Supplementary Materials for

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

Seung Jun Shin, Jasmina Bojadzieva, Louise C. Strong, Wenyi Wang and Ying Yuan

A Illustration of the Peeling Algorithm

For illustration, we consider a hypothetical family of three generations shown in Figure A.1.

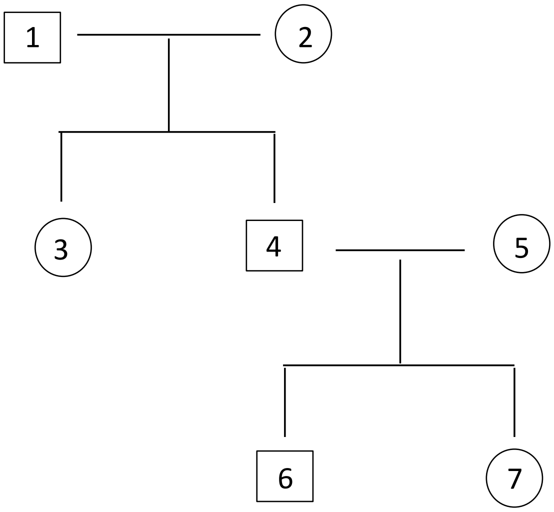


Figure A.1: A hypothetical pedigree for illustration of the Elston-Stewart algorithm for family-wise likelihood evaluation. The circle and square indicate the female and male members, respectively. Genotypes are all unknown except the 1st and 4th individuals.

Without loss of generality, we assume that $\mathbf{G}_{obs}^T = (G_1, G_4)$ and let $\mathbf{G}_{mis}^T = (G_2, G_3, G_5, G_6, G_7)$ and $\mathbf{H}^T = (H_1, \cdots, H_7)$ denote vectors of the unknown genotypes and the cancer history of the family, respectively. The peeling algorithm computes the family-wise likelihood $\Pr(\mathbf{H}|\mathbf{G}_{obs})$ as follows:

$$\Pr(\mathbf{H}|\mathbf{G}_{obs}) = \mathbb{A}_4(G_4|G_1, G_4)\Pr(H_4|G_4)\mathbb{P}_4(G_4|G_1, G_4) \tag{1}$$

Notice that the summation over G_4 in (1) is unnecessary since it is observed. Now, the anterior probability in (1) is

$$\begin{aligned}\mathbb{A}_4(G_4|G_1,G_4) = \mathbb{A}_1(G_1|G_1,G_4)\,\Pr(H_1|G_1) & \\ \times \sum_{G_2} \mathbb{A}_2(G_2|G_1,G_4)\,\Pr(H_2|G_2) \times \Pr(G_4|G_1,G_2,G_4) & \\ \times \sum_{G_3} \mathbb{P}_3(G_3|G_1,G_4)\,\Pr(H_3|G_3) \times \Pr(G_3|G_1,G_2) &\end{aligned}$$

Again, the summation over G_1 does not appear in the above since it is observed. Now, the anterior and posterior probabilities related to $\mathbb{A}_4(G_4|G_1,G_4)$ are computed as follows.

- $\mathbb{A}_1(G_1|G_1,G_4) = \Pr(\mathbf{H}_1^-, G_1|G_1,G_4) = 1$;
- $\mathbb{A}_2(G_2|G_1,G_4) = \Pr(\mathbf{H}_2^-, G_2|G_1,G_4) = \Pr(G_2|G_4)$, since 2 is a founder;
- $\mathbb{P}_3(G_3|G_1,G_4) = \Pr(\mathbf{H}_3^+|G_3,G_1,G_4) = 1$, since 3 is a founder.

Next, the posterior probability in (1) is

$$\begin{aligned}\mathbb{P}_4(G_4|G_1,G_4) = \sum_{G_5} \mathbb{A}_5(G_5|G_1,G_4)\,\Pr(H_5|G_5) & \\ \times \sum_{G_6} \mathbb{P}_6(G_6|G_1,G_4)\,\Pr(H_6|G_6)\,\Pr(G_6|G_4,G_5) & \\ \times \sum_{G_7} \mathbb{P}_7(G_7|G_1,G_4)\,\Pr(H_7|G_7)\,\Pr(G_7|G_4,G_5) &\end{aligned}$$

The anterior and posterior probabilities related to $\mathbb{P}_4(G_4|G_1,G_4)$ are given by

- $\mathbb{A}_5(G_5|G_1,G_4) = \Pr(\mathbf{H}_5^-, G_5|G_1,G_4) = \Pr(G_5)$ (prevalence), since 5 is a founder;
- $\mathbb{P}_6(G_6|G_1,G_4) = \Pr(\mathbf{H}_6^+|G_6,G_1,G_4) = 1$, since $\mathbf{H}_6^+ = \phi$;
- $\mathbb{P}_7(G_7|G_1,G_4) = \Pr(\mathbf{H}_7^+|G_7,G_1,G_4) = 1$, since $\mathbf{H}_7^+ = \phi$.

