Personalized Risk Prediction for Cancer Survivors: A Bayesian Semi-parametric Recurrent Event Model with Competing Outcomes

Nam H Nguyen[a,b], Seung Jun Shin[c], Elissa B Dodd-Eaton[a], Jing Ning[d], Wenyi Wang[a*]

[a]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX; [b]Department of Statistics, Rice University, Houston, TX; [c]Department of Statistics, Korea University, Seoul, Korea; [d]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX.

*Corresponding author: Wenyi Wang; Email: wwang7@mdanderson.org.

Abstract

Multiple primary cancers are increasingly more frequent due to improved survival of cancer patients. Recent studies established a link between the characteristics of the first primary cancer and the risk of developing subsequent primary cancers. Characterization of risk trajectory is still urgently needed for the healthcare policy development for cancer survivors. To fill this need, we propose a Bayesian semi-parametric framework, where the occurrence processes of the competing cancer types follow independent non-homogeneous Poisson processes that depend on patient-specific covariates, including the type and age at diagnosis of the first primary. We further propose two approaches to address different data types, an individual likelihood approach for unrelated patient data and a family likelihood approach for family history data. Upon applying our models to a historically collected dataset (2000-2015 at MD Anderson Cancer Center) from 429 families affected with Li-Fraumeni syndrome, which

is enriched with multiple primary tumors and diverse cancer types, we have derived a comprehensive suite of age-to-onset penetrance curves for cancer survivors which are covariate-specific, i.e., personalized. This includes the first set of penetrance curves for second primary lung cancer, which are potentially impactful to cancer screening decisions. Using Receiver Operating Characteristic (ROC) curves, we have validated the good predictive performance of our penetrance curves in predicting second primary lung cancer, sarcoma, breast cancer, and all other cancers combined, with areas under the curves (AUCs) at 0.89, 0.91, 0.76 and 0.68, respectively. Our framework successfully provides a first covariate-adjusted quantitative foundation for characterizing cancer risk trajectories among cancer survivors, hence moving a step closer to developing personalized health management for this unique population.

Keywords: Cancer survivors; Frailty modeling; Markov chain Monte Carlo; Personalized risk prediction; Recurrent event; Competing risk.

1 Introduction

The most recent report from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program estimated approximately 20% of all incident cancer cases in the United States are second or later primary cancers, i.e., additional cancers occurring in cancer survivors (Forjaz et al., 2022). By 2026, the total number of cancer survivors living in the US will exceed 20 million (Hulvat, 2020). Currently, the health care management of cancer survivors is largely indifferent from the management of either at-risk healthy individuals or new cancer patients, thus not yet adequately configured to deliver the desirable high-quality care in this growing population (Mayer et al., 2017). Recently, large population-based epidemiology studies have identified a few important features for cancer survivors: patients with bladder cancer present the highest risk of developing second cancer (Donin et al., 2016); patients with breast cancer have a higher risk of developing a second primary cancer, lung cancer in particular, than the general population (Bao et al., 2021; Wang et al., 2021); and 14% of patients with colorectal cancer had previously developed primary cancer, most frequently in the prostate or breast (Pruitt et al., 2022). Accurate characterization of the onset of second primary cancers is essential for cancer prevention strategy development, such as the United States Preventive Services Task Force (USPSTF) guidelines. One such example is the lung-cancer screening, which has been recommended to adults who are deemed at high-risk, because of the significant survival improvement with diagnosis at an early stage. However, the current recommendation does not include previous malignancy as a high-risk feature (Nobel et al., 2022), due to a lack of accurate personalized risk characterization in such a population.

The NCI SEER program provides a range of databases that allow researchers to directly query cancer-specific incidence rates, defined as the number of new cases per 100,000 people

per year, for the first primary cancer (www.seer.cancer.gov/seerstat version 8.4.0.1). These incidence rates can be segmented by 5-year age groups between age 0 and 85, as well as by basic demographic features such as gender. To address the increasing number of cancer survivors, the program recently updated its SEER*Stat software to allow for queries from patients that developed a specific cancer type as the second primary. Although the incidence rates for the second primary cancer can be approximated from the SEER data, such approximation is subject to biases, and hence does not accurately reflect age-at-onset penetrances. Furthermore, these rates are not fully personalized, i.e., they do not account for important person-specific features beyond age group and gender. For example, in the context of inherited cancer syndromes, where the personal risk of developing cancer is significantly increased by the carrier status of a genetic mutation (Parmigiani et al., 1998; Chen et al., 2006), the necessary covariate of mutation status is not captured by SEER.

Although covariate-adjusted risk characterization is urgently needed, unbiased and comprehensive data collection for cancer survivors, as needed for mathematical modeling, remains a major bottleneck. So far, a pan-cancer risk characterization study is not available except for the SEER study mentioned above. On the other hand, there exists a well-characterized and well-ascertained inherited and rare cancer syndrome, called Li-Fraumeni syndrome (LFS) (Li and Fraumeni, 1969), which is mainly caused by germline mutations in the tumor suppressor gene *TP53* (Malkin et al., 1990). LFS presents a wide spectrum of cancer types, as well as a much higher incidence rate of multiple primary cancers than the general population (50% as compared to 2-17%) (Mai et al., 2016; Vogt et al., 2017). We therefore resort to datasets that were collected from patients affected by LFS (see **Table 1** for a dataset example) for real-data motivation, model training and the model-based risk prediction application. People affected by LFS are more likely to develop a spectrum of

cancer types, including breast cancer, soft-tissue sarcoma, osteosarcoma, and others, and repeatedly over their lifetime (Li and Fraumeni, 1969). Because of the inheritance of *TP53* mutations within the family, the family members of patients diagnosed with LFS and the patients themselves live with a constant high level of anxiety, thus actively participating in risk counseling and cancer screening. This further highlights the need for cancer-specific risk prediction among those who survived the first primary cancer while addressing genetic mutations as covariates.

Other than risk modeling for individual data, we also take the opportunity to develop additional methods that address the family inheritance structure presented in the LFS datasets. There are several models in the literature that attempt personalized risk prediction given family data such as those collected from LFS. Chen et al. (2009) proposed a frailty model with two-phase case-control data. Choi (2012) introduced a shared frailty model for analyzing correlated time-to-event data arising from family-based studies. Both models, however, do not appropriately account for the pedigree structure of families that is needed for genetic inheritance models. To address this problem, Choi et al. (2016) took one step further by proposing a progressive three-state Markov model, in which a single-step EM algorithm is used to calculate genotype probabilities for individuals with no genetic testing results based on genotype data from other family members. Shin et al. (2019) and Shin et al. (2020) developed Bayesian semiparametric models, which introduced the peeling algorithm (Elston and Stewart, 1971) to take into account familial genetic structure, and the ascertainment-corrected joint (ACJ) approach (Iversen and Chen, 2005) to correct for ascertainment bias. However, none of these approaches have been able to model multiple cancer types beyond the first primary cancer.

	Positive families				Negative families			
	Wildtype	Mutation	Unknown	**Total**	Wildtype	Mutation	Unknown	**Total**
Male								
Healthy	112	41	2,227	**2,380**	7	0	2,221	**2,228**
FPC	11	50	339	**400**	47	0	408	**455**
SPC	1	39	17	**57**	32	0	43	**75**
Subtotal	124	130	2,583	**2,837**	86	0	2,672	**2,758**
Female								
Healthy	133	32	1,997	**2,162**	11	0	2,119	**2,130**
FPC	9	76	375	**460**	104	0	423	**527**
SPC	1	97	41	**139**	113	0	60	**173**
Subtotal	143	205	2,413	**2,761**	228	0	2,602	**2,830**
Total	267	335	4,996	**5,598**	314	0	5,274	**5,588**

Table 1: Categorization of family members in the MDACC cohort by gender, number of primary cancers and *TP53* mutation status. FPC: first primary cancer; SPC: second primary cancer. Families are grouped into two categories, positive: at least one family member with a *TP53* mutation, and negative: no mutation carriers.

In this paper, we introduce a novel Bayesian semiparametric framework that jointly models both multiple primary cancers and multiple cancer types. To achieve this, we utilize non-homogeneous Poisson processes to model the occurrence processes of primary cancers, each of which is characterized by an intensity function that is cancer-type-specific. This modeling approach allows predictions of the second primary to be dependent on the type and timing of the first primary. Under this novel framework, we derive an explicit expression for the individual likelihood contribution, and further a family-wise likelihood through integration across family members. We train and cross-validate our model on a patient cohort, collected from MD Anderson Cancer Center (MDACC) using clinical LFS criteria (Li et al., 1988; Chompret et al., 2001; Bougeard et al., 2015) from year 2000 to 2015 (**Table 1**).

The paper is organized as follows. In Section 2, we present the modeling and derivation of individual and family-wise likelihoods, techniques for addressing necessary ascertainment bias correction and incorporating frailty into family data, as well as simulation studies to

evaluate the parameter estimation performances. In Section 3, we demonstrate how the individual likelihood can be used to construct effective risk models, with a special emphasis on lung cancer. In Section 4, we apply our family-wise model to the LFS dataset to obtain novel penetrance estimates for *TP53* mutation carriers and non-carriers. Both sections include a cross-validation study. Section 5 concludes with future research directions.

