rare diseases such as LFS. Specifically, given a family's cancer history $\mathbf{H}_i$ and covariates $\mathbf{X}_i$, the risk that the jth individual in the ith family will develop the kth type of cancer by age t, $R_{ijk}(t|\mathbf{H}_i, \mathbf{X}_i)$, is predicted by

$$R_{ijk}(t|\mathbf{H}_i, \mathbf{X}_i) = \Pr(T_{ij} \leq t, D_{ij} = k|\mathbf{H}_i, \mathbf{X}_i) = \sum_{G_{ij} \in \{0, 1\}} \Pr(G_{ij}|\mathbf{H}_i, \mathbf{X}_i) q_k(t|G_{ij}, X_{ij}). \quad (19)$$

That is, the predicted cancer-specific risk is a weighted average of the cancer-specific penetrance $q_k(t|G_{ij}, X_{ij})$. The weight $\Pr(G_{ij}|\mathbf{H}_i, \mathbf{X}_i)$, also known as carrier probability, is the likelihood that the subject carries a specific genotype G_{ij}, given his/her family cancer history $\mathbf{H}_i$ and covariates $\mathbf{X}_i$. It can be routinely calculated using Bayes' rule and Mendelian laws of inheritance, see Supplementary Materials Section B for details. As we assume that the

subject's genotype G_{ij} is unknown, the calculation of the risk in (19) is marginalized over all possible values of G_{ij}.

To illustrate the utility of our method, consider two hypothetical families that have similar pedigree structures, but different genotypes and cancer histories, as shown in Figure 4. Family 1 does not carry the mutated allele and has three cases of cancer (two breast and one other cancers), and family 2 carries the mutated allele with four cases of cancer (one breast, two sarcoma and one other cancers). As mothers (the second generation) in both families had breast cancer, it is of great interest to predict the cancer risk for their daughters, referred to as counselees 1 and 2 in Figure 4. We consider two situations: the genotypes of the counselees are known or unknown. Specifically, when the genotypes of the counselees and their family are unknown, we predict the cancer risk for the counselees based on equation (19) with the cancer-specific penetrance estimated from the LFS data. When the genotypes of the conselees are known (i.e., conselee 1 is non-carrier and 2 is carrier), the risk prediction is straightforward and the cancer risk of the conselees is simply the estimated cancer-specific penetrances $q_k(t|G, X)$. Figure 5 shows the predicted cancer-specific risks of the counselees when their genotypes are known and unknown. Clearly, counselee 2 has a substantially higher risk of developing cancer than the counselee 1. Based on this result, we may recommend more frequent cancer screening for counselee 2. We note that counselee 2 has a very low risk of developing sarcoma although her family has two cases of sarcoma. This is because, as shown in Figure 3(b), the penetrance for sarcoma is high in male, but very low in female.

5.4 External and Interval Validation

As an external validation, we compare our estimates of non-carrier penetrance to those provided by the National Cancer Institute on the basis of the Surveillance, Epidemiology, and End Results (SEER) data. SEER is an authoritative source of information on cancer incidence and survival in the United States. It currently collects and publishes cancer incidence and survival data from population-based cancer registries that cover approximately 28% of the US population. SEER is the only comprehensive source of population-based information in the United States that includes the stage of cancer at the time of diagnosis and patient survival data. The SEER estimate can be regarded as a reference estimate for the normal US population (i.e., non-carrier). More details regarding SEER estimates can be found at http://seer.cancer.gov.

Figure 6 compares the penetrance of breast cancer, sarcoma, and all cancers for non-carriers to the most recent SEER estimates based on the data collected from 2008 to 2010. We can see that the estimates of non-carrier penetrance are generally consistent with the corresponding SEER estimates, suggesting that the proposed methodology performs well. For the purpose of comparison, we also show the estimate of the overall cancer penetrance based on the conventional Cox model for the time to cancer diagnosis using subjects with known genotypes. As shown in Figure 6, panel (c), the estimate of the overall cancer risk based on the proposed method is much closer to the SEER estimate than the estimate based on the Cox model.

We conduct internal validation through cross-validation. First, we randomly split the data (i.e., 186 families) into two halves. We use one half (i.e., 93 families) as the training families $\left\{\left(\mathbf{H}_i^{tr}, \mathbf{X}_i^{tr}, \mathbf{G}_{obs,i}^{tr}\right), i = 1, \cdots, 93\right\}$, and the other half as the test families, $\left\{\left(\mathbf{H}_{i'}^{tr}, \mathbf{X}_{i'}^{tr}, \mathbf{G}_{obs,i'}^{tr}\right), i' = 1, \cdots, 93\right\}$. Next, we estimate the cancer-specific penetrance using the training families, denoted by $\hat{q}_k^{tr}(t|G, X)$. Based on this estimate and equation (19), we predict the cancer-specific risk at a given age t_c for subjects in the test families, i.e., $R_{i'jk}\left(t_c\big|\mathbf{H}_{i'}^{\text{ts}}, \mathbf{X}_{i'}^{\text{ts}}\right)$. Given a certain risk cuto ψ, we predict that a subject will have kth type of cancers by age t_c if $R_{i'jk}\left(t_c\big|\mathbf{H}_{i'}^{\text{ts}}, \mathbf{X}_{i'}^{\text{ts}}\right) > \psi$. By varying the risk cuto ψ and comparing the predicted cancer status with the actually observed cancer status of the test families, we obtain the receiver operating characteristic (ROC) curves of our cancer risk prediction model. Figure 7 depicts the ROC curve of the predicted risk of the test family members at age 50 years for different cancer types. These results show reasonable performance, with the area under the ROC curves (AUC) being 0.773, 0.791 and 0.755 for predicting breast cancer, sarcoma and other cancers, respectively. For breast cancer, the ROC curves are generated from the females only since we assume no breast cancer for the males. We also consider the ROC curves for other caner-onset ages, $t_c = 30$, 40, and 60 years. The results are generally similar to that of $t_c = 50$ years, see Supplementary Materials Section F.