As a summary, the following table provides all the quantities to be recursively computed during the peeling algorithm.

Target	Step 1 (Pivot)	Step 2 (First-degree relatives)	
$\Pr(\mathbf{H}\|\mathbf{G}_{obs})$	$\mathbb{A}_4(G_4\|\mathbf{G}_{obs})$	$\mathbb{A}_1(G_1\|\mathbf{G}_{obs})$	Father
		$\mathbb{A}_2(G_2\|\mathbf{G}_{obs})$	Mother
		$\mathbb{P}_3(G_3\|\mathbf{G}_{obs})$	Sibling
	$\mathbb{P}_4(G_4\|\mathbf{G}_{obs})$	$\mathbb{A}_5(G_5\|\mathbf{G}_{obs})$	Mate
		$\mathbb{P}_6(G_6\|\mathbf{G}_{obs})$	Offspring 1
		$\mathbb{P}_7(G_7\|\mathbf{G}_{obs})$	Offspring 2

B Bayes-Mendel Model: Estimation of $P(G_j|\mathbf{H})$

We describe the Bayes-Mendel model (Chen et al.; 2004) that enables us to estimate the carrier $P(G_j|\mathbf{H})$ based on family cancer history, $\mathbf{H}$. The Bayes-Mendel model computes the carrier probability $P(G_j|\mathbf{H})$ as follows:

1. Bayesian updating step

$$\Pr(G_j|\mathbf{H}) = \frac{\overbrace{\Pr(G_j)}^{\text{Prevalence}} \Pr(\mathbf{H}|G_j)}{\sum_G \Pr(G) \Pr(\mathbf{H}|G)}$$

2. Integration step

$$\Pr(\mathbf{H}|G_j) = \sum_{\mathbf{G}_{-j}} \left\{ \Pr(\mathbf{H}|G_j, \mathbf{G}_{-j}) \cdot \Pr(\mathbf{G}_{-j}|G_j) \right\} = \sum_{\mathbf{G}_{-j}} \left[\left\{ \prod_{j=1}^{N} \underbrace{\Pr(H_j|G_j)}_{\text{Penetrance}} \right\} \cdot \underbrace{\Pr(\mathbf{G}_{-j}|G_j)}_{\text{Mendelian Prob}} \right]$$

   where $\mathbf{G}_{-j}$ denotes a genotype vector after the jth deleted.

Therefore, the carrier probability can be readily obtained from the prevalence $\Pr(H|G)$ and penetrance $\Pr(G)$ as long as the mode of inheritance is known.

C Baseline Hazard Model Comparison via Simulation

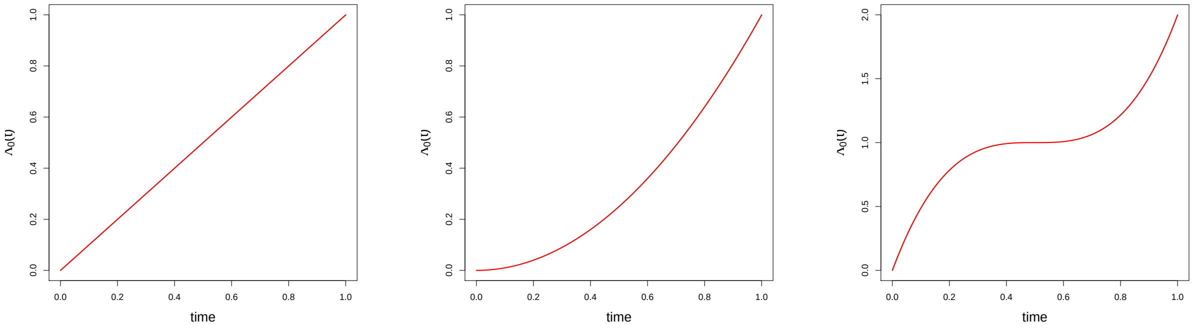
We conduct a simulation study to compare the performance of different baseline hazard models. We consider the conventional proportional hazard models $\lambda_k(t|X)$ for two competing risks:

$$\lambda_k(t|X) = \lambda_k(t)\exp(\beta_k X), k = 1, 2. \tag{2}$$

We first generate binary X_i, taking either 1 or 2 with equal probability. Given X, the time to the kth event, denoted by $T_{i,k}$, can be generated from $S_k(t|X_i) = \exp(-\Lambda_k(u|X_i)du)$. The censored time C_i is independently generated from the exponential distribution. Its rate parameter is chosen so as to achieve a 30% censoring proportion. Then we have $(T_i, D_i), i = 1, \cdots, n$ where $T_i = \min(T_{i,1}, T_{i,2}, C_i)$ and D_i is k if $T_i = T_i, k, k = 1, 2$ and 0 otherwise. We set $\beta_1 = \beta_2 = 1$ for the coefficient parameters for different competing risk models. For the cumulative baseline, $\Lambda_{k,0}(t) = \int_0^t \lambda_{k,0}(s)ds$, we consider the following three cases:

