(2018), **0**, 0, 1–14

Bayesian estimation of a semiparametric recurrent event model with application to the penetrance estimation of multiple primary cancers in Li-Fraumeni syndrome (Supplementary Materials)

SEUNG JUN SHIN

Department of Statistics, Korea University, Seoul, South Korea

JIALU LI

Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A

JING NING

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A

JASMINA BOJADZIEVA, LOUISE C. STRONG

Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A

WENYI WANG*

Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A

WWang7@mdanderson.org

APPENDIX

A. Computation of IPCW Kendall's τ

Letting (X_1, Y_1) and (X_2, Y_2) be two independent realizations of (X, Y), the first and second gap times, and letting $\psi_{12} = I\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - I\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$ indicate the concordant/discordant status of the pair, the Kendall's τ (Gibbons and Kendall, 1990) can be estimated from uncensored bivariate data $\{(X_i, Y_i), i = 1, \ldots, n\}$ by

$$\binom{n}{2}^{-1} \sum_{i<j} \psi_{ij}$$

. In the presence of censoring events (V_X, V_Y), respectively related to the two gap times, the estimation of τ can only be based on orderable pairs. Let one observation be denoted as $(\tilde{X}, \tilde{Y}, \delta_X, \delta_Y)$, where $\tilde{X} = \min(X, V_X)$, $\tilde{Y} = \min(Y, V_Y)$, $\delta_X = I(X < V_X)$ and $\delta_Y = I(Y < V_Y)$. Oakes (1982) showed that the pair (i, j) is orderable if $\{\tilde{X}_{ij} < \tilde{V}_{X_{ij}}, \tilde{Y}_{ij} < \tilde{V}_{Y_{ij}}\}$, where $\tilde{X}_{ij} = \min(X_i, X_j)$, $\tilde{Y}_{ij} = \min(Y_i, Y_j)$, $\tilde{V}_{X_{ij}} = \min(V_{X\,i}, V_{X\,j})$, and $\tilde{V}_{Y_{ij}} = \min(V_{Y\,i}, V_{Y\,j})$. Letting L_{ij} be the indicator of this event, and $\hat{p}_{ij}$ be an estimator of the probability of being orderable $p_{ij} = \Pr(V_X > \tilde{X}_{ij}; V_Y > \tilde{Y}_{ij} | \tilde{X}_{ij}, \tilde{Y}_{ij})$, Lakhal-Chaieb *and others* (2010) proposed the weighted estimate as

$$\hat{\tau}_m = \left(\sum_{i<j} \frac{L_{ij}}{\hat{p}_{ij}} \right)^{-1} \sum_{i<j} \frac{L_{ij}\psi_{ij}}{\hat{p}_{ij}}$$

To identify orderable pairs and estimate the corresponding p_{ij}, Lakhal-Chaieb *and others* (2010) showed that L_{ij} can be reduced to that X_i and X_j being uncensored, $\tilde{Y}_{ij}$ being observed, and that $\{V_{X\,i} > X_i + \tilde{Y}_{ij}; V_{X\,j} > X_j + \tilde{Y}_{ij}\}$. The conditional probability of a pair being orderable is then

$$\begin{aligned} p_{ij} &= \Pr\{V_{X\,i} > X_i + \tilde{Y}_{ij}; V_{X\,j} > X_j + \tilde{Y}_{ij} | X_i, X_j, \tilde{Y}_{ij}\} \\ &= G(X_i + \tilde{Y}_{ij}) \times G(X_j + \tilde{Y}_{ij}) \end{aligned}$$

The probability is estimated by

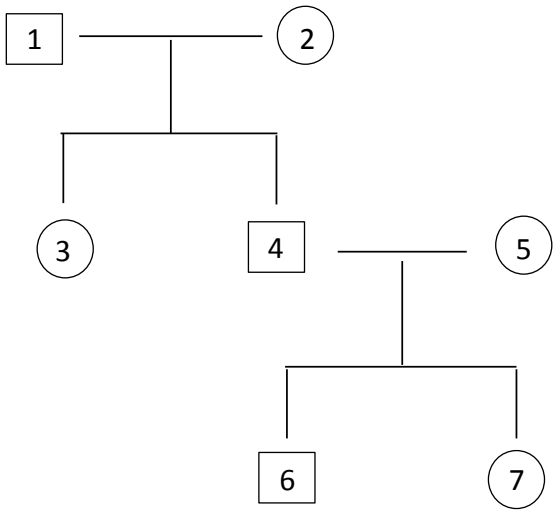
$$\hat{p}_{ij} = \hat{G}(X_i + \tilde{Y}_{ij}) \times \hat{G}(X_j + \tilde{Y}_{ij})$$

where $\hat{G}(.)$ is the Kaplan-Meier estimator of $G(.)$ based on $\{(\tilde{X}_k + \tilde{Y}_k, 1 - \delta_{Y_k}), k = 1, \cdots, n\}$. The standard error of the Kendall's τ is estimated by the jackknife technique.