2 Method

2.1 Age-at-onset penetrance of multiple cancer types in multiple primaries

We assume multiple cancer occurrences as recurrent events (Cook and Lawless, 2007). To describe the model, we begin by introducing some notations. Let $N(t)$ be a non-homogeneous Poisson process that counts the number of primary cancers by age t, and Z be the indicator of cancer type (i.e. $Z \in \{1, \ldots, K\}$) with K being the number of cancer types. In addition, $H(t)$ denotes the historical information of $N(t)$ up to time t, and $\boldsymbol{X}$ is a vector of patient-specific covariates, respectively. The kth cancer-specific intensity function, $\lambda_k(t|\boldsymbol{X}, H(t)), k = 1, 2, \cdots, K$, is defined as

$$\lambda_k(t|\boldsymbol{X}, H(t)) = \lim_{\Delta t \to 0^+} \frac{P[N(t + \Delta t) - N(t) > 0, Z = k|\boldsymbol{X}, H(t)]}{\Delta t}.$$

We note that the overall intensity of $N(t)$ is $\lambda(t|\boldsymbol{X}, H(t)) = \sum_{k=1}^{K} \lambda_k(t|\boldsymbol{X}, H(t))$. The notion of age-at-onset penetrance is defined as the probability that a patient develops the disease of interest by age t given his or her characteristics (Langbehn et al., 2004). Since then, the definition has been extended to accommodate more complex settings. For example, the kth cancer-specific age-at-onset penetrance is defined as the conditional probability

of having the kth type of cancer by age t prior to developing other cancer types. Compared to our previous model that estimates age-at-onset cancer-specific penetrance among the first primary cancers (Shin et al., 2020), here we focus on addressing the need of characterizing the kth cancer type among multiple primary cancers during the patient's lifetime. The desired solution will add another layer of complexity to further extend the definition of penetrance. To this end, let L be a non-negative integer that denotes the number of primary cancer occurrences. Then, we define the cancer-specific age-at-onset penetrance for multiple primary cancer as the probability of having the kth cancer type at the lth occurrence by age t given the $(l-1)$th occurrence at time u in the past, conditional on covariates representing patient characteristics and cancer history up to a time point u. Let T_l be the time and Z_l be the cancer type of the lth cancer diagnosis. The penetrance of interest, denoted by $q_{kl}(t|u, \boldsymbol{X}(u)), k = 1, \ldots, K; l = 1, \ldots, L$, is then given by

$$q_{kl}(t|u, \boldsymbol{X}, H(u)) = P[T_l \leq t, Z_l = k | T_{l-1} = u, \boldsymbol{X}, H(u)] \quad \text{for } t > u,$$

where $T_0 = 0$ for convenience. For $t > u$, we then have

$$q_{kl}(t|u, \boldsymbol{X}, H(u)) = \int_u^t \lambda_k(v|\ \boldsymbol{X}, H(u)) \prod_{a=1}^{K} S_a(v|u, \boldsymbol{X}, H(u))\ dv, \tag{1}$$

where $S_k(t|u, \boldsymbol{X}, H(u)) = \exp\Big[-\int_u^t \lambda_k(v|\boldsymbol{X}, H(u))\ dv\Big]$ for $k = 1, \cdots, K$.

For a patient with $l-1$ primary cancers, we are also interested in the probability of having the next cancer of any type by age t given the $(l-1)$th occurrence at time u. This penetrance can be obtained by summing over all cancer types, i.e., $q_l(t|u, \boldsymbol{X}, H(u)) = \sum_{k=1}^{K} q_{kl}(t|u, \boldsymbol{X}, H(u))$.

2.2 Intensity function modeling

One of the challenges in modeling recurrent events with competing risks is accounting for the correlations between the events. To address this, we introduce a binary variable $D_k(t)$ that indicates whether the patient had the kth cancer type as the first primary by time t. That is, $D_k(t) = 1$ if $T_1 < t$ and $Z_1 = k$, and $D_k(t) = 0$ otherwise. The variable $D_k(t)$ allows the risk of later primary to depend on the type of the first one. The relationship between the first and second primary has been established in a number of epidemiological studies (Travis et al., 2005; Mariotto et al., 2007; Donin et al., 2016). We define $D_{k,l}(t)$, $k = 1, \ldots, K$, as indicators of the lth primary cancer. That is, $D_{k,l}(t) = 1$ if $T_l < t$ and $Z_l = k$. By incorporating the set of covariates $\{D_{k,l}(t) : k = 1 \ldots, K \text{ and } l = 1, \ldots, L-1\}$ into the model specification, we allow the risk of the lth primary cancer to depend on the types of the previous $l-1$ occurrences. In our application, we consider only dependence on the type of the first primary cancer due to limited data availability. For each individual with an inherited cancer syndrome such as LFS, we let G be the mutation status (0 for wildtype, 1 for mutated) for the gene of interest, and S be the sex (0 for female, 1 for male). Therefore, the covariate vector is $\boldsymbol{X}(t) = \{G, S, D_1(t), \ldots, D_K(t)\}^T$.

We model the cancer-specific intensities using a proportional intensity model as follows:

$$\lambda_k(t|\boldsymbol{X}(t)) = \lambda_{0,k}(t) \exp(\boldsymbol{\beta}_k^T \boldsymbol{X}(t)), \qquad k = 1, \ldots, K \tag{2}$$

where $\boldsymbol{\beta}_k$ denotes the vector of coefficients, and $\lambda_{0,k}(t)$ is the baseline intensity function. Let $\Lambda_{0,k}(t) = \int_0^t \lambda_{0,k}(v)\, dv$ be the cumulative baseline intensity function. We model $\Lambda_{0,k}(t)$ using Bernstein polynomials, which are often used to approximate functions with constraints such as monotonicity (Lorentz, 1953; Shin et al., 2020). After rescaling t to $[0, 1]$, we can

approximate $\Lambda_{0,k}(t)$ using the Bernstein polynomials as follows.

$$\Lambda_{0,k}(t) \approx \sum_{m=1}^{M} \Lambda_{0,k}\left(\frac{m}{M}\right)\binom{M}{m} t^m (1-t)^{M-m},$$

where M denotes the order of the Bernstein polynomials. By reparameterizing $\gamma_{m,k} = \Lambda_{0,k}\left(\frac{m}{M}\right) - \Lambda_{0,k}\left(\frac{m-1}{M}\right)$ with $\Lambda_{0,k}(0) = 0$, it can be shown that

$$\lambda_{0,k}(t) \approx \sum_{m=1}^{M} \gamma_{m,k} f_{B(m,M-m+1)}(t),$$

where $f_{B(m,M-m+1)}$ denotes the beta density with parameters m and $M-m+1$ (Curtis and Ghosh, 2011). In practice, death from other causes is another competing risk that should be taken into account. One can model the hazard function for death in an identical way to the other cancer-specific intensities from (2).

2.3 Individual likelihood with known G

We first derive the full likelihood of an individual's observed cancer history, including the development of first, second, or more primary cancers over time, until censorship, when the genotype G is known. For individual j, where $j = 1, 2, \cdots, n$, let L_j be the number of observed cancer occurrences. Although we focus on $L_j \leq 2$ in the applications, we describe our model in a more general setting. We observe a dataset $\{\boldsymbol{t}_j, \boldsymbol{z}_j, c_j, g_j, s_j\}_{j=1}^n$, where $\boldsymbol{t}_j$ and $\boldsymbol{z}_j$ are L_j-dimensional vectors that indicate the times and cancer types at diagnoses; c_j is the censoring time; and g_j and s_j denote *TP53* mutation status and sex, respectively. Now, we introduce $d_{j,k}(t) = 1$ if $t_{j,1} < t$ and $z_{j,1} = k$, and $d_{j,k}(t) = 0$ otherwise. Finally, the vector of covariates is given by $\boldsymbol{x}_j(t) = (g_j, s_j, d_{j,1}(t), \ldots, d_{j,K}(t))^T$.