- Case I: $\Lambda_{k,0}(t) = \lambda_k t$ (linear)
- Case II: $\Lambda_{k,0}(t) = \lambda_k t^2$ (quadratic)
- Case III: $\Lambda_{k,0}(t) = \lambda_k((2t-1)^3 + 1)$ (complex)

and set $\lambda_1 = \lambda_2 = 1$. Notice that case I is a constant hazard model that corresponds to exponential survival times. Case II is a quadratic hazard model that corresponds to Weibull survival times. Case III does not satisfy the exponential nor the Weibull survival time assumption. The three true baseline functions are depicted in Figure C.1.



(a) Case 1

(b) Case 2

(c) Case 3

Figure C.1: true cumulative baseline hazard functions

We consider four different models for the baseline hazard functions, $\lambda_k(t), k = 1, 2$ including the exponential model, Weibull model, piecewise constant model, and the proposed Bernstein polynomial-based model. For the piecewise constant and the Bernstein polynomial-based models, the survival time T is rescaled so as to lie on $[0,1]$. For the piecewise constant model, we set 4 equally spaced knots to obtain 5 pieces of equal length. Similarly, we set $M = 5$ for the Bernstein polynomial-based model.

As a performance measure, we consider

$$\widehat{MSE}(\hat{\beta}_k) = \frac{1}{L}\sum_{\ell=1}^{L}(\hat{\beta}_{k,\ell} - \beta_k)^2$$

for the regression coefficient and

$$\widehat{MISE}(\hat{\Lambda}_{k,0}) = \frac{1}{L}\sum_{\ell=1}^{L}\int\Big(\hat{\Lambda}_{k,\ell}(t) - \Lambda_k(t)\Big)^2 dt, \quad k = 1, 2$$

for the baseline hazard. Here the subscript $\ell = 1, \cdots, L$ is used to represent the quantities obtained from the ℓth Monte Carlo (MC)iteration.

Case	k	Exponential	Weibull	Piecewise	Bernstein
I	β_1	0.018 (.006)	0.033 (.013)	0.023 (.007)	0.024 (.007)
	β_2	0.019 (.006)	0.032 (.013)	0.022 (.007)	0.022 (.008)
II	β_1	0.239 (.011)	0.025 (.008)	0.030 (.008)	0.027 (.007)
	β_2	0.248 (.012)	0.024 (.008)	0.034 (.009)	0.029 (.008)
III	β_1	0.229 (.024)	1.152 (.167)	0.033 (.008)	0.024 (.006)
	β_2	0.208 (.025)	1.114 (.169)	0.031 (.009)	0.023 (.007)

Table C.1: MSE of coefficient estimates. MC standard error estimates are in parentheses.

Case	k	Exponential	Weibull	Piecewise	Bernstein
I	$\Lambda_1(t)$	0.011 (.005)	0.020 (.009)	0.425 (.051)	0.206 (.031)
	$\Lambda_2(t)$	0.011 (.005)	0.017 (.008)	0.426 (.051)	0.182 (.032)
II	$\Lambda_1(t)$	0.263 (.035)	0.206 (.031)	0.207 (.038)	0.077 (.018)
	$\Lambda_2(t)$	0.264 (.036)	0.210 (.033)	0.222 (.039)	0.074 (.019)
III	$\Lambda_1(t)$	0.726 (.072)	1.247 (.253)	0.325 (.046)	0.247 (.034)
	$\Lambda_2(t)$	0.697 (.067)	1.160 (.237)	0.357 (.047)	0.226 (.033)

Table C.2: MISE of cumulative baseline hazard function estimates. MC standard error estimates are in parentheses.