5.5 Model Comparison

Due to the complicated structure of the LFS data (e.g., family structure, missing genotype, ascertainment bias and competing risks), standard model diagnosis tools for survival models, such as residuals (Schoenfeld; 1982; Therneau et al.; 1990) and chi-squared goodness-of-fit tests (Hjort; 1990; Hollander and Pena; 1992; Li and Doss; 1993), are not applicable here. We assess the adequacy of the proposed model through model comparison. We consider four alternative models. The first three models are obtained by replacing the Bayesian non-parametric baseline hazard model with three parametric models: the exponential, Weibull, and piecewise-constant models, respectively. For the piecewise-constant model, we use four equally spaced knots to obtain five partitions. The fourth model is obtained by removing the frailty $\xi_{i,k}$ from the competing risk model (5). We use two metrics to measure the goodness of fit of the models: the deviance information criterion (DIC) and conditional predictive ordinate (CPO, Ibrahim et al.; 2005). The DIC measures the overall goodness of fit of a model and the CPO measures the predictive ability of a model. The CPO for the ith family is defined as

$$\text{CPO}_i = \Pr\left(\mathbf{H}_i\big|\mathscr{D}_{(-i)}, \mathbf{G}_{i,obs}, \mathbf{X}_i, \boldsymbol{\xi}_i, \mathscr{A}_i\right) = \left[E\left(\frac{1}{\Pr\left(\mathbf{H}_i\big|\mathbf{G}_{i,obs}, \mathbf{X}_i, \boldsymbol{\xi}_i, \boldsymbol{\theta}, \mathscr{A}_i\right)}\right)\right]^{-1} \quad (20)$$

where $\mathscr{D}_{(-i)} = \left(\mathbf{H}_{(-i)}, \mathbf{G}_{(-i),obs}, \mathbf{X}_{(-i)}\right)$ represents the data with the ith family data deleted, and the expectation is made with respect to the posterior distribution of $\boldsymbol{\theta}$. The Monte Carlo approximation of (20) is given by

$$\widehat{\mathrm{CPO}}_i = \left[\frac{1}{L} \sum_{\ell = 1}^{L} \frac{1}{\Pr\left(\mathbf{H}_i \middle| \mathbf{G}_{i,obs}, \mathbf{X}_i, \widehat{\boldsymbol{\xi}}_{i,(l)}, \widehat{\boldsymbol{\theta}}_{(\ell)}, \mathscr{A}_i\right)} \right]^{-1}$$

where $\widehat{\boldsymbol{\xi}}_{i,(\ell)}$ and $\widehat{\boldsymbol{\theta}}_{(\ell)}$ denote posterior samples from the ℓth MCMC iteration, $\ell = 1, \cdots, L$.

Table 4 shows the DIC and $\sum_{i=1}^{I} \log \widehat{\mathrm{CPO}}_i$, known as the pseudo-marginal log-likelihood (PsML), for the different models. Smaller DIC values and larger PsML values suggest a better model. The proposed model based on Bernstein polynomials provides better goodness of fit and predictive ability than the models with exponential, Weibull, or piecewise-constant baseline hazards. The difference between the proposed model and the model without frailty is small, suggesting a weak within-family correlation. This is concordant with our finding that $\boldsymbol{\nu}$ estimates are large (see Table 3). For the purpose of comparison, we also perform the analysis based only on the subset of the data for whom the genotypes are observed, and the analysis without ascertainment bias correction. The estimates of cancer-specific penetrance under different approaches are provided in *Supplementary Materials* (Section D).

5.6 Sensitivity Analysis

We consider nine different combinations of priors for $\gamma_{m,k}$ and ν_k: three different priors for $\gamma_{m,k}$ including flat prior, *Gamma*(0.01, 0.01), and *Gamma*(1, 1); and three priors for $\nu_k \sim$ *Gamma*(0.01, 0.01), *Gamma*(0.1, 0.1) and *Gamma*(1, 1). The results (see Supplementary Materials Section E) show that the estimates are not particularly sensitive to the choice of priors.

6 Discussion

In the LFS study, estimating cancer-specific penetrance is not trivial under the presence of competing risks, but is essential for providing better treatment that is personalized to the patient's needs. We developed a cancer-specific age-at-onset penetrance model and proposed an associated Bayesian estimation scheme. The proposed method can incorporate all the family histories in the estimation by exploiting the family-wise likelihood. We also corrected the ascertainment bias, which is an important task in family data studies of rare diseases.

One detriment when modeling the cause-specific hazard in competing risk analysis is that covariate effects on the subdistribution (i.e., cancer-specific penetrance) are not interpretable. As an alternative, Fine and Gray (1999) proposed a proportional model for the subdistribution that enables us to directly assess the covariate effects on the corresponding cancer-specific penetrance. It is not diffcult to equivalently rewrite the individual likelihood in terms of the cancer-specific penetrance and the associated derivative (Maller and Zhou; 2002). The family-wise likelihood approach can be similarly applied to this alternative modeling approach.