Let $\boldsymbol{\beta} = \{\boldsymbol{\beta}_k : k = 1, \ldots, K\}$ and $\boldsymbol{\gamma} = \{\gamma_{m,k} : m = 1 \ldots, M; k = 1 \ldots, K\}$. For an

individual with cancer history $\boldsymbol{h}_j = \{\boldsymbol{t}_j, \boldsymbol{z}_j, c_j\}$, the likelihood can be expressed as a product of the observed cancer occurrences and the censoring event

$$P[\boldsymbol{h}_j|\boldsymbol{x}_j, \boldsymbol{\beta}, \boldsymbol{\gamma}] = P[T_{j,L_j+1} > c_j|T_{j,L_j} = t_{j,L_j}, \boldsymbol{x}_{ij}(t_{ij,L_j}), \boldsymbol{\beta}, \boldsymbol{\gamma}] \times \prod_{l=1}^{L_j} P[T_{j,l} = t_{j,l}, Z_{j,l} = z_{j,l}|T_{j,l-1} = t_{j,l-1}, \boldsymbol{x}_j(t_{j,l-1}), \boldsymbol{\beta}, \boldsymbol{\gamma}].$$

By noting that $d_k(t)$, $k = 1, \ldots, K$, are periodically fixed, we obtain the likelihood contribution from the lth cancer occurrence as follows.

$$P[T_{j,l} = t_{j,l}, Z_{j,l} = z_{j,l}|T_{j,l-1} = t_{j,l-1}, \boldsymbol{x}_j(t_{j,l-1}), \boldsymbol{\beta}, \boldsymbol{\gamma}] = \lambda_{z_{j,l}}(t_{j,l}|\boldsymbol{x}_j(t_{j,l})) \prod_{m=1}^{K} S_m(t_{j,l}|t_{j,l-1}, \boldsymbol{x}_j(t_{j,l-1})), \qquad l = 1, \ldots, L_j, \tag{3}$$

with $t_{j,0} = 0$. The likelihood contribution from the censored event is given by

$$P[T_{j,L_j+1} > c_j|T_{j,L_j} = t_{j,L_j}, \boldsymbol{x}_j(t_{j,L_j}), \boldsymbol{\beta}, \boldsymbol{\gamma}] = \prod_{k=1}^{K} S_k(c_j|t_{j,L_j}, \boldsymbol{x}_j(t_{j,L_j}))$$

Finally, the likelihood for the jth individual is

$$P[\boldsymbol{h}_j|\boldsymbol{x}_j, \boldsymbol{\beta}, \boldsymbol{\gamma}] = \prod_{k=1}^{K} S_k(c_j|t_{j,L_j}, \boldsymbol{x}_j(t_{j,L_j})) \times \left\{ \prod_{l=1}^{L_j} \lambda_{z_{j,l}}(t_{j,l}|\boldsymbol{x}_j(t_{j,l})) \prod_{m=1}^{K} S_m(t_{j,l}|t_{j,l-1}, \boldsymbol{x}_j(t_{j,l-1})) \right\}. \tag{4}$$

Since we assume the individuals are independent, the overall likelihood of the dataset is simply given by the product of all the individual likelihoods. The complete derivations of the equations above are relegated to **Supplementary Material Section A**.

2.4 Family-wise likelihood with unknown G

For inherited cancer syndromes in the general population, the total number of carriers is very low, e.g., ∼1 out of 1,300 for *TP53* mutations (Gao et al., 2020). This rare condition means even one of the most extensively collected datasets would still have just a few hundred mutation carriers ($n = 335$, see **Table 1**) to be used for estimating a series of penetrance curves, in order to capture dynamic cancer outcomes within the first and the second primary cancers (referred to as SPC and FPC from now on). On the other hand, data collection for inherited cancer syndromes, such as LFS, usually includes cancer history information from family members. Accordingly, we have data from an additional tens of thousands of family members (e.g., $n = 10,270$, see **Table 1**), who have all the required information except for mutation status. Many of these family members can carry the mutation due to their genetic relationship with the carriers. Therefore, we introduce a model that includes all family members, including those with unknown G.

2.4.1 Frailty modeling for the family units

Let $i = \{1, \ldots, I\}$ denote the families. In order to account for often-observed non-genetic correlations between family members, after conditioning on $\boldsymbol{X}(t)$, as a random effect, we introduce a frailty term (Hougaard, 1995), denoted by $\xi_{i,k}$, into the modeling of the cancer-specific intensities (2). Importantly, this family-wise random effect may vary across families and across different cancer types. Hence we have

$$\lambda_k(t|\xi_{i,k}, \boldsymbol{X}(t)) = \xi_{i,k}\lambda_{0,k}(t)\exp(\boldsymbol{\beta}_k^T\boldsymbol{X}(t)), \qquad k = 1, \ldots, K. \tag{5}$$

For each cancer type k, we assume the frailty term to follow a gamma distribution with the same shape and scale parameters, i.e., $\xi_{1,k}, \ldots, \xi_{I,k} \overset{iid}{\sim} Gamma(\phi_k, \phi_k)$. With the pres-

ence of the frailties, the age-at-onset penetrance defined in eq.(1) is further modified by marginalizing over the frailties $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_K\}$. For $t > u$, we then have

$$q_{kl}(t|u, \boldsymbol{X}(u)) = \int_u^t \lambda_{0,k}(v) \exp(\boldsymbol{\beta}_k^T \boldsymbol{X}(u)) R_k(v|u, \boldsymbol{X}(u)) \prod_{a=1}^{K} R_a(v|u, \boldsymbol{X}(u))^{\phi_a} \, dv, \qquad (6)$$

where $R_k(t|u, \boldsymbol{X}(u)) = \phi_k / \big\{\phi_k + \int_u^t \lambda_{0,k}(v) \exp(\boldsymbol{\beta}_k^T \boldsymbol{X}(u)) \, dv\big\}$.

2.4.2 Peeling algorithm to model the Mendelian inheritance

Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$ be the full set of model parameters. The likelihood contribution of the jth individual in the ith family, which we call family-wise likelihood, can be derived similarly to (4), when the covariate vector $\boldsymbol{x_{ij}}$ is fully observed for all family member j in family i:

$$P[\boldsymbol{h}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}, \boldsymbol{\xi}_i] = \prod_{k=1}^{K} S_k(c_{ij}|\xi_{i,k}, t_{ij,L_{ij}}, \boldsymbol{x}_{ij}(t_{ij,L_{ij}})) \times \left\{ \prod_{l=1}^{L_{ij}} \lambda_{z_{ij,l}}(t_{ij,l}|\xi_{i,z_{ij,l}}, \boldsymbol{x}_{ij}(t_{ij,l})) \prod_{n=1}^{K} S_n(t_{ij,l}|\xi_{i,n}, t_{ij,l-1}, \boldsymbol{x}_{ij}(t_{ij,l-1})) \right\}. \qquad (7)$$

Let $\boldsymbol{x}_i = \{\boldsymbol{x}_{ij} : j = 1, \ldots, n_i\}$ and $\boldsymbol{h}_i = \{\boldsymbol{h}_{ij} : j = 1, \ldots, n_i\}$ be the aggregated covariate vector and cancer history across all family members of the ith family, with $\boldsymbol{h}_{ij} = \{\boldsymbol{t}_{ij}, \boldsymbol{z}_{ij}, c_{ij}\}$. Direct evaluation of (7), however, is not possible in our application since the *TP53* mutation status is unknown for most family members.

To tackle this, we employ the peeling algorithm (Elston and Stewart, 1971) as described in the following. Let $\boldsymbol{g}_i = \{g_{ij} : j = 1, \ldots, n_i\}$ be the set of genotype information. We then partition the covariate vector into genotype and other non-genetic covariates, hence $\boldsymbol{x}_i = \boldsymbol{g}_i \cup \boldsymbol{g}_i^C$. Let $\boldsymbol{g}_{i,obs} = \{g_{ij} : g_{ij} \text{ is known}\}$ be the observed part and $\boldsymbol{g}_{i,mis}$ the missing part of the genotype data for the ith family. We then have $\boldsymbol{g}_i =$

$\boldsymbol{g}_{i,mis} \cup \boldsymbol{g}_{i,obs}$. We sum over all possible genotype configurations for the family members who have missing genotype information to calculate the family-wise likelihood as $P[\boldsymbol{h}_i|\boldsymbol{g}_{i,obs}] = \sum_{g_{i,mis}} P[\boldsymbol{h}_i|\boldsymbol{g}_{i,mis}, \boldsymbol{g}_{i,obs}]P[\boldsymbol{g}_{i,mis}|\boldsymbol{g}_{i,obs}]$. Although this summation procedure seems computationally intractable as the family size increases, the peeling algorithm can solve it very efficiently(Shin et al., 2019, 2020). The details of the implementation of the peeling algorithm are provided in **Supplementary Material Section B**.

2.4.3 Ascertainment bias correction for families with LFS

The ascertainment bias is inevitable and its correction is essential in studying rare diseases like LFS since the data can only be collected from high-risk populations. Introducing an ascertainment indicator for the ith family $\mathcal{A}_i$ that takes 1 if the ith family is ascertained and 0 otherwise, the ascertainment-corrected joint (ACJ) likelihood (Iversen and Chen, 2005) is

$$P[\boldsymbol{h}_i, \boldsymbol{g}_{i,obs}|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i, \mathcal{A}_i = 1] = \frac{P[\mathcal{A}_i = 1|\boldsymbol{h}_i, \boldsymbol{g}_{i,obs}, \boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]P[\boldsymbol{h}_i|\boldsymbol{g}_{i,obs}, \boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]P[\boldsymbol{g}_{i,obs}|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]}{P[\mathcal{A}_i = 1|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]}.$$

In practice, the ascertainment decision is often made based on the phenotype of the proband in a deterministic way. If so, the first term in the numerator, $P[\mathcal{A}_i = 1|\boldsymbol{h}_i, \boldsymbol{g}_{i,obs}, \boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]$ is reduced to $P[\mathcal{A}_i = 1|\boldsymbol{h}_{i,1}]$ which is independent of the model parameters. In addition, it is not practical for any model to predict genotype data from non-genetic covariates, hence we can reasonably assume that $P[\boldsymbol{g}_{i,obs}|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]$ is simplified to $P[\boldsymbol{g}_{i,obs}]$. We then have the following ACJ likelihood up to a constant of proportionality

$$P[\boldsymbol{h}_i, \boldsymbol{g}_{i,obs}|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i, \mathcal{A}_i = 1] \propto \frac{P[\boldsymbol{h}_i|\boldsymbol{g}_{i,obs}, \boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]}{P[\mathcal{A}_i = 1|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]}.$$

