Supplemental Materials

Deep Learning of Semi-Competing Risk Data
via a New Neural Expectation-Maximization Algorithm

Contents

A Illness-Death Model

A.1 Notation

Let T_1 and T_2 denote times to a non-terminal and terminal event, respectively. Let $\lambda_1(t_1)$ denote the hazard of the non-terminal event at time t_1, $\lambda_2(t_2)$ denote the hazard of the terminal event at t_2 without experiencing the non-terminal event, and $\lambda_3(t_2 \mid t_1)$ denote the hazard of the terminal event at t_2 given the non-terminal event at $t_1 \leq t_2$. These hazard rates, corresponding to the transitions between states, are defined as

$$\lambda_1(t_1) = \lim_{\Delta \to 0} \Pr[T_1 \in [t_1, t_1 + \Delta) \mid T_1 \geq t_1, T_2 \geq t_1]/\Delta; \tag{A.1}$$

$$\lambda_2(t_2) = \lim_{\Delta \to 0} \Pr[T_2 \in [t_2, t_2 + \Delta) \mid T_1 \geq t_2, T_2 \geq t_2]/\Delta; \tag{A.2}$$

$$\lambda_3(t_2 \mid t_1) = \begin{cases} \lim_{\Delta \to 0} \Pr[T_2 \in [t_2, t_2 + \Delta) \mid T_1 = t_1, T_2 \geq t_2]/\Delta, & t_2 > t_1 > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{A.3}$$

Note that the definitions of $\lambda_1(t_1)$ and $\lambda_2(t_2)$ mirror that of the cause-specific hazards under a competing risks framework, where they describe the hazards of first observing either the non-terminal or terminal event. Under semi-competing risks, observing the non-terminal event is subject to observing the terminal event, but not vice versa. Hence, $\lambda_3(t_2 \mid t_1)$ describes the hazards of observing the terminal event at t_2 after having observed the non-terminal event at t_1. As we cannot observe the non-terminal event after the terminal event has been observed, the space of (T_1, T_2) is restricted to the so-called 'upper wedge' of the first quadrant where $t_1 \leq t_2$, and the non-terminal event is said to be dependently censored by the terminal event. To incorporate this dependence, we parameterize (A.1) - (A.3) by extending the Cox proportional hazards model [1] to a shared gamma-frailty conditional Markov model, given by

$$\lambda_1(t_1 \mid \gamma) = \gamma \lambda_{01}(t_1) \exp\{h_1(\boldsymbol{x})\}; \tag{A.4}$$

$$\lambda_2(t_2 \mid \gamma) = \gamma \lambda_{02}(t_2) \exp\{h_2(\boldsymbol{x})\}; \tag{A.5}$$

$$\lambda_3(t_2 \mid t_1, \gamma) = \begin{cases} \gamma \lambda_{03}(t_2) \exp\{h_3(\boldsymbol{x})\}, & t_2 > t_1 > 0; \\ 0, & \text{otherwise,} \end{cases} \tag{A.6}$$

where γ is a random effect, referred to as a subject's *frailty*, $\lambda_{0g}(t); g \in \{1, 2, 3\}$ are the baseline hazards for the three state transitions, $\boldsymbol{x}$ is a p-vector of covariates, and $h_g(\boldsymbol{x})$ are log-risk functions which relate the covariates to the hazard rates for each transition. Based on (A.4) - (A.6), survival functions of interest are

$$S(t_1, t_1 \mid \gamma) = \Pr(T_1 > t_1, T_2 > t_1 \mid \gamma) = \exp\left\{-\gamma\left[\Lambda_{01}(t_1) \exp\left\{h_1(\boldsymbol{x})\right\} + \Lambda_{02}(t_1) \exp\left\{h_2(\boldsymbol{x})\right\}\right]\right\} \tag{A.7}$$

$$S_{2|1}(t_2 \mid t_1, \gamma) = \Pr(T_2 > t_2 \mid T_1 = t_1, \gamma) = \exp\left\{-\gamma[\Lambda_{03}(t_2) - \Lambda_{03}(t_1)] \exp\left\{h_3(\boldsymbol{x})\right\}\right\}; \; t_2 > t_1 > 0, \tag{A.8}$$

where $\Lambda_{0g}(t) = \int_0^t \lambda_{0g}(u)du$ are the cumulative baseline hazards for each transition. The joint survival function evaluated at $t_2 = t_1$ takes the form in (A.7) due to the competing nature of the state transitions from no event to the first of either the non-terminal or the terminal event. The survival function of t_2 conditional on t_1 and the frailty in (A.8) is defined in terms of the difference in time between events.

A.2 Conditional Likelihood

As both events are subject to censoring, we do not fully observe (T_1, T_2). With censoring time, C, we observe

$$\mathcal{D} = \{(Y_{i1}, \delta_{i1}, Y_{i2}, \delta_{i2}, \boldsymbol{x}_i); \; i = 1, \ldots, n\},$$

where $Y_{i2} = \min(T_{i2}, C_i)$, $\delta_{i2} = I(T_{i2} \leq C_i)$, $Y_{i1} = \min(T_{i1}, Y_{i2})$, $\delta_{i1} = I(T_{i1} \leq Y_{i2})$, $\boldsymbol{x}_i$ is a p-vector of covariates, $I(\cdot)$ is the indicator function, and i indexes an individual in the study, $i = 1, \ldots, n$. There are four potential observation types for an individual during a finite period of follow-up, as given in Table A.1.