In the LFS study, a patient can have multiple primary cancers during his or her lifetime. In the current approach, we consider only the first cancer that occurred and discard all the

subsequent cancer history. In order to incorporate a longitudinal history that may involve multiple cancers, our approach can be extended to the so-called multi-state model (Putter et al.; 2007) to recurrently observe multiple failures. In theory, the multi-state model can be regarded as an extended version of the competing risk model. However, it is practically challenging to collect data for a suffcient number of subjects who have multiple primary cancers in order to attain an appropriate level of estimation accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

References

Abel L, Bonney GE and Rao D (1990). A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases, Genetic epidemiology 7(6): 391–407. [PubMed: 2292365]

Berry DA, Iversen ES, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, Lerman C, Watson P, Lynch HT and Hilsenbeck SG (2002). Brcapro validation, sensitivity of genetic testing of brca1/ brca2, and prevalence of other breast cancer susceptibility genes, Journal of Clinical Oncology 20(11): 2701–2712. [PubMed: 12039933]

Birch JM, Alston RD, McNally R, Evans D, Kelsey AM, Harris M, Eden OB and Varley JM (2001). Relative frequency and morphology of cancers in carriers of germline tp53 mutations., Oncogene 20(34): 4621–4628. [PubMed: 11498785]

Carnicer JM and Peña JM (1993). Shape preserving representations and optimality of the bernstein basis, Advances in Computational Mathematics 1(2): 173–196.

Chang I-S, Hsiung CA, Wu Y-J and Yang C-C (2005). Bayesian survival analysis using bernstein polynomials, Scandinavian journal of statistics 32(3): 447–466.

Chatterjee N, Hartge P and Wacholder S (2003). Adjustment for competing risk in kin-cohort estimation, Genetic epidemiology 25(4): 303–313. [PubMed: 14639700]

Curtis MS and Ghosh SK (2011). A variable selection approach to monotonic regression with bernstein polynomials, Journal of Applied Statistics 38(5): 961–976.

Duchateau L and Janssen P (2007). The frailty model, Springer.

Elston RC and Stewart J (1971). A general model for the genetic analysis of pedigree data, Human heredity 21(6): 523–542. [PubMed: 5149961]

Ewens W and Shute NC (1986). A resolution of the ascertainment sampling problem i. theory, Theoretical Population Biology 30(3): 388–412. [PubMed: 3810507]

Fernando R, Stricker C and Elston R (1993). An e cient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops, Theoretical and Applied Genetics 87(1–2): 89–93. [PubMed: 24190198]

Fine JP and Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk, Journal of the American Statistical Association 94(446): 496–509.

Gauderman W and Faucett C (1997). Detection of gene-environment interactions in joint segregation and linkage analysis, American Journal of Human Genetics 61: 1189–1199. [PubMed: 9345092]

Gelfand AE and Mallick BK (1995). Bayesian analysis of proportional hazards models built from monotone functions, Biometrics pp. 843–852. [PubMed: 7548703]

Gorfine M and Hsu L (2011). Frailty-based competing risks model for multivariate survival data, Biometrics 67(2): 415–426. [PubMed: 20707868]

Gorfine M, Hsu L, Zucker DM and Parmigiani G (2014). Calibrated predictions for multivariate competing risks models, Lifetime data analysis 20(2): 234–251. [PubMed: 23737081]

Hashemian AH, Hajizadeh E, Kazemnezad A, Meshkani MR and Mehdipour P (2009). Kin-cohort estimate of penetrance with piecewise weibull model, World Applied Sciences Journal 6(1): 77–82.

Hjort NL (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates, The Annals of Statistics pp. 1221–1258.

Hollander M and Pena EA (1992). A chi-squared goodness-of-fit test for randomly censored data, Journal of the American Statistical Association 87(418): 458–463.

Hwang S-J, Lozano G, Amos CI and Strong LC (2003). Germline p53 mutations in a cohort with childhood sarcoma: sex differences in cancer risk, The American Journal of Human Genetics 72(4): 975–983. [PubMed: 12610779]

Ibrahim JG, Chen M-H and Sinha D (2005). Bayesian survival analysis, Wiley Online Library.

Iversen ES and Chen S (2005). Population-calibrated gene characterization: estimating age at onset distributions associated with cancer genes, Journal of the American Statistical Association 100(470): 399–409. [PubMed: 18418465]

Kraft P and Thomas D (2000). Bias and e ciency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods, American Journal of Human Genetics 66: 1119–1131. [PubMed: 10712222]

Lalloo F, Varley J, Ellis D, Moran A, O'Dair L, Pharoah P and Evans DGR (2003). Prediction of pathogenic mutations in patients with early-onset breast cancer by family history, The Lancet 361(9363): 1101–1102.

Lange K and Elston R (1975). Extensions to pedigree analysis, Human Heredity 25(2): 95–105. [PubMed: 1150306]

Li G and Doss H (1993). Generalized pearson-fisher chi-square goodness-of-fit tests, with applications to models with life history data, The Annals of Statistics pp. 772–797.

Lorentz GG (1953). Bernstein polynomials, American Mathematical Soc.

Lustbader E, Williams W, Bondy M, Strom S and Strong L (1992). Segregation analysis of cancer in families of childhood soft-tissue-sarcoma patients., American journal of human genetics 51(2): 344. [PubMed: 1642235]

Malkin D, Li FP, Strong LC, Joseph F Fraumeni J, Nelson CE, Kim DH, Kassel J, Gryka MA, Bischo FZ, Tainsky MA and Friend SH (1990). Germline p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms, Science 250(4985): 1233–1238. [PubMed: 1978757]

Maller RA and Zhou X (2002). Analysis of parametric models for competing risks, Statistica Sinica 12(3): 725–750.