Note that the numerator is the likelihood contribution of the ith family without considering the ascertainment bias, which can be obtained by the peeling algorithm as described in Section 2.4.2. For most LFS studies, the probands can be ascertained due to the diagnosis of a variety of cancer types, in particular those that belong to the LFS spectrum. The ascertainment probability for LFS studies, where the ascertainment decision is based on the first primary cancer diagnosis of the proband, is given by

$$P[\mathcal{A}_i = 1|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i] = \sum_{k=1}^{K} P[\mathcal{A}_i|T_{ip,1} = t_{ip,1}, Z_{ip,1} = k]P[T_{ip,1} = t_{ip,1}, Z_{ip,1} = k|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i],$$

where the subscript p denotes the proband. We are interested in the ACJ likelihood up to proportionality, hence only the ratio between the terms $P[\mathcal{A}_i|T_{ip,1} = t_{ip,1}, Z_{ip,1} = k]$, $k = 1, \ldots, K$, matters. Denoting these terms by $a_1, \ldots, a_K$, where $a_K = 1$, it follows that

$$\begin{aligned} P[\mathcal{A}_i = 1|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i] &\propto \sum_{k=1}^{K} a_k P[T_{ip,1} = t_{ip,1}, Z_{ip,1} = k|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i] \\ &= \sum_{k=1}^{K} \sum_{g_{ip} \in \{0,1\}} a_k P[T_{ip,1} = t_{ip,1}, Z_{ip,1} = k|g_{ip}, \boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i] P[g_{ip}|\boldsymbol{g}_i^C]. \end{aligned}$$

In our application, the *TP53* mutation prevalence is known to be independent of the non-genetic covariates which yields $P[g_{ip}|\boldsymbol{g}_i^C] = P[g_{ip}]$. This probability can be calculated readily from the mutated allele frequency κ_A: $P[G_{ip} = 0] = (1 - \kappa_A)^2$, and $P[G_{ip} = 1] = 1 - (1 - \kappa_A)^2$. In the Western population, we have $\kappa_A = 0.0006$ (Lalloo et al., 2003). In Section 4, we will show that our model is not sensitive to the choice of a_k, $k = 1, \ldots, K$.

We denote the overall likelihood of the dataset by $P[\boldsymbol{h}, \boldsymbol{g}_{obs}|\boldsymbol{g}^C, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{A}]$, where the family subscript i is dropped to represent aggregation across all the I families (e.g., $\boldsymbol{h} = \{\boldsymbol{h}_i : i = 1, \ldots, I\}$). Once the ACJ likelihood has been obtained for each family, the overall likelihood

can easily be evaluated as a product as follows, since the families are independent

$$P[\boldsymbol{h}, \boldsymbol{g}_{obs}|\boldsymbol{g}^C, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{A}] \propto \prod_{i=1}^{I} \frac{P[\boldsymbol{h}_i|\boldsymbol{g}_{i,obs}, \boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]}{P[\mathcal{A}_i = 1|\boldsymbol{g}_i^C, \boldsymbol{\theta}, \boldsymbol{\xi}_i]}.$$

2.5 Markov chain Monte Carlo (MCMC) estimation

We assume a normal prior distribution for $\boldsymbol{\beta}_k$: $\boldsymbol{\beta}_k \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$, $k = 1, \ldots, K$, where $\mathbf{0}$ and $\boldsymbol{I}$ denote the zero vector and the identity matrix respectively, and σ is set to be large enough (e.g., 100) to reflect little prior knowledge about $\boldsymbol{\beta}_k$. For the baseline intensity function, we set $\gamma_{m,k} \sim Gamma(0.01, 0.01)$, $m = 1, \ldots, M$ and $k = 1, \ldots, K$. This distribution is non-negative as required of $\gamma_{m,k}$, and the corresponding variance is 100 to reflect little prior knowledge. We assume the same gamma prior for ϕ_k, $k = 1, \ldots, K$. It has been shown previously that the penetrance estimates for real data with *TP53* mutations are not particularly sensitive to the choice of parameter for the gamma prior through a sensitivity analysis (Shin et al., 2020). We denote the priors of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\phi}$ by $P[\boldsymbol{\beta}]$, $P[\boldsymbol{\gamma}]$, and $P[\boldsymbol{\phi}]$, respectively. Then, the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ is given by

$$P[\boldsymbol{\theta}, \boldsymbol{\xi}|\boldsymbol{h}, \boldsymbol{g}_{obs}, \boldsymbol{g}^C, \mathcal{A}] \propto P[\boldsymbol{h}, \boldsymbol{g}_{obs}|\boldsymbol{g}^C, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathcal{A}] \cdot P[\boldsymbol{\beta}].P[\boldsymbol{\gamma}] \cdot P[\boldsymbol{\xi}|\boldsymbol{\phi}] \cdot P[\boldsymbol{\phi}]$$

We use a random walk Metropolis-Hastings-within-Gibbs algorithm to generate posterior samples, of which the first part is discarded as burn-in. The MCMC is coded in `R` with the peeling algorithm implemented in `C++` and linked via the `Rcpp` library.

2.6 Simulation study

To test the individual likelihood model, we simulated 20 datasets, each consisting of 10,000 individuals with known genotypes. We generated the cancer outcome data for these in-

dividuals based on a given set of model parameters with $K = 2$. The MCMC samples are converged well and the corresponding 95% credible intervals cover corresponding true parameters In addition, the estimates show promising performance with low bias and mean squared errors. We refer to **Supplementary Materials Section C** for complete details about the simulation for the individual likelihood model.

For the family-wise likelihood model in Section 2.4, we simulated 20 datasets, each consisting of 300 families with 30 family members spanning three generations. We generated the mutation and the cancer outcome data for all family members, as well as an ascertainment scheme to collect families that meet certain criteria until the total number reaches 300. We include the frailty term and apply the ACJ likelihood for the bias correction. Similar to the individual likelihood case, MCMC estimates converge very well with promising performance. The $\boldsymbol{\beta^G}$ parameters showed the highest absolute biases, 0.35 for $\beta_1^G = 3$ and 0.36 for $\beta_2^G = 2$. Considering the high complexity of the family-based likelihood model, we are satisfied with its overall performance. **Supplementary Materials Section D** provides full details about the simulation for the family-wise likelihood model.

3 Application of the individual likelihood

We apply our models to the MDACC patient cohort as introduced in Section 1. We refer to **Table E.1** (**Supplementary Materials Section E**) for more details about about the first and second primary cancers in the dataset. Since the positive families have characteristics that are distinct from the negative ones (e.g., more mutation carriers, more family members with SPC), we put these two types of families into two separate strata. For the cross-validation study, we split the full dataset, randomly with each strata, into two subsets of equal size to have training and validation datasets comparable in cancer-type distributions.

See **Table E.2** (**Supplementary Materials Section E**).

To apply the individual likelihood model to individuals with known G in the training set, we focused on all family members who were tested for *TP53*, plus those with a predicted mutation probability greater than 0.9 (inferred mutation carriers) or less than 0.01 (inferred non-carriers). These probabilities were calculated using the peeling algorithm as described in Section 2.4.2. To utilize the individual likelihood-based penetrance estimates in the validation set, we expanded our validation subjects to include family members with a predicted mutation probability greater than 0.9 or less than 0.01. Related summaries of the training and validation sets for the individual likelihood model are provided in **Tables E.3 and E.4** (**Supplementary Materials Section E**).

We also compare our penetrance estimates with those that are imputed from the most recent SEER Research Plus database (2022). The database provides access to incidence data across eight registries in the US from 1975 to 2019. See **Table E.5** (**Supplementary Materials Section E**) for the summary of the SEER data.

3.1 Model specifications

We choose three individual cancer-type models to demonstrate utility, with each focused on a different cancer type group among the second primary cancers: i) breast cancer, ii) sarcoma, and iii) lung cancer. Here we group all the other cancer types in contrast with the chosen one to maximize sample size and reduce the number of parameters to be estimated. Our model design is based on the frequency of cancer types among SPC (**Tables E.1 - E.4, Supplementary Materials Section E**) as well as the potential relevance to public health (e.g., lung cancer screening). In each model, we categorize the outcomes as breast cancer (or sarcoma, lung cancer), other cancers, and death. For clarity of notations, we let $k \in \{br/sa/lu, ot, d\}$, instead of integers, where br, sa, lu, ot, and d

refer to breast cancer, sarcoma, lung cancer, other cancers, and death respectively. For the latter two groups, the intensity/hazard functions take the form given by Equation (4), where $\boldsymbol{X}(t) = \{G, S, D(t)\}^T$, and $D(t)$ is an indicator of previous cancer occurrences (i.e., $D(t) = 1$ if the person has had at least one cancer in the past, and $D(t) = 0$ otherwise). Here we use a common indicator to alleviate limited data availability, but cancer-specific indicators, as used in the simulation study, is still applicable to obtain insights on interactions between cancer types given enough data. Given this covariate vector $\boldsymbol{X}(t)$, we write $\boldsymbol{\beta}_k = (\beta_k^G, \beta_k^S, \beta_k^D)^T$ for the regression coefficients in the intensity of cancer type k. Our dataset has no male patients diagnosed with breast cancer. Hence we restrict the intensity of breast cancer to 0 for male as $\lambda_{br}(t|\boldsymbol{X}(t)) = \lambda_{0,br}(t)\exp(\boldsymbol{\beta}_{br}^T\boldsymbol{X}(t))I\{S = 0\}$. The sarcoma model categorizes the outcomes into sarcoma, other cancers, and death (i.e., $k \in \{sa, ot, d\}$). Lastly, the lung cancer model considers lung cancer, other cancers, and death (i.e., $k \in \{lu, ot, d\}$). For both models, we do not impose any restrictions such as the indicator function above on their intensity functions.