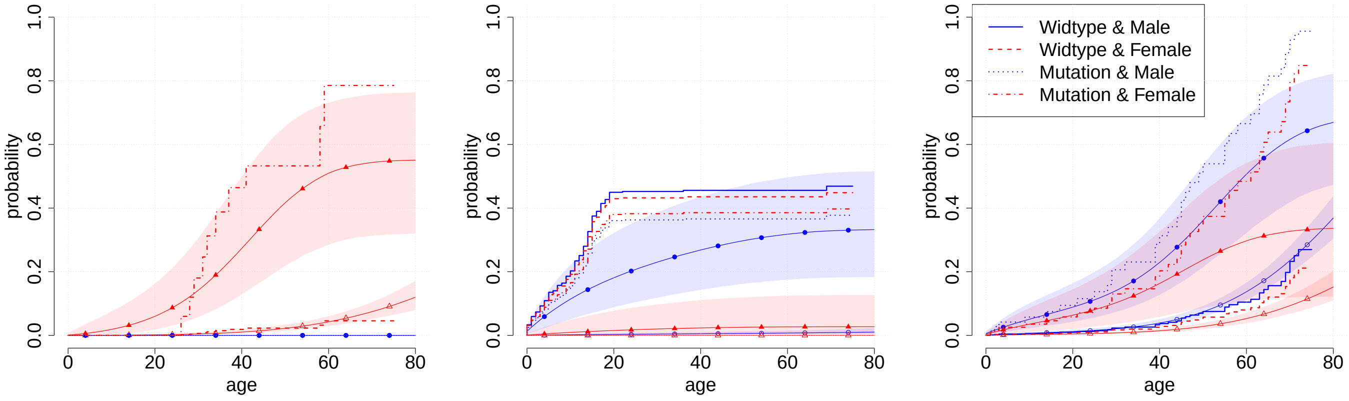
Then, we generate 200 survival times from the competing risk model under the three cases. The results based on $L = 1,000$ independent MC repetitions are summarized in Tables C.1 and C.2. Under case 1, the exponential model performs best with the smallest MSE and MISE because it is correctly specified. Under case 2, the exponential model performs worst with the largest MSE and MISE. Bernstein polynomial and Weibull model are comparable and outperform the piecewise constant model. Under case 3, the exponential model and Weibull model lead to large MSE and MISE. The piecewise constant model performs better than the the exponential model and Weibull model. Bernstein polynomial performs best with the smallest MSE and MISE. Overall, Bernstein polynomial outperforms the other three models and is robust to different shapes of hazard, which partially justifies the use of Bernstein polynomial to model the baseline hazard.

D Comparison of Cancer-Specific Penetrance Estimates of LFS

We compare the penetrance estiamtes of LFS obtained from different models.

D.1 Genotype-Observed Subjects Only

We compare the results to the most naive estimates from the conventional proportional hazard models using the genotype-observed subjects only.