B. An example of using the peeling algorithm to calculate the familywise likelihood

Supplementary Figure 1 shows an example of a hypothetical family with 3 generations. 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 peels through the family by considering individuals 1, 2, 3 as anterior and individuals 5, 6, 7 as posterior of individual 4. We can then compute the family-wise likelihood $\Pr(\mathbf{h}|\mathbf{g}_{obs})$ as follows:

$$
\begin{aligned}
&\Pr(\mathbf{h}|\mathbf{g}_{obs}) \\
&\quad = \Pr(h_4|\mathbf{g}_{obs}) \times \Pr(h_1, h_2, h_3|\mathbf{g}_{obs}) \times \Pr(h_5, h_6, h_7|\mathbf{g}_{obs}) \\
&\quad = \Pr(h_4|g_4) \times \Pr(h_1|g_1) \cdot \Pr(h_2, h_3|g_1, g_4) \times \Pr(h_5, h_6, h_7|g_1, g_4) \\
&\quad = \Pr(h_4|g_4) \times \Pr(h_1|g_1) \cdot \left[\sum_{g_2} \Pr(h_2|g_2) \Pr(h_3|g_1, g_2, g_4) \Pr(g_2|g_1, g_4) \right] \\
&\qquad \times \left[\sum_{g_5} \Pr(h_5|g_5) \Pr(h_6, h_7|g_1, g_4, g_5) \Pr(g_5|g_1, g_4) \right] \\
&\quad = \Pr(h_4|g_4) \times \Pr(h_1|g_1) \cdot \left[\sum_{g_2} \Pr(h_2|g_2) \Pr(g_2|g_4) \left\{ \sum_{g_3} \Pr(h_3|g_3) \Pr(g_3|g_1, g_2, g_4) \right\} \right] \\
&\qquad \times \left[\sum_{g_5} \Pr(h_5|g_5) \Pr(g_5) \left\{ \sum_{g_6} \Pr(h_6|g_6) \Pr(h_7|g_4, g_5) \Pr(g_6|g_4, g_5) \right\} \right] \\
&\quad = \Pr(h_4|g_4) \times \Pr(h_1|g_1) \cdot \left[\sum_{g_2} \Pr(h_2|g_2) \Pr(g_2|g_4) \left\{ \sum_{g_3} \Pr(h_3|g_3) \Pr(g_3|g_1, g_2, g_4) \right\} \right] \\
&\qquad \times \left[\sum_{g_5} \Pr(h_5|g_5) \Pr(g_5) \left\{ \sum_{g_6} \Pr(h_6|g_6) \Pr(g_6|g_4, g_5) \left(\sum_{g_7} \Pr(h_7|g_7) \Pr(g_7|g_4, g_5) \right) \right\} \right].
\end{aligned}
$$

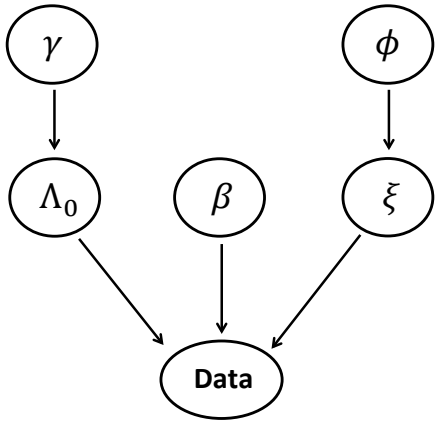


Supplementary Figure 1. A hypothetical pedigree to illustrate the likelihood calculation using the Elston-Stewart algorithm. The family consists of three generations. The circle indicates the female member while the square indicates the male. The horizontal lines indicate marriage and vertical lines indicate the next generation. In this example, the genotype is assumed unknown for every members except the 1st and 4th individuals.

All probabilities in the last equation are straightforward to compute when the mode of inheritance is known.

C. Bayesian estimation procedure

In this study, we used the MCMC algorithm to generate posterior distributions for model parameter estimation. The algorithm integrates the Metropolis-Hastings algorithm, which draws posterior samples by comparing posterior densities from two adjacent iterations, with the Gibbs sampling scheme, which allows for sampling multiple model parameters within an iteration by utilizing the full conditional likelihood. More details about the MCMC algorithm can be found in Hoff (2009); Gelman *and others* (2014). Here, we show the Bayesian inference in the frailty model. The inference of the final model we used for the LFS study can be made by simply removing the part for the frailty estimation.