Nichols KE, Malkin D, Garber JE, Fraumeni JF and Li FP (2001). Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers, Cancer Epidemiology Biomarkers & Prevention 10(2): 83–87.

Pfei er RM, Gail MH and Pee D (2001). Inference for covariates that accounts for ascertainment and random genetic e ects in family studies, Biometrika 88(4): 933–948.

Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell V and Breslow N (1978). The analysis of failure times in the presence of competing risks, Biometrics pp. 541–554. [PubMed: 373811]

Putter H, Fiocco M and Geskus R (2007). Tutorial in biostatistics: competing risks and multi-state models, Statistics in medicine 26(11): 2389–2430. [PubMed: 17031868]

Schoenfeld D (1982). Partial residuals for the proportional hazards regression model, Biometrika 69(1): 239–241.

Srivastava S, Zou Z, Pirollo K, Blattner W and Chang EH (1990). Germ-line transmission of a mutated p53 gene in a cancer-prone family with li–fraumeni syndrome, Nature 348(6303): 747–749. [PubMed: 2259385]

Strong LC, Williams WR and Tainsky MA (1992). The li–fraumeni syndrome: from clinical epidemiology to molecular genetics, American journal of epidemiology 135(2): 190–199. [PubMed: 1536134]

Therneau TM, Grambsch PM and Fleming TR (1990). Martingale-based residuals for survival models, Biometrika 77(1): 147–160.

Tsiatis A (1975). A nonidentifiability aspect of the problem of competing risks, Proceedings of the National Academy of Sciences 72(1): 20–22.

Wang Y, Clark LN, Marder K and Rabinowitz D (2007). Nonparametric estimation of age-at-onset distributions from censored kin-cohort data, Biometrika 94(2): 403–114.

Wu C-C, Strong LC and Shete S (2010). Effects of measured susceptibility genes on cancer risk in family studies, Human genetics 127(3): 349–357. [PubMed: 20039063]
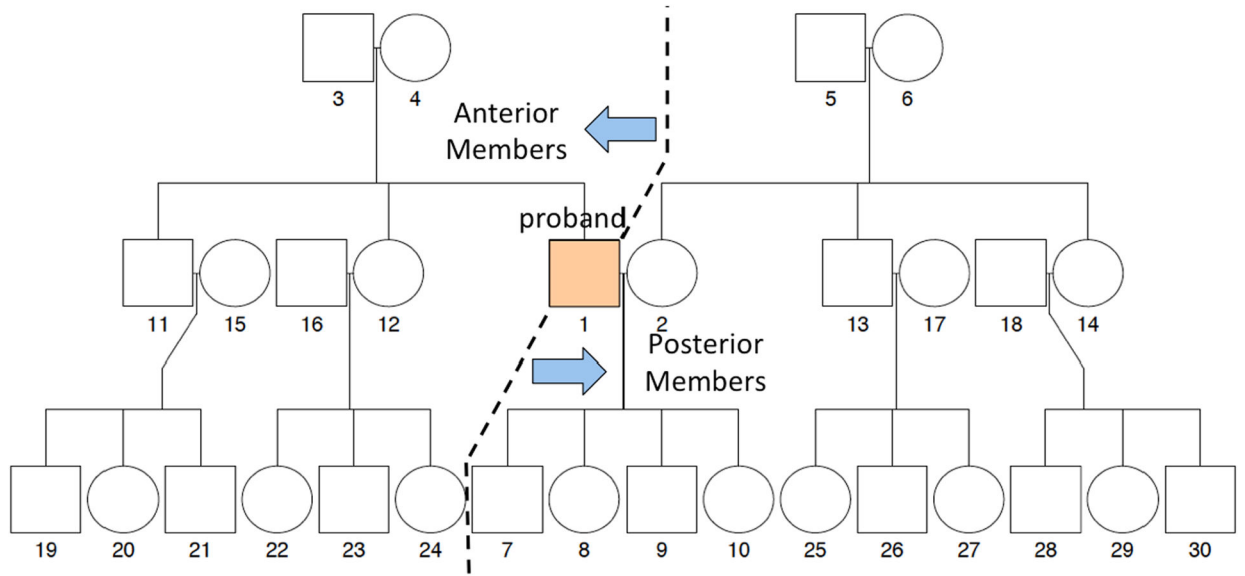
Figure 1:
Pedigree of the simulated family of three generations with 30 members.