3.2 Model estimation

Using the random-walk Metropolis-Hastings-within-Gibbs algorithm, we obtain 10,000 posterior samples, of which the first 5,000 are discarded as burn-in. Each MCMC iteration takes about two minutes on a high-performance computing cluster at MDACC (407 nodes, each of which has 28 CPU cores with Intel Skylake 2.6Ghz). Posterior samples for all model parameters converged well within the first 10,000 iterations (**Figures F.1-F.3, Supplementary Material Section F**).

Table F.4 (**Supplementary Material Section F**) shows the estimated parameters. We observe that β_{br}^G, β_{sa}^G, and β_{lu}^G are significantly positive across all models as expected, since breast cancer, sarcoma, and lung cancer all belong to the LFS spectrum. The pa-

rameter βde^D is also significantly positive in all three models as patients with one cancer in the past are at higher risk of mortality. In the lung cancer model, it is interesting to see that β_{ot}^S is significantly negative, suggesting that females are more likely to develop cancers outside of the lung. In light of our competing risk framework, it may be due to the prevalence of breast cancer, which is exclusive to females in our training data, within the group of other cancer types. This hypothesis is further confirmed by observing that β_{sa}^S in the sarcoma model is not significant, indicating that sarcoma, the second most popular cancer type in LFS, does not present gender disparity.

3.3 Age-at-onset penetrance

For each of the three models in Section 3.1, our primary goal is to estimate the cancer-specific age-at-onset penetrances. Using the notations introduced in Section 2, for a patient with *TP53* mutation status g and sex s, the set of cancer-specific penetrances to the first primary is denoted by $q_{k1}(t|0, \boldsymbol{X})$, $k = 1, 2, 3$, where $\boldsymbol{X} = \{g, s, 0\}^T$. For a patient with a first primary of any type at age 10, the set of cancer-specific penetrances to the second primary is given by $q_{k2}(t|10, \boldsymbol{X})$, $k = 1, 2, 3$, where $\boldsymbol{X} = \{g, s, 1\}^T$. While we can construct penetrance curves for other cancers and death, the primary purpose of the Breast Cancer model is to estimate penetrance for breast cancer (e.g., $q_{br,1}(t|0, \boldsymbol{X})$ and $q_{br,2}(t|10, \boldsymbol{X})$), which is of clinical relevance. Similarly, we estimate the age-at-onset penetrances to the first and second primary for sarcoma, and lung cancer, respectively.

The estimated age-at-onset penetrance to the first primary is presented in **Figure G.1** (**Supplementary Materials Section G**). For comparison, we also impute probability density curves using the SEER incidence rates (2022). Given the rare prevalence of *TP53* mutations, one would expect the SEER estimates, which are computed from the general population, to closely match our model-based penetrance estimates for individuals who

carry wildtype *TP53*.

The major contribution of our approach, however, is the set of covariate-adjusted penetrance estimates for the second primary cancer, which has never been studied in the literature. **Figure 1** shows the corresponding cancer-specific penetrance estimates to the second primary, given a first primary cancer of any type at age 10, from the three individual likelihood models. In contrast, to produce the SEER estimates that match the covariate values, we have to perform the following steps manually. We first select SEER individuals who were diagnosed with first primary cancers of any type at age 10-14 during the period 1975-1989. We then prospectively follow these individuals until 2019 to monitor the occurrences of second primary cancers. The cancer-specific incidence rates at gap times of 5, 10, 15, 20, 25, and 30 years are calculated as the proportions of these individuals who further developed the corresponding cancer type at ages 15-19, 20-24, 25-29, 30-34, 35-39, and 40-44 respectively. In **Figure 1**, our penetrance estimates for individuals with wildtype *TP53* approximates the SEER estimates reasonably well in breast cancer and sarcoma, less well in lung cancer, which has much lower penetrance among SPCs as compared to the other two cancer types. The penetrance estimates of breast cancer and sarcoma are much higher for mutation carriers compared to wildtypes as expected. The penetrance estimates of lung cancer are also higher for mutation carriers, but the difference is much smaller. Although lung cancer belongs to the LFS spectrum, this small difference is not surprising given the very strong competing risks from breast cancer and sarcoma (i.e., mutation carriers are much more likely to develop these two cancer types before lung cancer). While lung cancer has been extensively studied by medical researchers, our study is the first to report age-at-onset penetrance estimates for this cancer type among SPCs. In **Figures G.2-G.3** (**Supplementary Materials Section G**), we provide additional penetrance estimates for

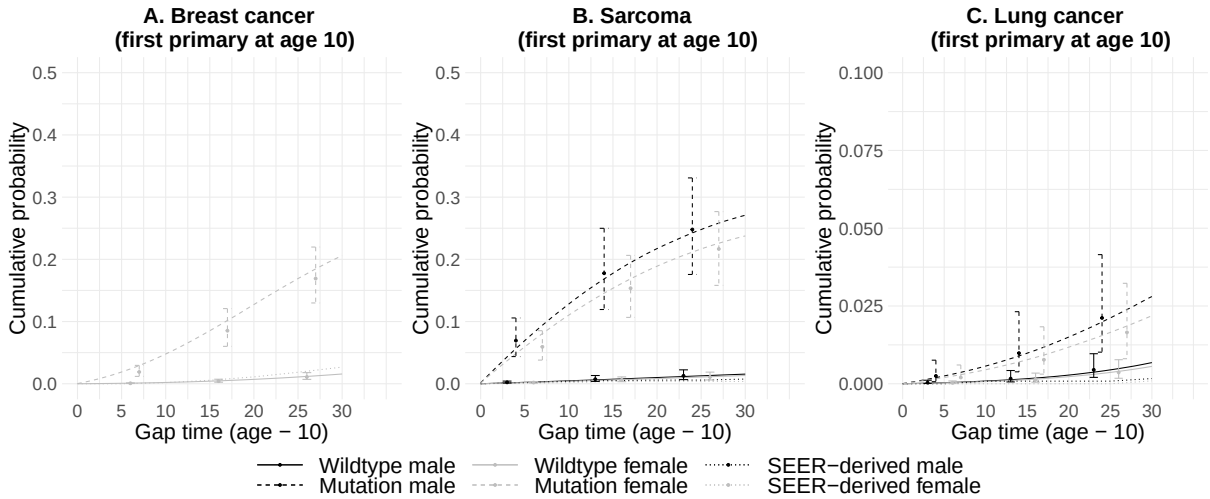
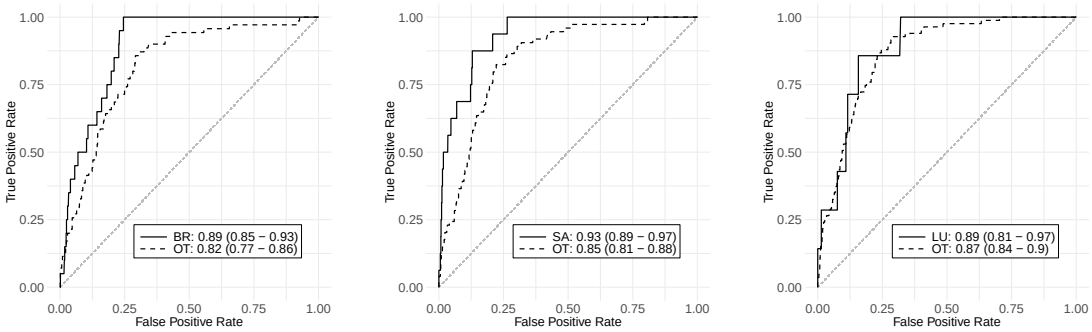


Figure 1: Individual likelihood models: estimates of cancer-specific age-at-onset penetrance to the second primary cancer, given the first primary at age 10. The 95% credible intervals are given at gap times of 5, 15, and 25 years. Cumulative cancer-specific incidence rates from SEER are shown for comparison.

two hypothetical patients: one has a first primary cancer at age 20, and the other has a first primary cancer at age 30. These penetrance curves collectively describe how the risk landscapes of patients depend on the ages at diagnosis of their first primary cancers, and demonstrate the utility of our novel statistical framework that formally accounts for such covariates.