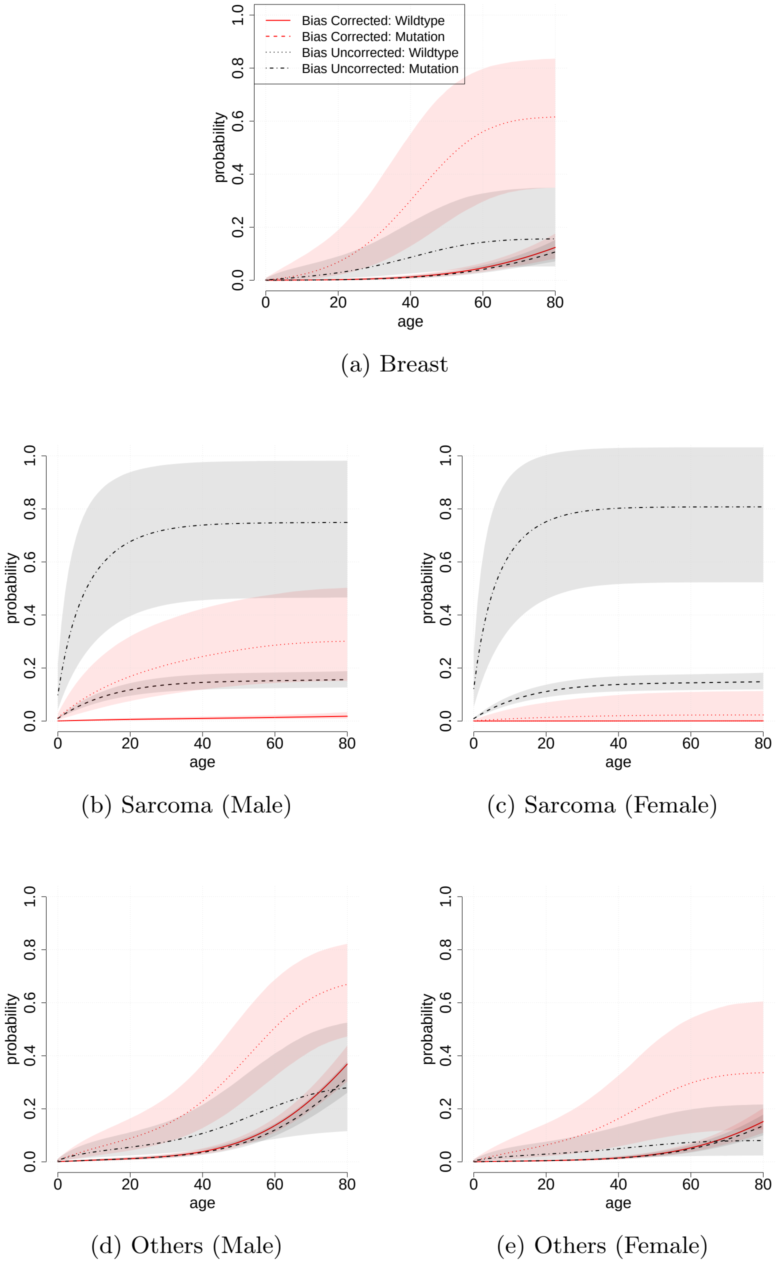


(a) Breast (b) Sarcoma (c) Others

Figure D.1: Comparison to the proportional hazard model based on genotype-observed subjects only

D.2 Without Ascertainment Bias Correction

We compare the results with and without ascertainment bias correction.



(a) Breast

(b) Sarcoma (Male)

(c) Sarcoma (Female)

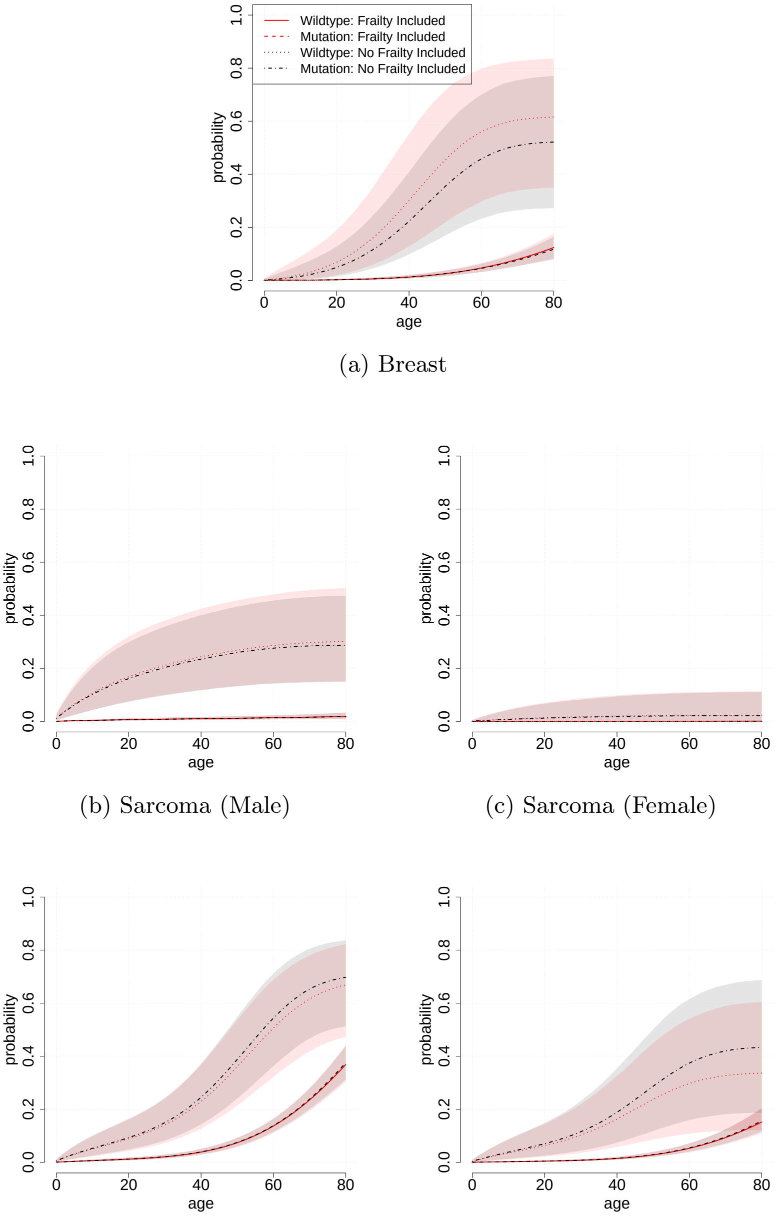
(d) Others (Male)

(e) Others (Female)

Figure D.2: Comparison to the model without ascertainment bias correction

D.3 Without Frailty

We compare the results with and without random frailty



(a) Breast

(b) Sarcoma (Male)