(a) Breast   (b) Sarcoma   (c) Others
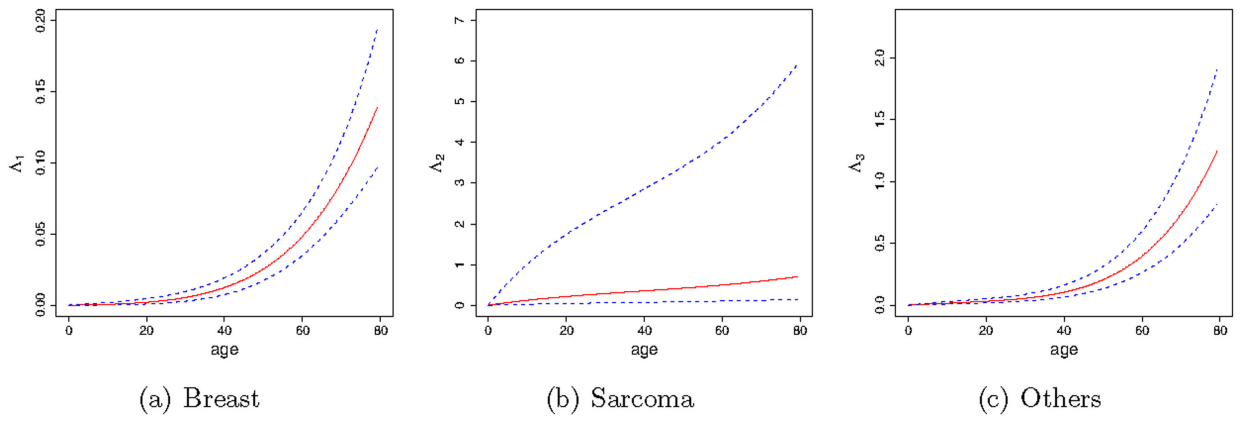
Figure 2:
Posterior estimates of the cancer-specific cumulative baseline hazard functions for breast cancer (a), sarcroma (b) and other cancers (c). Dashed lines indicate 95% credible band of the estimates.
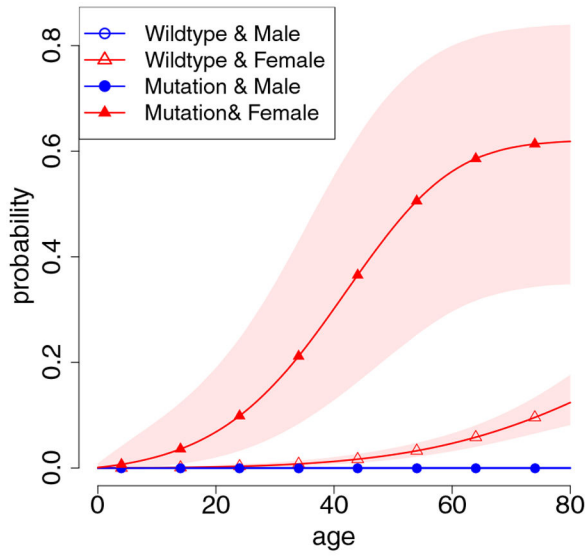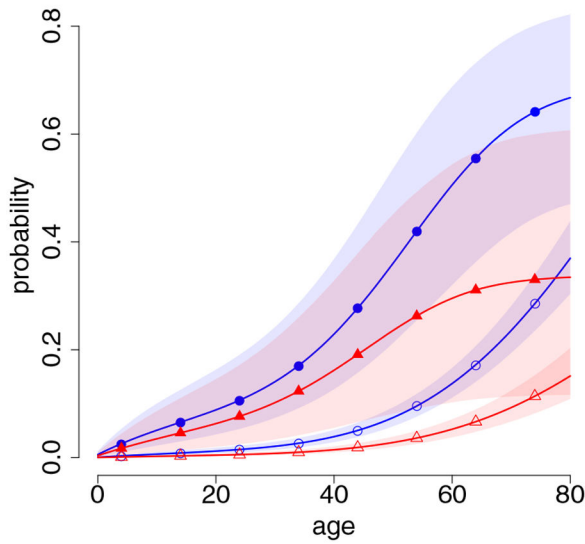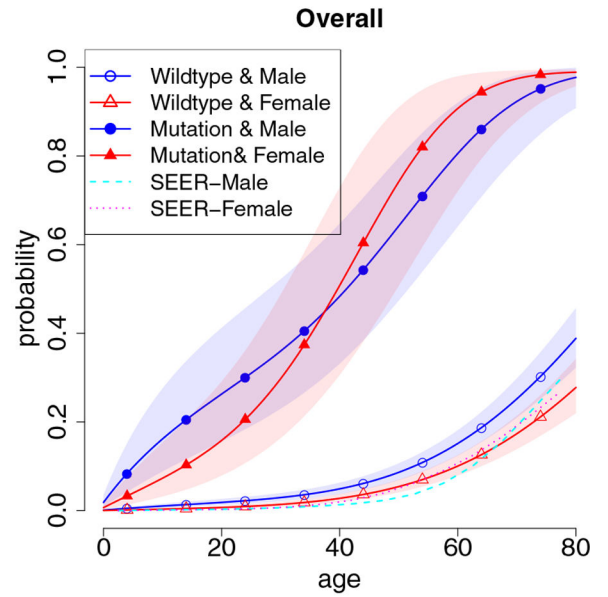
(a) Breast, $q_1(t|G, X)$
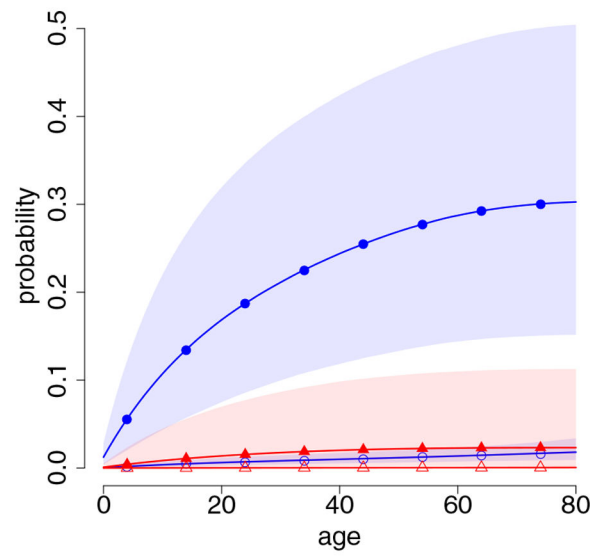
(b) Sarcoma, $q_2(t|G, X)$