Supplementary Figure 2. Graphical representation of the Bayesian frailty model. Λ_0 is the cumulative baseline rate function; ϕ is the hyper-parameter of frailty ξ.

Supplementary Figure 2 shows the frailty model represented by a directed graph that connects the observed data, model parameters and the hyper-parameter, and details about MCMC algorithm is summarized in the following:

- **Prior setting**

  $\beta \sim N(0, 100^2)$; γ: flat prior; $\phi \sim Gamma(.01, .01)$

- **Proposal setting**

  Given $\theta^{(t-1)}$, generate $\theta^* \sim q(\theta^{(t-1)})$

- **Iterative updating:**

  1) Compute proposal adjustment $adj = \frac{q(\theta^{(t-1)}|\theta^*)}{q(\theta^*|\theta^{(t-1)})}$;

  2) Let $\mathbf{h}$ denote the cancer phenotype (or survival) data, and $p(\mathbf{h}|\theta^*, others)$ denote the full conditional distribution of θ^*, and compute the acceptance ratio

$$r = min\left(\frac{p(\mathbf{h}|\theta^*, others)p(\theta^*)}{p(\mathbf{h}|\theta^{(t-1)}, others)p(\theta^{(t-1)})} * adj, 1\right)$$

  3) Take

$$\theta^{(t)} = \begin{cases} \theta^*, & \text{with probability r} \\ \theta^{(t-1)}, & \text{with probability } 1-r \end{cases}$$

4) Sample $u \sim Uniform(0,1)$, and set $\theta^{(t)} = \theta^*$ if $u < r$ or $\theta^{(t)} = \theta^{(t-1)}$ otherwise.

Since we have parameters (e.g., γ, ξ and ϕ) that only take positive values, we employ a log-normal proposal. Suppose $\gamma^{(t-1)} \in (0,+\infty) \sim \log N(\mu,\sigma)$, and $\log \gamma^{(t-1)} \in (-\infty,+\infty) \sim N(\mu',\sigma')$. To propose a new sample, we generate $\log \gamma^* = \log \gamma^{(t-1)} + \epsilon$ where $\epsilon \sim N(0,1)$, by which we can obtain $\gamma^* = \exp(\log \gamma^*) \in (0,+\infty)$. To adjust the asymmetric proposal density, we calculate

$$adj = \frac{lnN(\gamma^{(t-1)}|ln\gamma^*)}{lnN(\gamma^*|ln\gamma^{(t-1)})} = \frac{\frac{1}{\gamma^{(t-1)}\sigma\sqrt{2\pi}} \exp[-\frac{(ln\gamma^{(t-1)}-ln\gamma^*)^2}{2\sigma^2}]}{\frac{1}{\gamma^*\sigma\sqrt{2\pi}} \exp[-\frac{(ln\gamma^*-ln\gamma^{(t-1)})^2}{2\sigma^2}]} = \frac{\gamma^*}{\gamma^{(t-1)}}$$

which is simply the ratio of the proposed samples.