3.4 Validation of the penetrance

Let $r_{kl}(t|u, \boldsymbol{X}(u), \boldsymbol{H})$, where $\boldsymbol{X}(u) = \{G, S, D(u)\}^T$ and $\boldsymbol{H}$ is the family history, be the risk of developing cancer of type k as the lth primary by age t given the $(l-1)$th cancer occurrence at age u. That is, $r_{kl}(t|u, \boldsymbol{X}(u), \boldsymbol{H}) = P(T_l \leq t, Z_l = k|T_{l-1} = u, \boldsymbol{X}(u), \boldsymbol{H})$. For patients with known *TP53* mutation status, $\boldsymbol{H}$ can be removed based on the conditional independence assumption. Since we select only family members with measured genotypes, $r_{kl}(t|u, \boldsymbol{X}(u), \boldsymbol{H}) = P(T_l \leq t, Z_l = k|T_{l-1} = u, \boldsymbol{X}(u))$, is equal to the age-at-onset penetrance $q_{kl}(t|u, \boldsymbol{X}(u))$. For each cancer model, we evaluate its prediction performance for (1) the first primary, and (2) the second primary for individuals who already had one primary



(a) Breast Cancer model **(b)** Sarcoma model **(c)** Lung Cancer model

Figure 2: ROC curves, along with the AUCs and their 95% confidence intervals, for cancer-specific prediction of the second primary based on the individual likelihood models. SA = sarcoma, BR = breast cancer, LU = lung cancer, OT = other cancer. Sample size: **(a)** n(BR) = 20, n(OT) = 70; **(b)** n(SA) = 16, n(OT) = 74; **(c)** n(LU) = 7, n(OT) = 83.

cancer. Receiver Operating Characteristic (ROC) curves for the first evaluation are given in **Supplementary Materials Section H**. The second evaluation is our focus, and the results are shown in **Figure 2**. All models displayed good predictive performances, with areas under the ROC curve (AUCs) being between 0.82 and 0.93, and further presenting small 95% bootstrap confidence intervals.

4 Application of the family-wise likelihood

4.1 Model Specifications

Aiming to address the data sparsity issue presented in the application of our individual model to the real dataset, our family-wise model systematically infers missing genotypes and correspondingly includes all family members in the likelihood calculation. Therefore, we fit this model to the full training set across everyone reported. Motivated by both the biologically meaningful outcomes as well as the need for a sufficient sample size for each category, we focus on three groups of cancer diagnosis: (1) sarcoma, (2) breast cancer, and (3) all other cancers combined, which include both LFS-related (i.e., more frequently observed in LFS) and non-LFS malignancies. We include death from other causes as an additional

competing risk. For clarity of notations, we set $k \in \{sa, br, ot, de\}$, for sarcoma, breast cancer, other cancers, and death respectively. We have $\boldsymbol{X}(t) = \{G, S, D_{sa}(t), D_{br}(t), D_{ot}(t)\}^T$ and $\boldsymbol{\beta}_k = (\beta_k^G, \beta_k^S, \beta_k^{D_{sa}}, \beta_k^{D_{br}}, \beta_k^{D_{ot}})^T$ for the regression coefficients in the intensity of cancer type k. For sarcoma, other cancers, and death, the intensity functions take the form given by Equation (5). In this dataset, no male patients were diagnosed with breast cancer. Hence, we impose the following restriction on the intensity function of breast cancer: $\lambda_{br}(t|\xi_{i,br}, \boldsymbol{X}(t)) = \xi_{i,br}\lambda_{0,br}(t)\exp(\boldsymbol{\beta}_{br}^T\boldsymbol{X}(t))I\{S = 0\}$. For the ascertainment bias correction, we set $a_{sa} = 10$, $a_{br} = 2$ and $a_{ot} = 1$ as we observe sarcoma to be more indicative of *TP53* mutation.

4.2 Model estimation

In addition to the settings described in Section 2.5, we set $M = 5$ for the degree of the Bernstein polynomial to ensure sufficient flexibility while keeping the computational cost reasonable. Using the random walk Metropolis-Hastings-within-Gibbs algorithm, we generate 30,000 posterior samples, of which the first 5,000 are discarded. Overall, the estimation is computationally intensive due to a large number of parameters, with each MCMC iteration taking about 45 seconds on the data when both frailty and ascertainment bias correction is applied. Trace plots of the posterior samples confirm good convergence within 30,000 iterations (**Figures I.1-I.4, Supplementary Material Section I**).

Table 2 displays all estimates of the regression coefficients from our model. All estimates are significantly different from zero at 95% significance level, except $\hat{\beta}_{sa}^S$, which agrees with a previous study on sarcoma by Amadeo et al. (2020). The $\hat{\beta}_{sa}^G$, $\hat{\beta}_{br}^G$ and $\hat{\beta}_{ot}^G$ are positive, which is expected, as people with *TP53* mutations are much more susceptible to LFS-related cancer types. Parameters of the form $\beta_p^{D_q}$, $p \in \{sa, br, ot\}$ and $q \in \{sa, br, ot\}$, are positive with $\beta_{br}^{D_{br}}$ being the sole exception, indicating that patients with a first primary

are more likely to develop a second primary later in their lives. A major contributor for this phenomenon is the method of treatment used for the first primary. Radiotherapy and chemotherapy have long been linked to increased risk in cancer survivors (Boice et al., 1985; Berrington de Gonzalez et al., 2011). Bhatia and Sklar (2002) found a strong relationship between radiation-related tumors and the amount of exposure to radiation, and that the second cancers typically develop within the radiation field after a latency period. They also reported some chemotherapeutic agents, such as alkylating agents and topoisomerase II inhibitors, that increase the risk of developing second cancer. Sarcoma is particularly likely to occur in patients who were treated with radiotherapy (Patel, 2000), which may explain the largest effect for the onset of second primary cancers is presented by $\beta_{sa}^{D_{br}}$. Although data on treatment methods were not collected for our cohort, they can readily be incorporated into our model as an additional covariate in future studies.

Estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are provided in **Tables I.5-I.6** (**Supplementary Material Section I**). To evaluate the sensitivity of our model to the weights a_k, we also report the estimates of the regression coefficients when a_{sa}, a_{br} and a_{ot} are set to 4, 2 and 1 respectively (**Table I.7**, **Supplementary Material Section I**). In this case, sarcoma is still more indicative of *TP53* mutation compared to other cancer types, but to a lesser extent. We notice that the estimates do not change much from those in **Table 2**, hence our model is not overly sensitive to the choice of a_k.

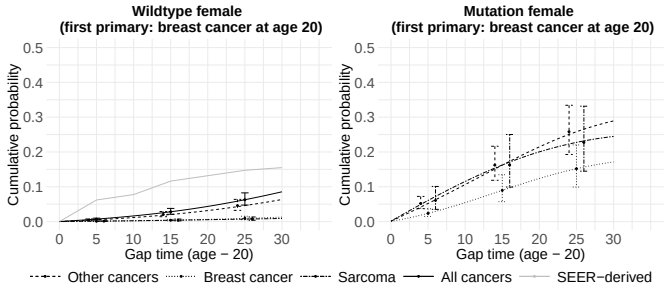
4.3 Age-at-onset penetrance

We present penetrance estimates for FPC and those for SPC with example onsets of the first primary in sarcoma and other cancers in **Tables J.1-J.5** (**Supplementary Materials Section J**). For comparison, we also show the penetrances that are imputed from the incidence rates provided by SEER (2022). Here we focus on discussing the penetrance

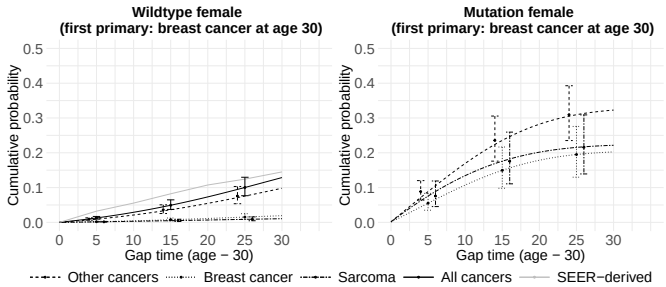
Event	Parameter	Mean	Median	SD	2.5%	97.5%
Sarcoma	β_{sa}^{G}	4.063	4.063	0.223	3.645	4.494
	β_{sa}^{S}	0.426	0.426	0.182	0.073	0.789
	$\beta_{sa}^{D_{sa}}$	1.416	1.417	0.327	0.753	2.043
	$\beta_{sa}^{D_{br}}$	1.491	1.493	0.316	0.877	2.110
	$\beta_{sa}^{D_{ot}}$	1.716	1.717	0.267	1.181	2.232
Breast	β_{br}^{G}	3.516	3.514	0.201	3.118	3.908
	$\beta_{br}^{D_{sa}}$	1.222	1.235	0.408	0.365	1.988
	$\beta_{br}^{D_{br}}$	-0.785	-0.781	0.252	-1.300	-0.314
	$\beta_{br}^{D_{ot}}$	0.861	0.860	0.197	0.472	1.238
Other	β_{ot}^{G}	2.410	2.413	0.137	2.139	2.668
	β_{ot}^{S}	0.392	0.392	0.080	0.239	0.545
	$\beta_{ot}^{D_{sa}}$	1.183	1.186	0.216	0.762	1.596
	$\beta_{ot}^{D_{br}}$	0.641	0.644	0.169	0.295	0.964
	$\beta_{ot}^{D_{ot}}$	0.783	0.783	0.113	0.557	0.993
Death	β_{de}^{G}	0.841	0.836	0.132	0.578	1.102
	β_{de}^{S}	0.327	0.328	0.056	0.219	0.435
	$\beta_{de}^{D_{sa}}$	2.309	2.314	0.157	1.987	2.614
	$\beta_{de}^{D_{br}}$	1.573	1.576	0.112	1.345	1.790
	$\beta_{de}^{D_{ot}}$	2.139	2.138	0.061	2.016	2.259