(c) Sarcoma (Female)

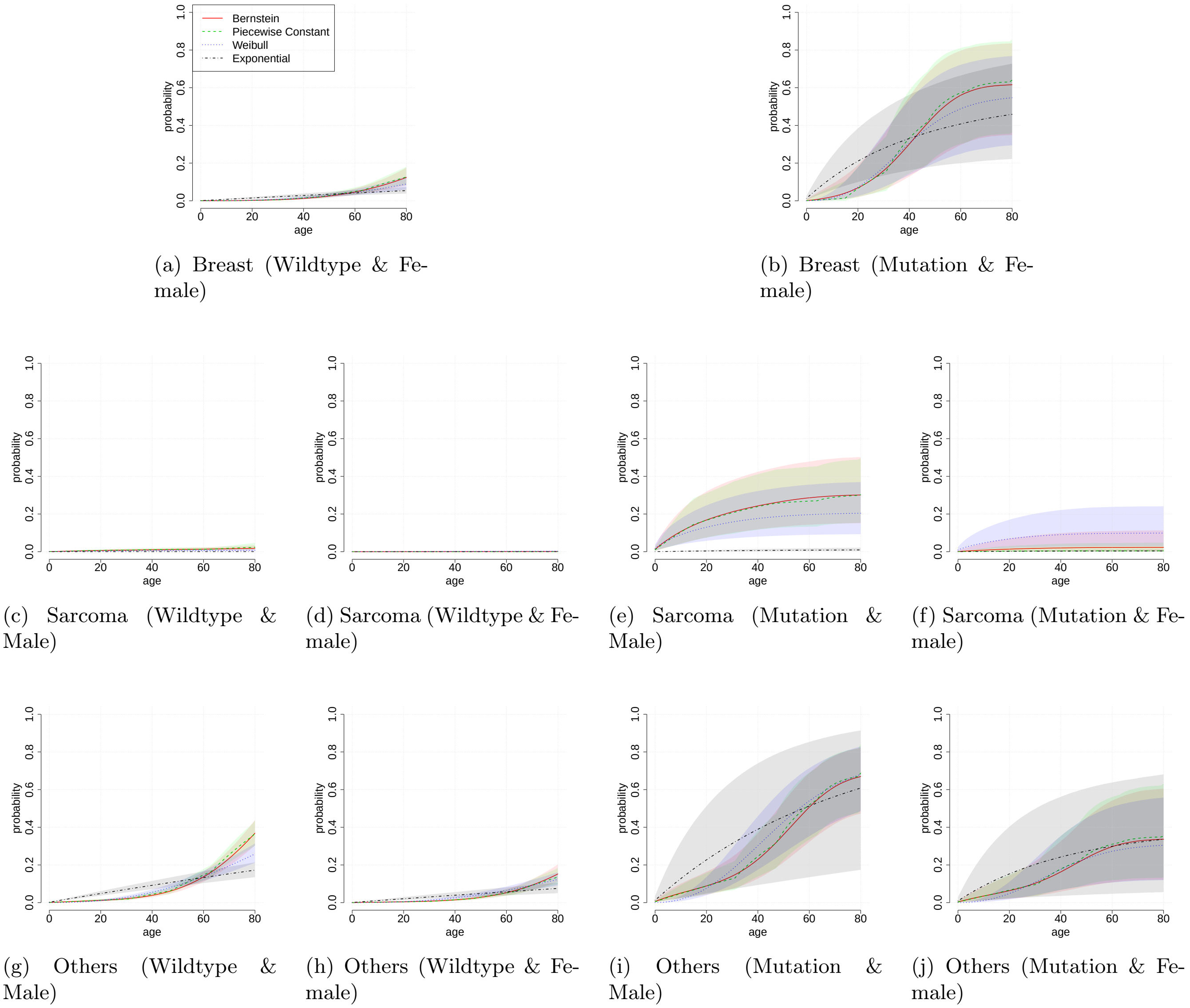
(d) Others (Male)

(e) Others (Female)

Figure D.3: Comparison to the model without frailty

D.4 Different Baseline Hazard Models

We compare the results from the different baseline hazard models used in Section C.