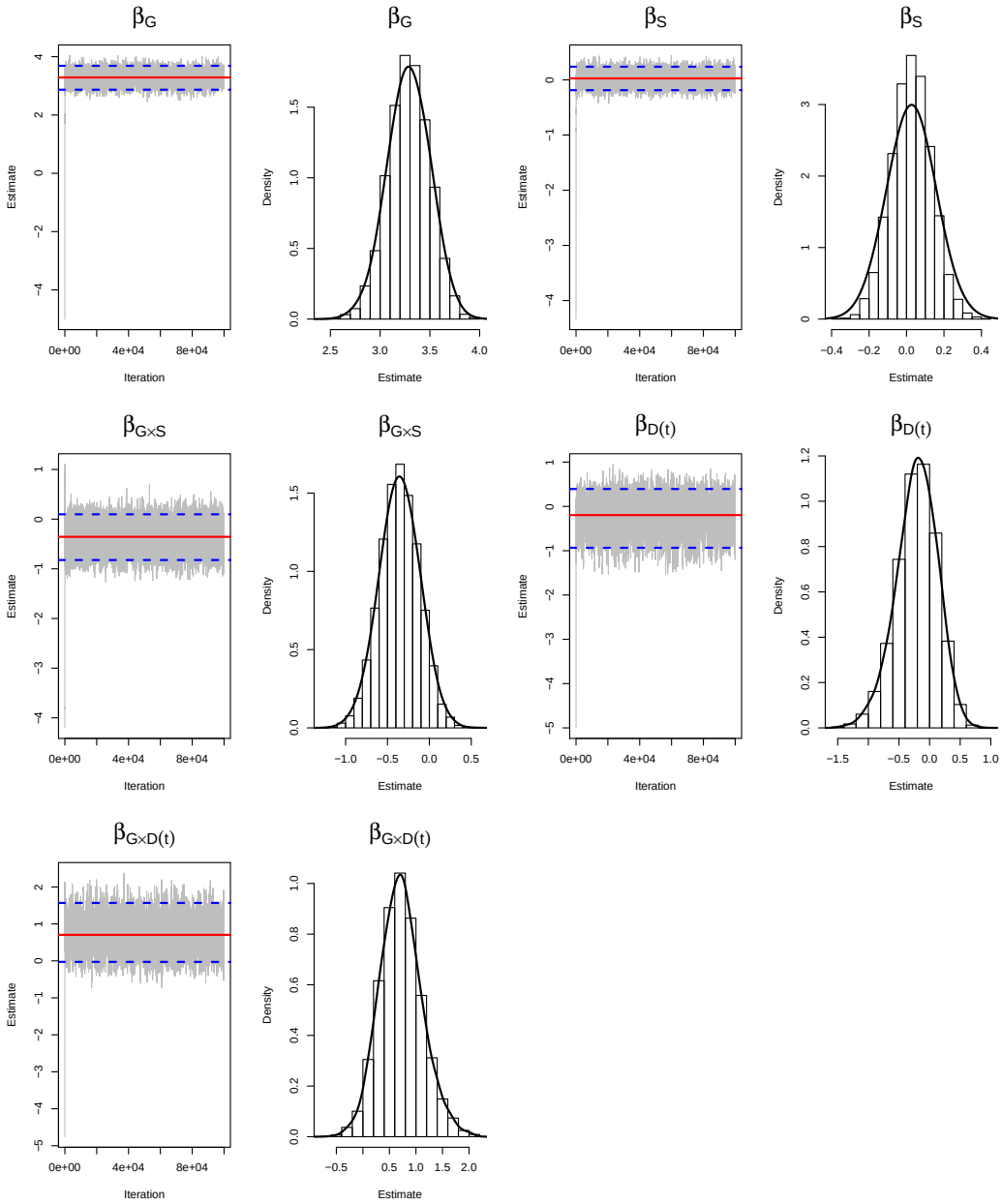
The posterior density for ϕ was constructed as previously described (Clayton, 1991). In brief, let $\phi \sim Gamma(\nu_a,\nu_b)$, or $f(\phi|\nu_a,\nu_b) = \frac{\nu_b^{\nu_a}}{\Gamma\{\nu_a\}}\phi^{\nu_a-1}\exp\{-\nu_b\phi\}$, where ν_a,ν_b are the shape and rate of the Gamma distribution, respectively. The posterior density of ϕ is then

$$\begin{aligned}
\Pr(\phi|\boldsymbol{\xi}) &\propto \Pr(\boldsymbol{\xi}|\phi)\Pr(\phi|\nu_a,\nu_b) \\
&= \prod_i^I \frac{\phi^\phi \xi_i^{(\phi-1)}\exp(-\phi\xi_i)}{\Gamma(\phi)} \frac{\nu_b^{\nu_a}\phi^{(\nu_a-1)}\exp(-\nu_b\phi)}{\Gamma(\nu_a)} \\
&= \frac{\phi^{I\phi+\nu_a-1}\exp(-\nu_b\phi)\exp\left(\left[(\phi-1)\log\prod_i^I\xi_i - \phi\sum_i^I\xi_i\right]\right)}{\Gamma(\phi)^I}.
\end{aligned}$$