Table 2: Summary of estimated $\boldsymbol{\beta}_k$, $k \in \{sa, br, ot, de\}$, based on the last 25,000 posterior samples. The weights a_{sa} for sarcoma, a_{br} for breast cancer, and a_{ot} for other cancers are set to 10, 2, and 1 respectively.

estimates for SPC, which have never been reported in the literature. For a female patient with a first primary of breast cancer at age 20, the set of cancer-specific penetrances to the second primary cancer is given by $q_{k2}(t|20, \boldsymbol{X})$, where $\boldsymbol{X} = \{g, 0, 0, 1, 0\}^T$. **Figure 3a** shows the penetrance curves for such a patient with $g = 0$ and $g = 1$. The SEER curve was constructed by the following steps. First we selected individuals who were diagnosed with first primary breast cancers at age 20-24 during the period 1975-1989, and prospectively followed them until 2019. We then computed incidence rates at gap times of 5, 10, 15, 20, 25, 30 years as the proportions of individuals who developed second primary cancers of any type during the corresponding age intervals. Similar to **Figure 3b**, we changed the onset of the first primary breast cancer to age 30. With the females who do not carry *TP53* mutations, the SEER estimate approximates closely for those with FPC at age 30 but is



(a) First primary breast cancer at age 20



(b) First primary breast cancer at age 30

Figure 3: Estimates of cancer-specific age-at-onset penetrance to the second primary cancer, given that the first primary is breast cancer. The 95% credible intervals at gap times of 5, 15, and 25 years are shown. Cumulative incidence rates across all cancer types from SEER are shown for comparison.

much higher than those with FPC at age 20. This highlights an advancement in our model to explicitly adjust for mutation status as a covariate, whereas curves derived from the SEER registry cannot differentiate those with or without *TP53* mutations. For individuals with FPC of breast at age 20, their chances of carrying germline mutations are high. Hence the corresponding SEER curve should approximate more closely to the penetrance estimate for mutation carriers (right panel) than noncarriers (left panel) in **Figure 3a**.

4.4 Validation of the penetrance

A majority of family members of patients with LFS do not undergo genetic testing for various reasons. Our family-wise likelihood is particularly advantageous in this case because it incorporates all family members' cancer history into the computation of the time-to-event coefficients. Let $\boldsymbol{X}_G(u)$ be the set of covariates without the mutation status (i.e.,

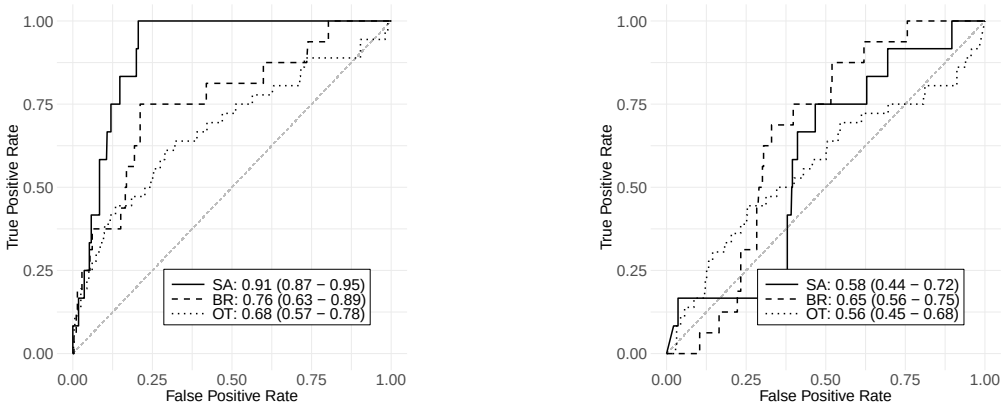
$\boldsymbol{X}_G(u) = \boldsymbol{X}(u) \setminus \{G\}$). By conditioning on the unobserved G, the risk r_{kl} is given by a weighted average: $r_{kl}(t|u, \boldsymbol{X}_G(u), \boldsymbol{H}) = \sum_{G \in \{0,1\}} q_{kl}(t|u, \boldsymbol{X}(u)) P(G|T_{l-1} = u, \boldsymbol{X}_G(u), \boldsymbol{H})$, where the mutation probability $P(G|T_{l-1} = u, \boldsymbol{X}_G(u), \boldsymbol{H})$ is calculated recursively using the peeling algorithm (Elston and Stewart, 1971).

We first evaluate the predictive performance of our joint model in predicting outcomes, such as the *TP53* mutation status, number of primaries, or cancer types within the first primary, which are achievable by previous models (Shin et al. (2019, 2020)). The joint model achieved comparable performances **Supplementary Materials Section K**.

We then move further to predict a unique outcome made available by the joint model, the type of cancer within the second primary. This evaluation is particularly noteworthy as it sets our model apart from others in the literature. Given the limited sample size, we include individuals with mutation probabilities greater than 0.99 as inferred carriers, and smaller than 0.01 as inferred wildtypes. For comparison, we consider a naïve model that assumes independence and identical distributions of the first and second primary. Under this assumption, we can show that the cancer-specific age-at-onset penetrances to the second primary is given by

$$q_{k2}(t|u, \boldsymbol{X}(u)) = \frac{P(T_2 \leq t|T_{l-1} = u, \boldsymbol{X}(u)) P(T_1 \leq t - u, Z_1 = k|\boldsymbol{X}(u))}{P(T_1 \leq t - u|\boldsymbol{X}(u))}.$$

The second term in the numerator is given by a cancer-specific model (Shin et al., 2020), while the other two are given by an MPC model (Shin et al., 2019). We use the penetrance estimates produced by this naive model to perform risk prediction on the validation cohort. **Figure 4** compares the validation results of our model with the naive model. For our model, the AUCs are 0.91, 0.76 and 0.68 for sarcoma, breast cancer, and other cancers, respectively. Furthermore, the ROC curves of the naive model hoover around the diagonal



(a) Cancer-specific model for multiple primary cancer

(b) Naïve model

Figure 4: ROC curve, along with the AUCs and their 95% confidence intervals, for cancer-specific prediction of the second primary in the test dataset. SA = sarcoma, BR = breast cancer, OT = other cancer. Sample size: n(SA) = 12, n(BR) = 16, n(OT) = 36.

line, with AUC not significantly different from 0.5 for sarcoma and other cancer types. This indicates poor predictive performance and further confirms the importance of accounting for the type and age-at-onset for the first primary when predicting the second primary.

5 Conclusion

With the advancements in oncology, cancer patients are living longer, resulting in a sharp increase in the number of patients diagnosed with multiple primary cancer. Our Bayesian semiparametric framework that jointly models both multiple primary cancers and multiple cancer types represent one of the first attempts to build a quantitative foundation for characterizing cancer risk trajectories among cancer survivors. We provide explicit expressions of both an individual likelihood and a family-wise likelihood, allowing us to evaluate cancer risk trajectories among both independent individuals and family members. The individual likelihood can be used for risk characterization of a general population whose cancer history is captured by registries, while family-wise likelihood can be used for a given inherited disease where family history data are routinely collected. With estimated coefficients by the

likelihoods, a suite of penetrance curves can be built, which are cancer-type-specific age-at-onset probabilities for both first and second primary cancers. These probabilities serve as the basis for calculating personalized risk estimates, some of which, for first primary cancer only, are already used in clinical practice to facilitate decision-making (Parmigiani et al., 1998). The joint model presented in this paper has the potential to move the needle in clinical management by expanding the current practice to the population of cancer survivors.

We have validated the performance of our models using both simulation studies and cross-validation for risk prediction in an extensively collected real dataset (429 families ascertained and followed up between 2000 and 2015 at MD Anderson). Using the individual likelihood approach, we accurately characterized the risk of second primary lung cancer (SPLC) among individuals without *TP53* mutation in our study population. Our estimated penetrance curves for SPLC is important for the future update of the USPSTF recommendation for lung cancer screening. Using the family-wise likelihood approach, we have obtained a new set of penetrance estimates for genetic and risk counseling with families with LFS. A constant question during the counseling sessions from these families is when my child will develop next cancer and what it will be. Our framework is general and can accommodate any number of cancer types and the number of primary cancers, as well as additional covariates such as treatment information, given additional data availability. Given enriched data with a long-term follow-up in the future, we can apply our model to reveal a complete picture of the risk landscape in cancer patients.

One limitation of our methods is the relatively high computation cost. Training the family-wise model required about three weeks on a high-performance computing cluster, and training individual likelihood models took even longer. To tackle this issue, We will

explore moving the computation to Stan, a programming language written in `C++` that has been fine-tuned for performance in Bayesian inference.

Acknowledgments

The authors thank Gang Peng for providing the computer code to implement the peeling algorithm.

Funding

Cancer Prevention and Research Institute of Texas [RP200383], National Institutes of Health [R01CA239342, R01CA269696, P30 CA016672], National Research Foundation of Korea [2022M3J6A1063595].

References

(National Cancer Institute, April 2022). "Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Plus Data, 8 Registries, Nov 2021 Sub (1975-2019) - Linked To County Attributes - Total U.S., 1969-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2022, based on the November 2021 submission".