(a) Breast (Wildtype & Female)

(b) Breast (Mutation & Female)

(c) Sarcoma (Wildtype & Male)

(d) Sarcoma (Wildtype & Female)

(e) Sarcoma (Mutation & Male)

(f) Sarcoma (Mutation & Female)

(g) Others (Wildtype & Male)

(h) Others (Wildtype & Female)

(i) Others (Mutation & Male)

(j) Others (Mutation & Female)

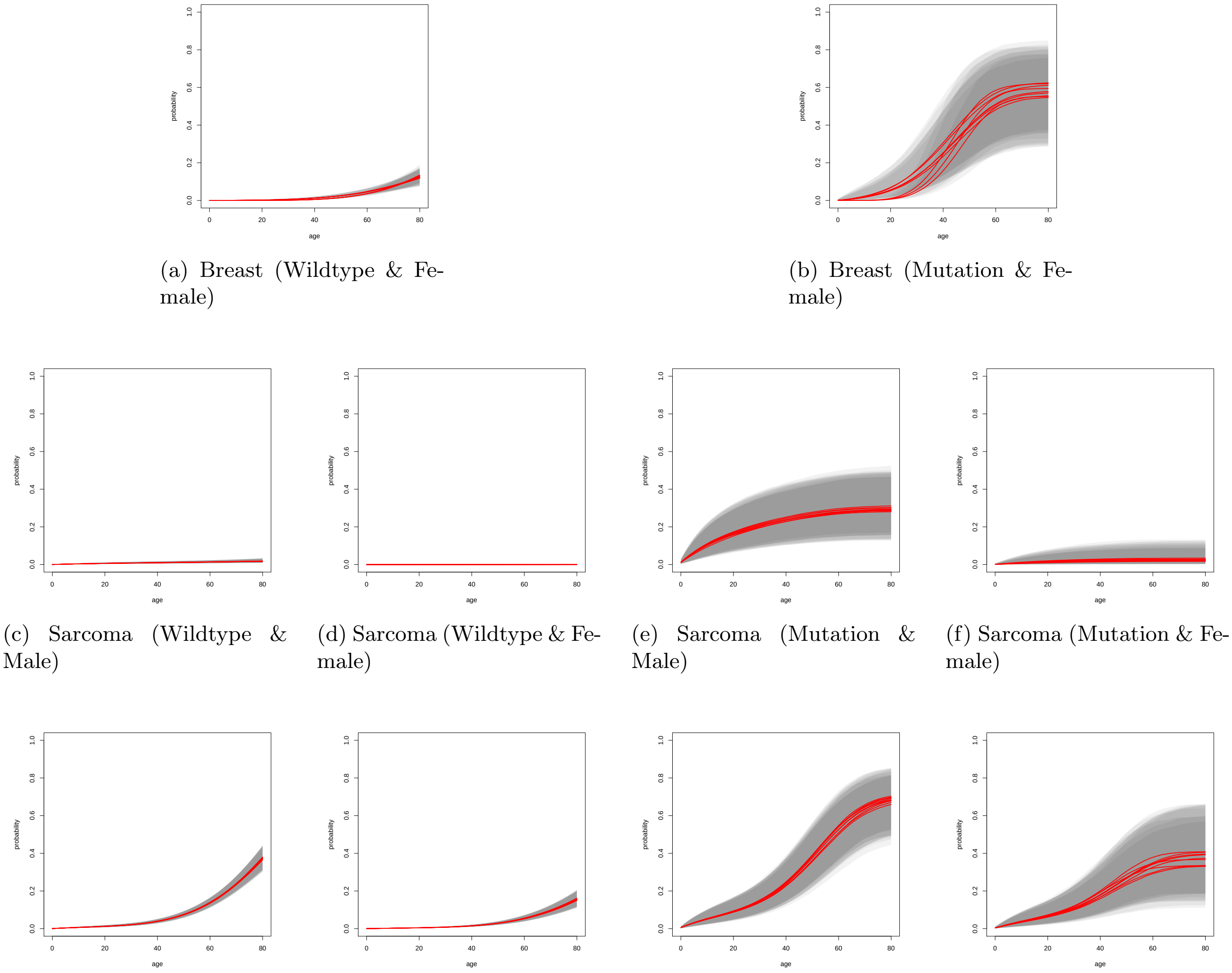
Figure D.4: Different Baseline hazard Models.

E Sensitivity Analysis

We consider the following nine combinations of prior settings.

- $\gamma_k \sim \text{Flat}; Gamma(0.01, 0.01); Gamma(1, 1)$
- $\nu_k \sim Gamma(0.01, 0.01); Gamma(0.1, 0.1); Gamma(1, 1)$.

Figure E.1 depicts penetrance estimates from different prior settings. We observe that the penetrance estimates are not particularly sensitive to the choice of hyperparameters.