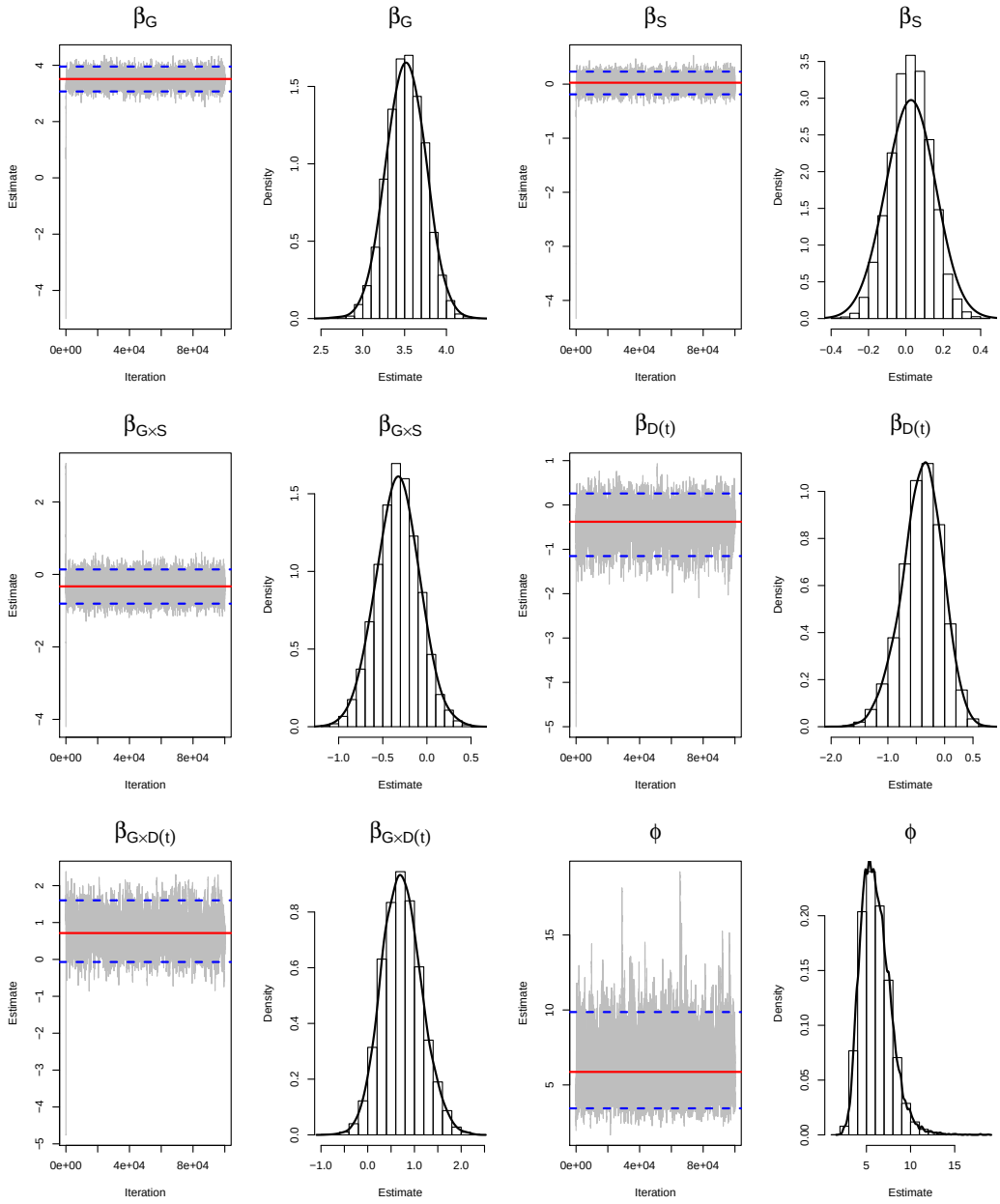
where I denotes the number of families.

Finally, we implemented this MCMC algorithm in R as follows.

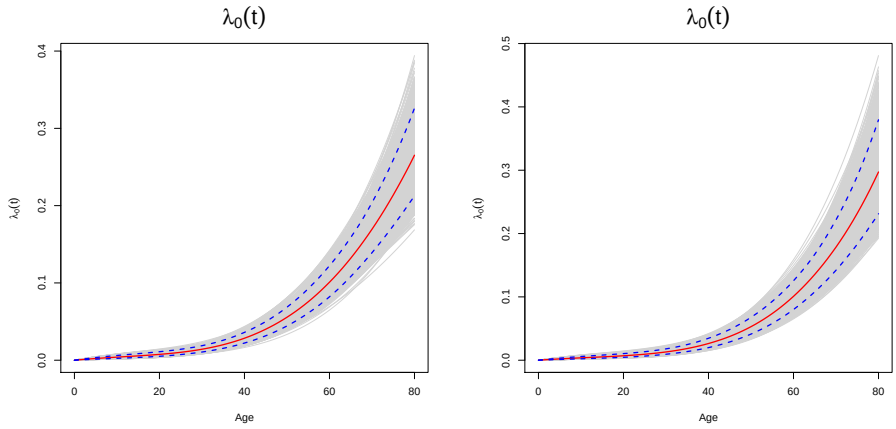
To check the convergence of the algorithm, we applied the proposed models both with and without frailty term to the real data. Supplementary Figure 3, Supplementary Figure 4, and Supplementary Figure 5 show the results. Both models converges well and the results are nearly identical.



Supplementary Figure 3. Trace plots and density distribution of posterior samples (after removing burn-in) from the proposed method. The red line indicates posterior median estimate. The density distribution is estimated based on the histogram.



Supplementary Figure 4. Trace plots and density distribution of posterior samples (after removing burn-in) from the frailty model. The red line indicates posterior median estimate. The density distribution is estimated based on the histogram.