Amadeo, B., Penel, N., Coindre, J.-M., Ray-Coquard, I., Ligier, K., Delafosse, P., et al. (2020). "Incidence and Time Trends of Sarcoma (2000-2013): Results from the French Network of Cancer Registries (FRANCIM)". *BMC Cancer*, 20.

Bao, S., Jiang, M., Wang, X., Hua, Y., Zeng, T., Yang, Y., et al. (2021). "Non-metastatic Breast Cancer Patients Subsequently Developing Second Primary Malignancy: A population-based Study". *Cancer Medicine*, 10(23):8662–8672.

Berrington de Gonzalez, A., Curtis, R. E., Kry, S. F., Gilbert, E., Lamart, S., Berg, C. D., et al. (2011). "Proportion of Second Cancers Attributable to Radiotherapy Treatment in Adults: A Cohort Study in the US SEER cancer Registries". *The Lancet Oncology*, 12(4):353–360.

Bhatia, S. and Sklar, C. (2002). "Second Cancers in Survivors of Childhood Cancer". *Nature Reviews Cancer*, 2(2):124–132.

Boice, John D., J., Day, N. E., Andersen, A., Brinton, L. A., Brown, R., Choi, N. W., et al. (1985). "Second Cancers Following Radiation Treatment for Cervical Cancer. An International Collaboration Among Cancer Registries". *JNCI: Journal of the National Cancer Institute*, 74(5):955–975.

Bougeard, G., Renaux-Petel, M., Flaman, j.-m., Charbonnier, C., Fermey, P., Belotti, M., et al. (2015). "Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers". *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 33.

Chen, L., Hsu, L., and Malone, K. (2009). "A Frailty-Model-Based Approach to Estimating the Age-Dependent Penetrance Function of Candidate Genes Using Population-Based Case-Control Study Designs: An Application to Data on the BRCA1 Gene". *Biometrics*, 65:1105–14.

Chen, S., Wang, W., Lee, S., Nafa, K., Lee, J., Romans, K., et al. (2006). "Prediction of Germline Mutations and Cancer Risk in the Lynch Syndrome". *JAMA*, 296(12):1479–1487.

Choi, Y.-H. (2012). "A Frailty-Model-Based Method for Estimating Age-Dependent Penetrance from Family Data". *Journal of biometrics & biostatistics*, Suppl 4.

Choi, Y.-H., Briollais, L., Win, A., Hopper, J., Buchanan, D., Jenkins, M., and Lakhal-Chaieb, L. (2016). "Modelling of Successive Cancer Risks in Lynch Syndrome Families in the Presence of Competing Risks Using Copulas". *Biometrics*, 73.

Chompret, A., Abel, A., Stoppa-Lyonnet, D., Brugières, L., Pagès, S., Feunteun, J., et al. (2001). "Sensitivity and Predictive Value of Criteria for p53 Germline Mutation Screening". *Journal of Medical Genetics*, 38(1):43–47.

Cook, R. and Lawless, J. (2007). *"The Statistical Analysis of Recurrent Events"*.

Curtis, S. and Ghosh, S. (2011). "A Variable Selection Approach to Monotonic Regression with Bernstein Polynomials". *Journal of Applied Statistics*, 38:961–976.

Donin, N., Filson, C., Drakaki, A., Tan, H.-J., Castillo, A., Kwan, L., et al. (2016). "Risk of Second Primary Malignancies Among Cancer Survivors in the United States, 1992 Through 2008". *Cancer*, 122.

Elston, R. and Stewart, J. (1971). "A General Model for the Genetic Analysis of Pedigree Data". *Human heredity*, 21:523–542.

Forjaz, G., Howlader, N., Scoppa, S., Johnson, C. J., and Mariotto, A. B. (2022). "Impact of Including Second and Later Cancers in Cause-specific Survival Estimates Using Population-based Registry Data". *Cancer*, 128(3):547–557.

Gao, F., Pan, X., Dodd, E., Recio, C., Montierth, M., Bojadzieva, J., et al. (2020). "A Pedigree-based Prediction Model Identifies Carriers of Deleterious de novo Mutations in Families with Li-Fraumeni Syndrome". *Genome Research*, 30.

Hougaard, P. (1995). "Frailty Models for Survival Data". *Lifetime Data Analysis*, 1:255–273.

Hulvat, M. C. (2020). "Cancer Incidence and Trends". *Surgical Clinics of North America*, 100(3):469–481.

Iversen, E. and Chen, S. (2005). "Population-Calibrated Gene Characterization: Estimating Age at Onset Distributions Associated With Cancer Genes". *Journal of the American Statistical Association*, 100:399–409.

Lalloo, F., Varley, J., Ellis, D., Moran, A., O'Dair, L., Pharoah, P., et al. (2003). "Prediction of Pathogenic Mutations in Patients with Early-onset Breast Cancer by Family History". *The Lancet*, 361(9363):1101–1102.

Langbehn, D., Brinkman, R., Falush, D., Paulsen, J., Hayden, M., and on behalf of an International Huntington's Disease Collaborative Group (2004). "A New Model for Prediction of the Age of Onset and Penetrance for Huntington's Disease Based On CAG Length". *Clinical Genetics*, 65(4):267–277.

Li, F. and Fraumeni, J. J. (1969). "Soft-Tissue Sarcomas, Breast Cancer, and Other Neoplasms". *Annals of Internal Medicine*, 71(4):747–752.

Li, F. P., Fraumeni, J. F., Mulvihill, J. J., Blattner, W. A., Dreyfus, M. G., Tucker, M. A., et al. (1988). "A Cancer Family Syndrome in Twenty-four Kindreds". *Cancer Research*, 48(18):5358–5362.

Lorentz, G. (1953). *"Bernstein Polynomials"*. Mathematical expositions. University of Toronto Press.

Mai, P. L., Best, A. F., Peters, J. A., DeCastro, R. M., Khincha, P. P., Loud, J. T., et al. (2016). "Risks of First and Subsequent Cancers Among TP53 Mutation Carriers in the National Cancer Institute Li-Fraumeni Syndrome Cohort". *Cancer*, 122(23):3673–3681.

Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Nelson, C. E., Kim, D. H., et al. (1990). "Germline p53 Mutations in a Familial Syndrome of Breast Cancer, Sarcomas, and Other Neoplasms". *Science*, 250(4985):1233–1238.

Mariotto, A. B., Rowland, J. H., Ries, L. A., Scoppa, S., and Feuer, E. J. (2007). "Multiple Cancer Prevalence: A Growing Challenge in Long-term Survivorship". *Cancer Epidemiology and Prevention Biomarkers*, 16(3):566–571.

Mayer, D. K., Nasso, S. F., and Earp, J. A. (2017). "Defining Cancer Survivors, Their Needs, and Perspectives on Survivorship Health Care in the USA". *The Lancet Oncology*, 18(1):e11–e18.

Nobel, T. B., Carr, R. A., Caso, R., Livschitz, J., Nussenzweig, S., Hsu, M., et al. (2022). "Primary Lung Cancer in Women After Previous Breast Cancer". *BJS Open*, 5(6).

Parmigiani, G., Berry, D. A., and Aguilar, O. (1998). "Determining Carrier Probabilities for Breast Cancer–Susceptibility Genes BRCA1 and BRCA2". *The American Journal of Human Genetics*, 62(1):145–158.

Patel, S. R. (2000). "Radiation-induced Sarcoma". *Current Treatment Options in Oncology*, 1:258–261.

Pruitt, S. L., Tavakkoli, A., Zhu, H., Heitjan, D. F., Gerber, D. E., Singal, A. G., et al. (2022). "Survival of Cancer Survivors with a New Pancreatic Cancer Diagnosis". *Cancer Medicine*.

Shin, S. J., Li, J., Ning, J., Bojadzieva, J., Strong, L. C., and Wang, W. (2020). "Bayesian Estimation of a Semiparametric Recurrent Event Model with Applications to the Pene-

trance Estimation of Multiple Primary Cancers in Li-Fraumeni Syndrome". *Biostatistics*, 21(3):467–482.

Shin, S. J., Yuan, Y., Strong, L. C., Bojadzieva, J., and Wang, W. (2019). "Bayesian Semi-parametric Estimation of Cancer-Specific Age-at-Onset Penetrance with Application to Li-Fraumeni Syndrome". *Journal of the American Statistical Association*, 114(526):541–552.

Travis, L. B., Fosså, S. D., Schonfeld, S. J., McMaster, M. L., Lynch, C. F., Storm, H., et al. (2005). "Second Cancers Among 40,576 Testicular Cancer Patients: Focus on Long-term Survivors". *JNCI: Journal of the National Cancer Institute*, 97(18):1354–1365.

Vogt, A., Schmid, S., Heinimann, K., Frick, H., Herrmann, C., Cerny, T., et al. (2017). "Multiple Primary Tumours: Challenges and Approaches, A Review". *ESMO Open*, 2.

Wang, Y., Li, J., Chang, S., Dong, Y., and Che, G. (2021). "Risk and Influencing Factors for Subsequent Primary Lung Cancer After Treatment of Breast Cancer: A Systematic Review and Two Meta-Analyses Based on Four Million Cases". *Journal of Thoracic Oncology*, 16(11):1893–1908.