(c) Others, $q_3(t|G, X)$

(d) Overall, $q(t|G, X) = \sum_k q_k(t|G, X)$

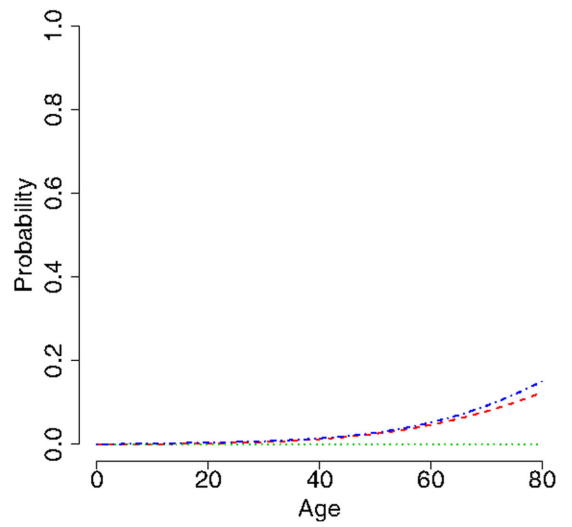**Figure 3:**
Cancer-specific age-at-onset penetrances $q_k(t|G, X)$, $k = 1, 2, 3$, are depicted in (a)–(c), and the overall cancer penetrance $q(t|G, X)$ is given in (d). Shaded areas denote the 95% credible bands.

(a) Family 1   (b) Family 2
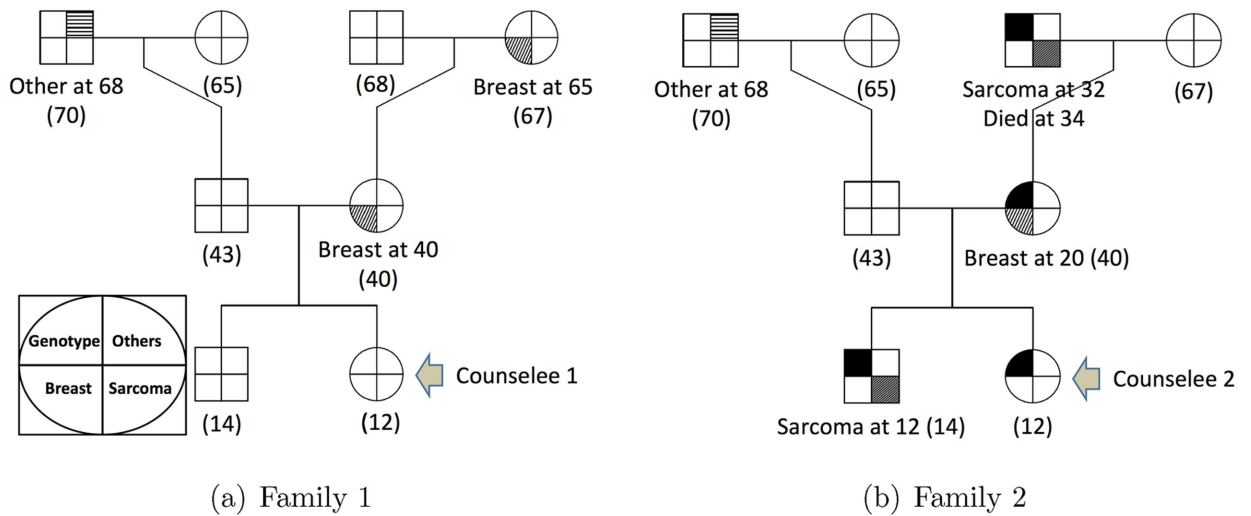
Figure 4:
Pedigrees of two families, where square and circle represent male and female subjects, respectively. The symbol is partitioned into four sections, which represent statuses of genotype (topleft), breast cancer (bottom left), sarcoma (bottom right), and other cancers (topright). Filled sections represent that the subject carries a mutated allele or had a certain type of cancer. The number in the parentheses is subject's current age.
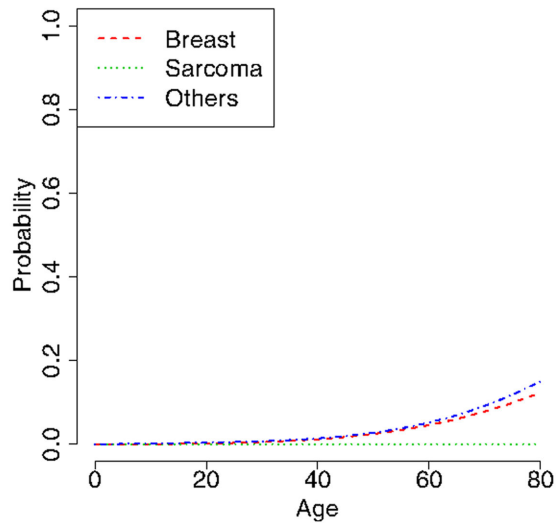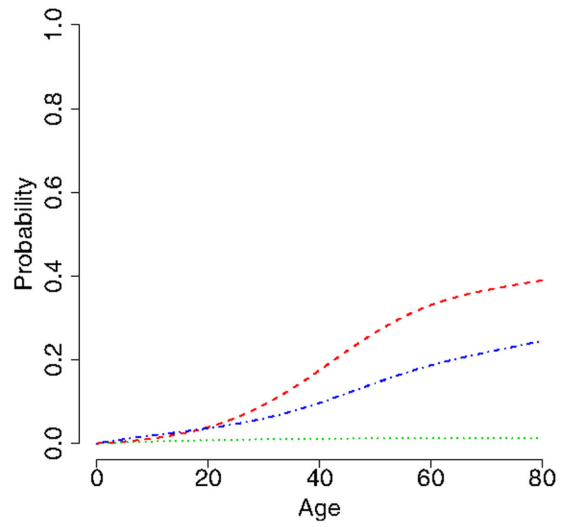
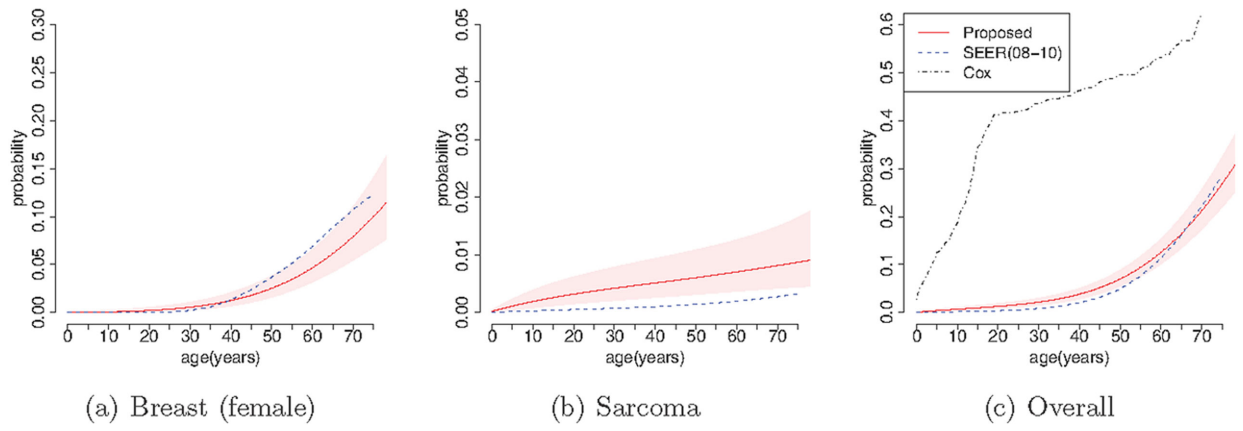(a) Counselee 1 (G is known)