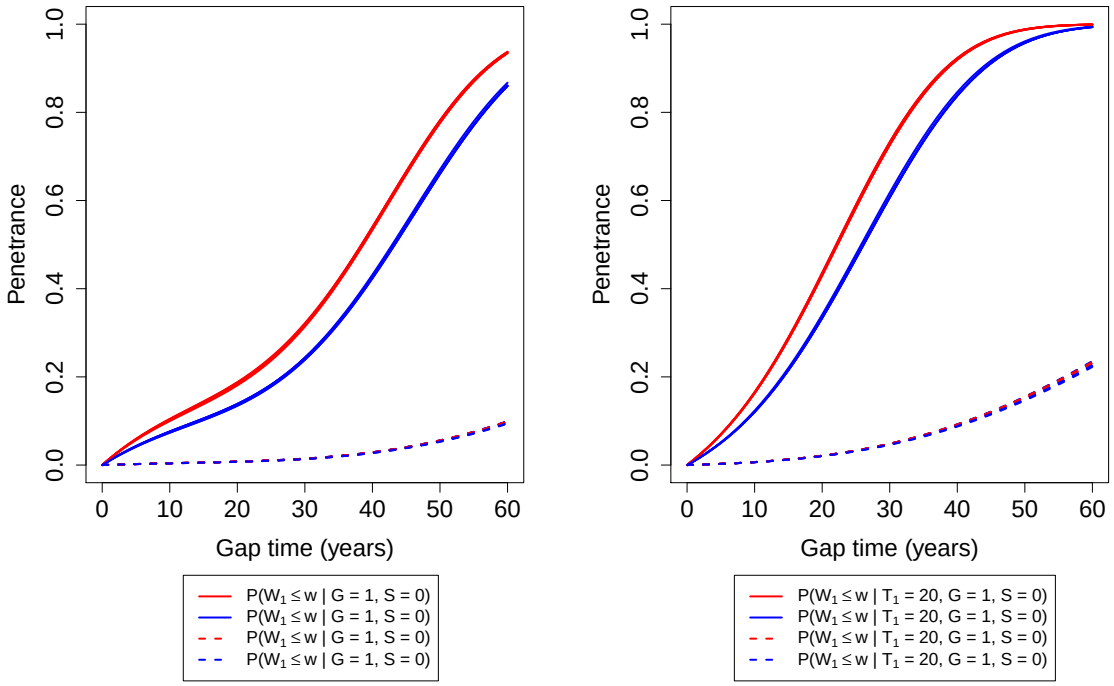


(a) No Frailty (b) Frailty

Supplementary Figure 5. Comparison of Baseline Estimates for frailty vs. no frailty models.

D. Sensitivity prior analysis

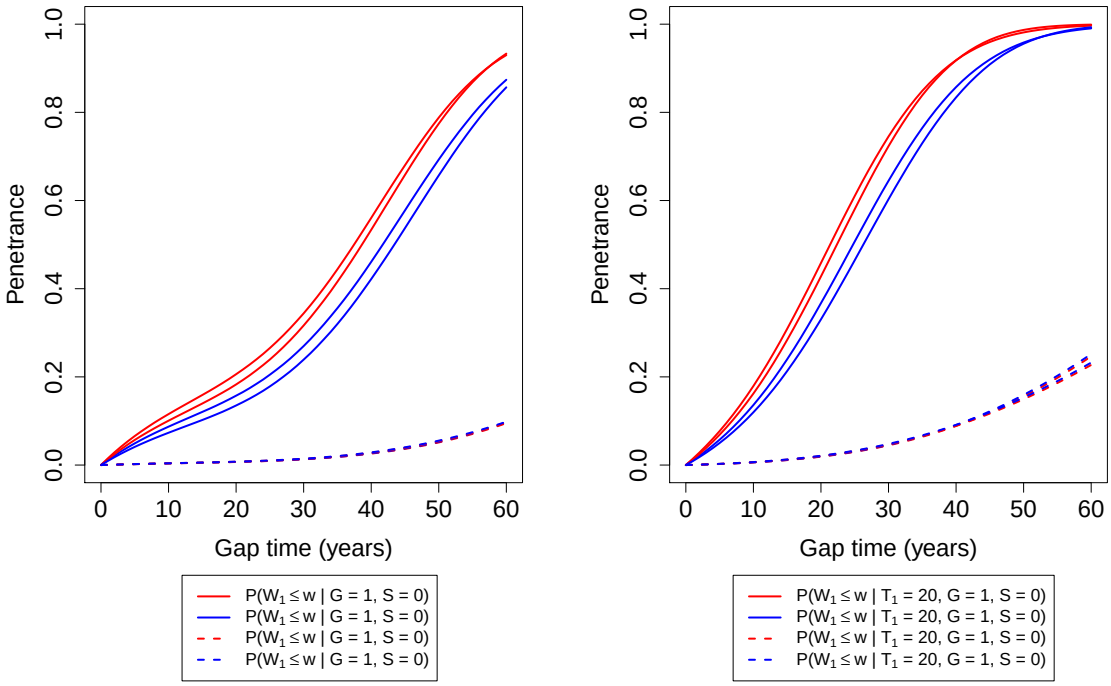
We performed sensitivity analysis by comparing penetrance estimates under different prior settings. We tested 6 combinations of priors for β and γ: three different priors for β, including $Nomral(0, 100^2)$, $Normal(0, 10^2)$ and a flat prior, and three different priors for γ including $Gamma(0.1, 0.1)$ and a flat prior. Supplementary Figure 6 shows their penetrance estimates for the first or the second primary cancers for each subgroup.