(a) Breast (Wildtype & Female)

(b) Breast (Mutation & Female)

(c) Sarcoma (Wildtype & Male)

(d) Sarcoma (Wildtype & Female)

(e) Sarcoma (Mutation & Male)

(f) Sarcoma (Mutation & Female)

(g) Others (Wildtype & Male)

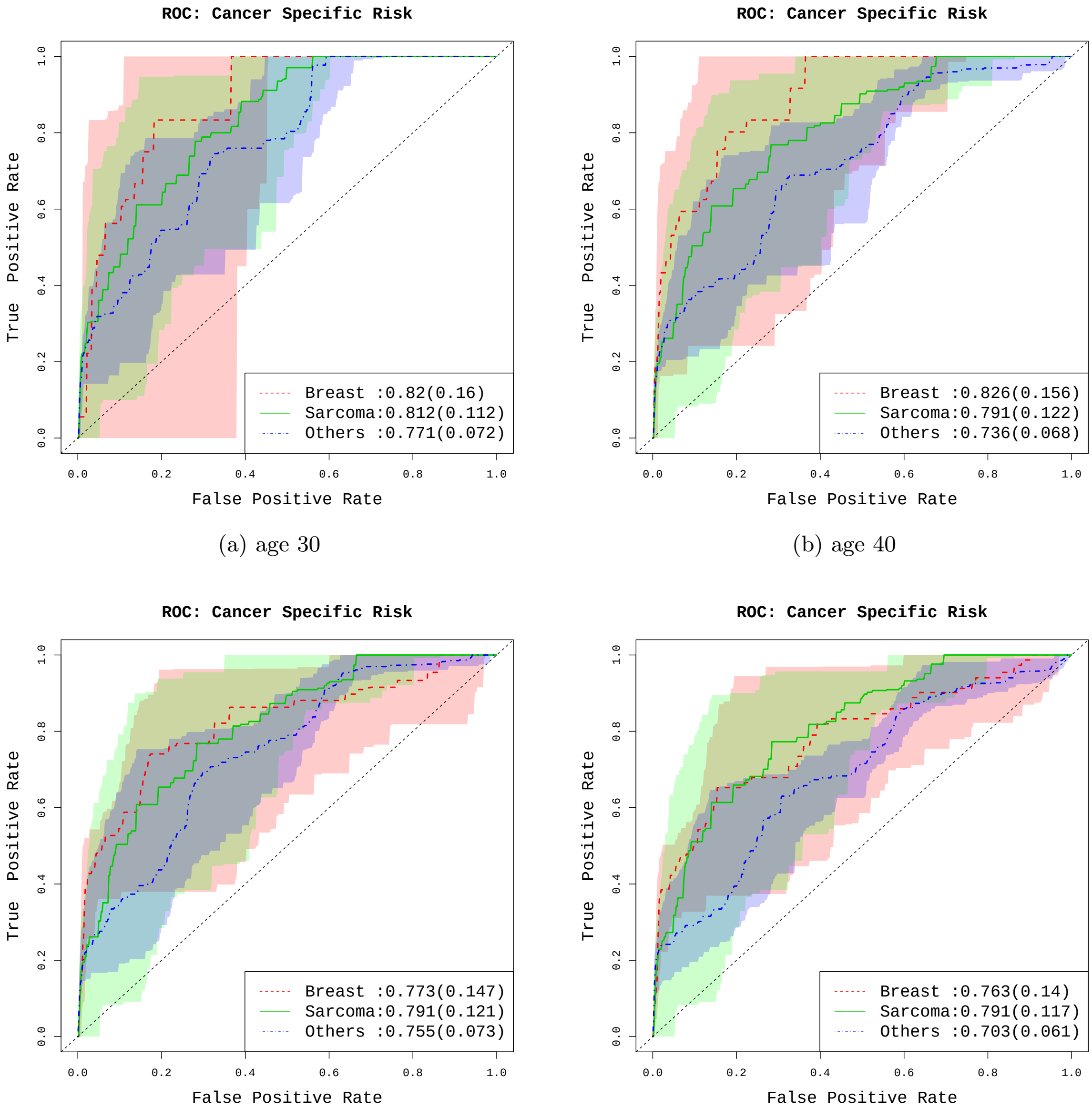
(h) Others (Wildtype & Female)

(i) Others (Mutation & Male)

(j) Others (Mutation & Female)

Figure E.1: Sensitivity analysis

F Cross-validated ROC curves for the cancer-specific risk prediction.



(a) age 30

(b) age 40

(c) age 50

(d) age 60

Figure F.1: Cross validated ROC curves for the cancer-specific risk prediction evaluated at different ages $t_c = 30, 40, 50$ and 60.

References

Chen, S., Wang, W., Broman, K., Katki, H. A. and Parmigiani, G. (2004). Bayesmendel: an r environment for mendelian risk prediction, *Statistical Application in Genetics and Molecular Biology* **3**(1): Article 21.