(b) Counselee 1 (G is unknown)
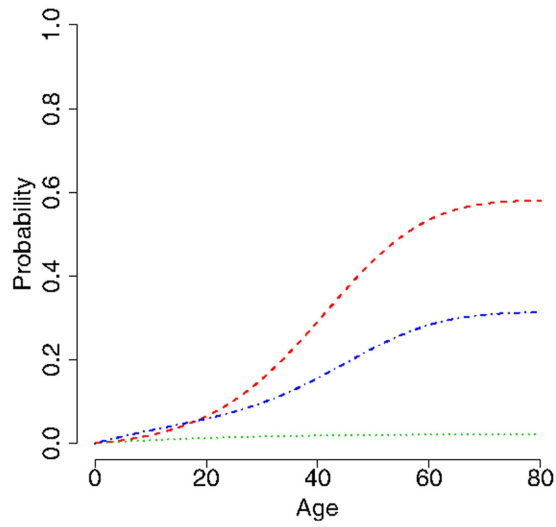
(c) Counselee 2 (G is known)

(d) Counselee 2 (G is unknown)

Figure 5:
Predicted cancer-specific risk for counselees 1 and 2 when their genotypes G are known or unknown.

(a) Breast (female)   (b) Sarcoma   (c) Overall

Figure 6:
External validation of the estimated penetrance for non-carrier (solid lines) through comparison to the SEER estimates (dashed lines). Shaded areas represent the 95% credible bands. For panel (c), the dotted line indicates the estimate based on the Cox model.
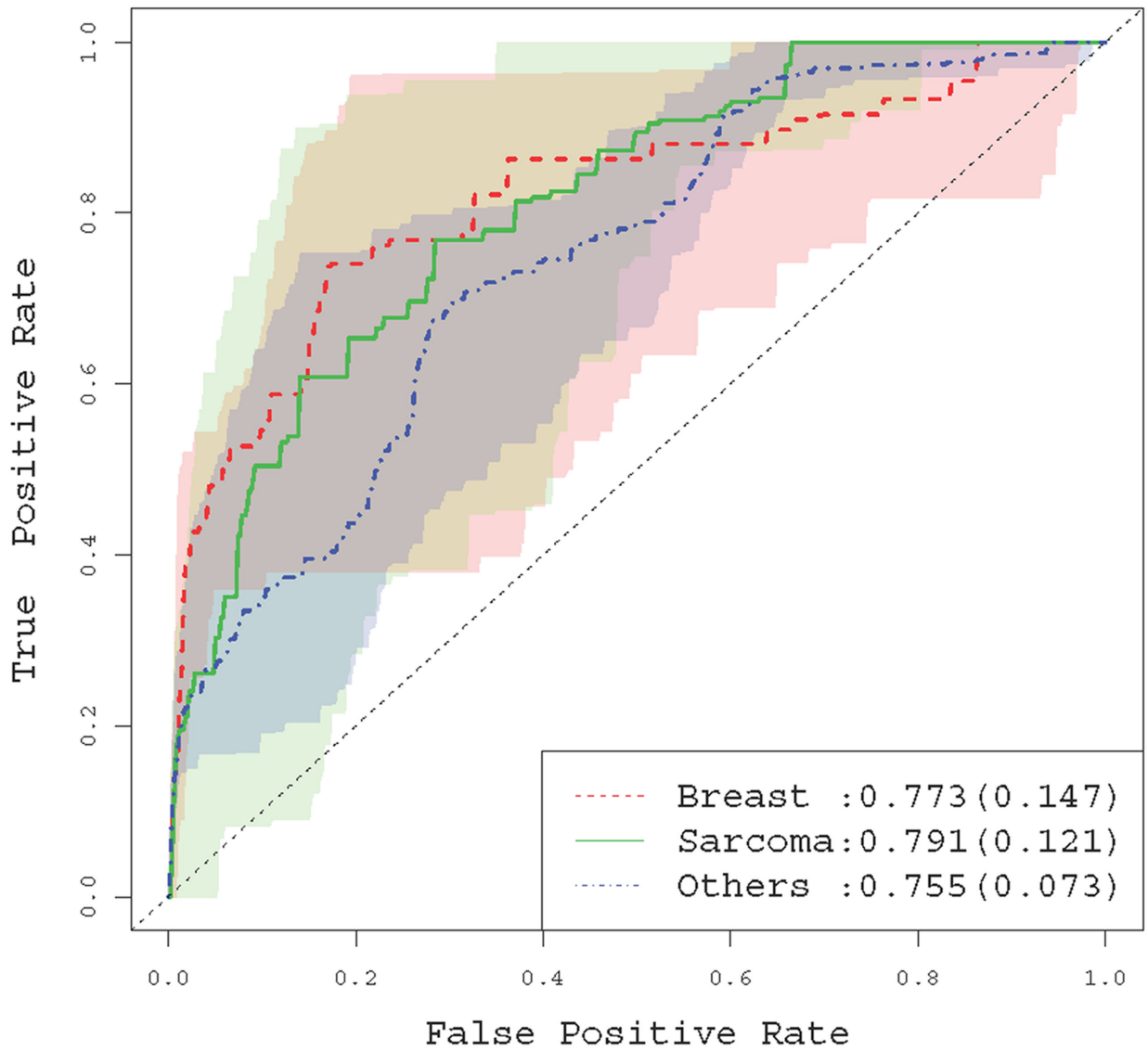
Figure 7:
The ROC curve for the cancer-specific risk prediction at age 50. Values in the parentheses are standard deviations of the AUC. Shaded areas represent pointwise 95% variations of the ROC curves from different random partitions.

Table 1:

Frequency table for LFS data.

Gender	Genotype	Breast	Sarcoma	Others	Censored	Total
	Unknown	0	11	130	1275	1416
Male	Wildtype	0	76	30	295	401
	Mutation	0	12	27	9	48
	Subtotal	0	99	187	1579	1865
	Unknown	39	4	62	1204	1309
Female	Wildtype	8	96	17	343	464
	Mutation	19	12	7	10	48
	Subtotal	66	112	86	1557	1821
Total		66	211	273	3136	3686

Table 2:

Absolute biases and standard deviations (in parentheses) of estimates based on 100 simulations.

	Genotype	No bias correction				Bias correction			
		No fraility		Frailty		No fraility		Frailty	