Supplementary Figure 6. Penetrance estimates from sensitivity prior analysis for the first (left) or the second primary cancer (right). Penetrances estimated from the different combinations of prior settings are shown with the same color and line type for each subgroup.

E. Penetrance Estimates from the Frailty Model

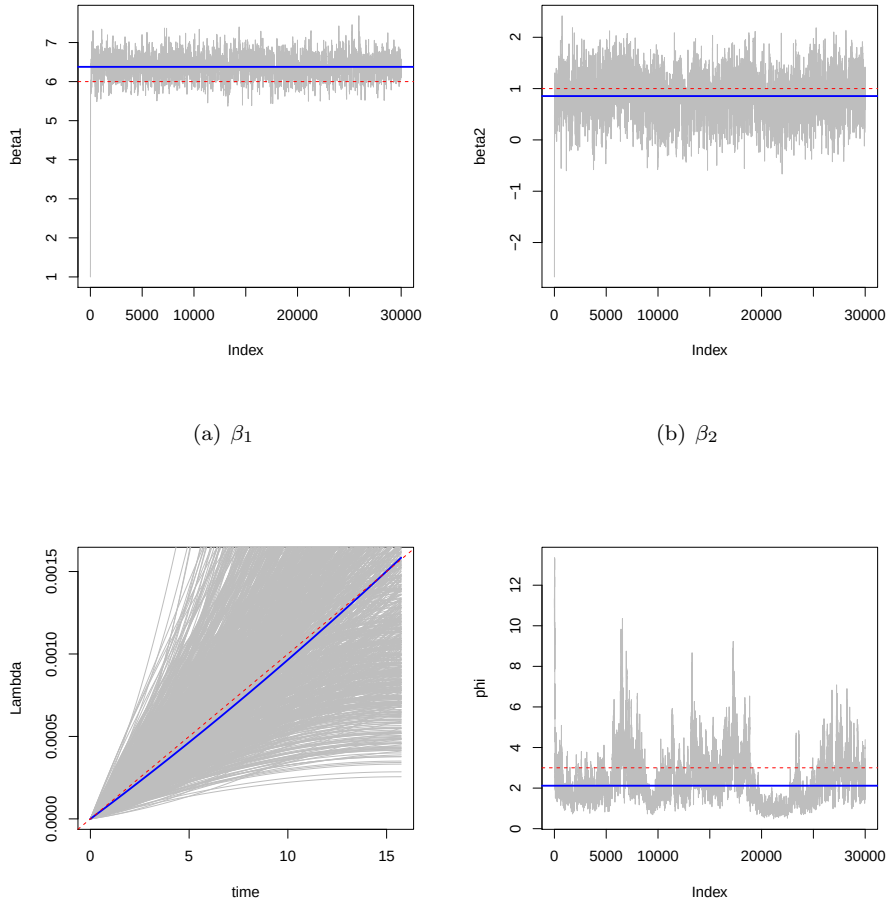
Penetrance estimates from the frailty model and the model without frailty are shown in Supplementary Figure 7. There is no obvious difference between the two sets of estimates.



Supplementary Figure 7. Comparison of penetrance estimates generated from frailty model and model without frailty.

F. Illustration of R-code

We provide estimation results for a simulated dataset with 50 families. The data generation procedure is described in Section 4. As shown in Supplementary Figure 8, our code successfully recovers the true values of all parameters. The complete set of source code, including the set that reproduces the results presented in this section, is available at `http://github.com/wwylab/MPC`.