β_1	No missing	1.1968	(.3608)	0.7818	(.3385)	1.0667	(.3633)	0.4905	(.3420)
	Missing	1.4363	(.4117)	1.2627	(.3421)	1.2824	(.4120)	0.8681	(.3942)
β_2	No missing	5.5993	(.1973)	4.7515	(.2051)	0.8659	(.3220)	0.2347	(.4627)
	Missing	5.7480	(.2225)	5.4368	(.2267)	1.2012	(.3016)	0.4764	(.3293)
$\lambda_{0,1}(t)$	No missing	0.0227	(.0366)	0.0194	(.0409)	0.0190	(.0394)	0.0116	(.0479)
	Missing	0.0184	(.0374)	0.0167	(.0398)	0.0166	(.0395)	0.0123	(.0449)
$\lambda_{0,2}(t)$	No missing	0.1025	(.0505)	0.0914	(.0518)	0.0004	(.0004)	0.0043	(.0038)
	Missing	0.1340	(.0641)	0.1116	(.0540)	0.0010	(.0008)	0.0053	(.0045)

Table 3:

Posterior estimates of regression coeffcients β and inverse variances of the frailty ν.

Cancer	Parameter	Mean	SD	2.5%	97.5%
Breast	Genotype	3.560	0.516	2.541	4.544
	(Frailty Var)$^{-1}$	6.126	1.850	3.185	10.347
Sarcoma	Genotype	2.464	0.895	0.675	4.182
	Female	−3.677	1.077	−6.176	−1.902
	Interaction	0.971	0.548	−0.110	2.040
	(Frailty Var)$^{-1}$	6.574	1.990	3.490	11.227
Others	Genotype	1.576	0.769	0.072	3.072
	Female	−0.993	0.186	−1.366	−0.647
	Interaction	0.559	0.574	−0.620	1.628
	(Frailty Var)$^{-1}$	7.148	2.001	3.986	11.857

Table 4:

Comparison of the proposed model with four alternative models.

Model	Baseline hazard	Frailty included	DIC	PsML
1	Exponential	Yes	3273.7	−1657.120
2	Weibull	Yes	3020.2	−1512.252
3	Piecewise	Yes	3010.3	−1513.405
4	Bernstein.	No	2989.3	−1499.735
Proposed	Bernstein	Yes	2983.7	−1499.689