(a) β_1 (b) β_2

(c) $\Lambda_0(t)$ (d) ϕ

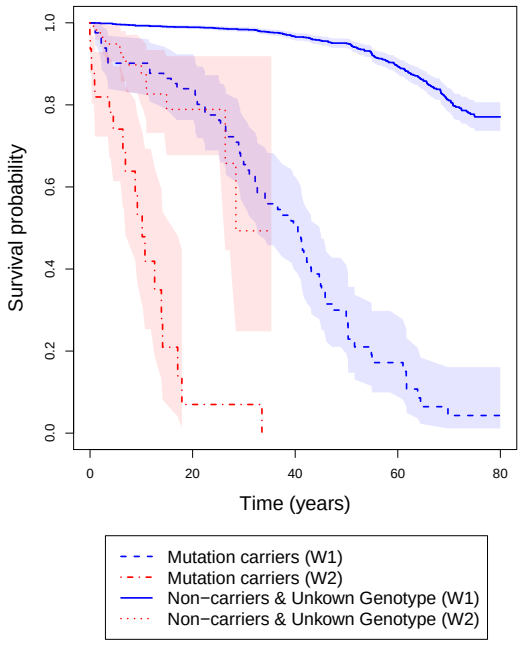
Supplementary Figure 8. Our code successfully recovers the true values of all parameters. Here (blue) solid lines represent posterior estimates and (red) dashed lines represent true values.

G. Additional Supplementary Figures and Tables

This section contains addition figures and tables referred to in the main manuscript of this article.

Supplementary Table 1. Summary of the LFS data referred in Section 2.1. "W/ carriers", family with at least one mutation carrier; "W/O carriers", family with no observed mutation carriers.

	W/ carriers	W/O carriers	total
Number of families	17	172	189
Number of individuals	2,409	1,297	3,706
Average family size	142	8	20

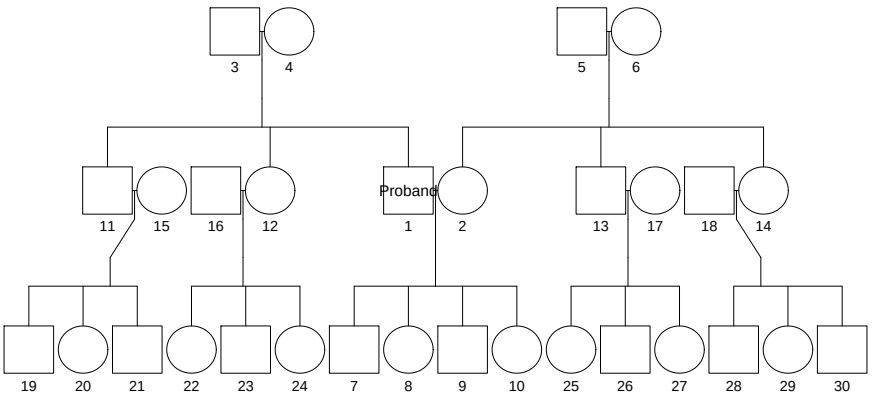


Supplementary Figure 9. Kaplan-Meier estimates of the survival distributions for the first or the second gap times of the LFS dataset without probands, referred in Section 2.2. The solid lines denote mutation carriers. The dotted lines denote individuals either with a wildtype or without any genotype information. Blue denotes the first gap time W_1 and pink denotes the second gap time W_2. The shaded areas are the 95% confidence bounds. A log-rank test gave p-values $< 10^{-7}$ comparing the first and second gap time distributions for individuals that are *TP53* mutation carriers, or otherwise, respectively.

References

Clayton, David G. (1991). A monte carlo method for bayesian inference in frailty models. *Biometrics*, 467–485.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki and



Supplementary Figure 10. Illustration of the artificial pedigree structure used for the simulation study in Section 6.

Supplementary Table 2. Summary of deviance information criterion (DIC) for model selection referred in Section 7.1. *This model is selected.

Model	Covariates	DIC
(M1)	$\{G, S, D(t)\}$	3469.75
(M2)	$\{G, S, D(t), G \times S\}$	3529.36
(M3)	$\{G, S, D(t), G \times D(t)\}$	3526.03
(M4)*	$\{G, S, D(t), G \times S, G \times D(t)\}$	3478.01
(M5)	$\{G, S, D(t), G \times S, G \times D(t), S \times D(t)\}$	3499.61

Rubin, Donald B. (2014). *Bayesian data analysis*, Volume 2. CRC press Boca Raton, FL.

Gibbons, Jean D and Kendall, MG. (1990). Rank correlation methods. *Edward Arnold*.

Hoff, Peter D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Lakhal-Chaieb, Lajmi, Cook, Richard J and Lin, Xihong. (2010). Inverse probability of censoring weighted estimates of kendall's τ for gap time analyses. *Biometrics* **66**(4), 1145–1152.

Oakes, David. (1982). A concordance test for independence in the presence of censoring. *Biometrics*, 451–455.