Table A.1: Potential event progressions and corresponding observed data

	Observed Event Progression		Observed Data			
Case	Non-Terminal	Terminal	Y_{i1}	δ_{i1}	Y_{i2}	δ_{i2}
1	✓	✓	T_{i1}	1	T_{i2}	1
2	✗	✓	T_{i2}	0	T_{i2}	1
3	✓	✗	T_{i1}	1	C_i	0
4	✗	✗	C_i	0	C_i	0

To construct the likelihood conditional on the subject-specific frailties, we multiply the likelihood contributions under each case in Table A.1, raised to the appropriate event indicators, and taken over the n subjects. Define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ to be the n-vector of latent frailties, and let $\boldsymbol{\psi} = \{\Lambda_{01}, \Lambda_{02}, \Lambda_{03}, \theta\}$ denote the collection of model parameters to be learned. The likelihood function is

$$
\begin{aligned}
L(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} & \left[\lambda_1(Y_{i1}) S(Y_{i1}, Y_{i1} \mid \gamma_i) \times \lambda_3(Y_{i2}) S_{2|1}(Y_{i2} \mid Y_{i1}, \gamma_i)\right]^{\delta_{i1}\delta_{i2}} \times \left[\lambda_2(Y_{i2}) S(Y_{i1}, Y_{i1} \mid \gamma_i)\right]^{(1-\delta_{i1})\delta_{i2}} \\
& \times \left[\lambda_1(Y_{i1}) S(Y_{i1}, Y_{i1} \mid \gamma_i) \times S_{2|1}(Y_{i2} \mid Y_{i1}, \gamma_i)\right]^{\delta_{i1}(1-\delta_{i2})} \times \left[S(Y_{i1}, Y_{i1} \mid \gamma_i)\right]^{(1-\delta_{i1})(1-\delta_{i2})} \\
= \prod_{i=1}^{n} & \left[\gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\left\{h_1(\boldsymbol{x}_i)\right\} + \Lambda_{02}(Y_{i1}) \exp\left\{h_2(\boldsymbol{x}_i)\right\}\right]\right\}\right. \\
& \left. \times \gamma_i \lambda_{03}(Y_{i2}) \exp\{h_3(\boldsymbol{x}_i)\} \exp\left\{-\gamma_i [\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] \exp\left\{h_3(\boldsymbol{x}_i)\right\}\right\}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \left[\gamma_i \lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{x}_i)\} \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\left\{h_1(\boldsymbol{x}_i)\right\} + \Lambda_{02}(Y_{i1}) \exp\left\{h_2(\boldsymbol{x}_i)\right\}\right]\right\}\right]^{(1-\delta_{i1})\delta_{i2}} \\
& \times \left[\gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\left\{h_1(\boldsymbol{x}_i)\right\} + \Lambda_{02}(Y_{i1}) \exp\left\{h_2(\boldsymbol{x}_i)\right\}\right]\right\}\right. \\
& \left. \times \exp\left\{-\gamma_i [\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] \exp\left\{h_3(\boldsymbol{x}_i)\right\}\right\}\right]^{\delta_{i1}(1-\delta_{i2})} \\
& \times \left[\exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\}\right]\right\}\right]^{(1-\delta_{i1})(1-\delta_{i2})} \\
= \prod_{i=1}^{n} & \left[\gamma_i^2 \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} \lambda_{03}(Y_{i2}) \exp\{h_3(\boldsymbol{x}_i)\} \times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\left\{h_1(\boldsymbol{x}_i)\right\}\right.\right.\right. \\
& \left.\left.\left. + \Lambda_{02}(Y_{i1}) \exp\left\{h_2(\boldsymbol{x}_i)\right\} + \left\{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\right\} \exp\left\{h_3(\boldsymbol{x}_i)\right\}\right]\right\}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \left[\gamma_i \lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{x}_i)\} \times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\left\{h_1(\boldsymbol{x}_i)\right\} + \Lambda_{02}(Y_{i1}) \exp\left\{h_2(\boldsymbol{x}_i)\right\}\right]\right\}\right]^{(1-\delta_{i1})\delta_{i2}} \\
& \times \left[\gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} \times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\left\{h_1(\boldsymbol{x}_i)\right\} + \Lambda_{02}(Y_{i1}) \exp\left\{h_2(\boldsymbol{x}_i)\right\}\right.\right.\right. \\
& \left.\left.\left. + \left\{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\right\} \exp\left\{h_3(\boldsymbol{x}_i)\right\}\right]\right\}\right]^{\delta_{i1}(1-\delta_{i2})} \\
& \times \left[\exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\}\right]\right\}\right]^{(1-\delta_{i1})(1-\delta_{i2})} \\
= \prod_{i=1}^{n} & \left[\gamma_i \lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\}\right]^{\delta_{i1}} \times \left[\gamma_i \lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{x}_i)\}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\gamma_i \lambda_{03}(Y_{i2}) \exp\{h_3(\boldsymbol{x}_i)\}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\}\right]\right\}^{(1-\delta_{i1})} \\
& \times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\} + \{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\} \exp\{h_3(\boldsymbol{x}_i)\}\right]\right\}^{\delta_{i1}}.
\end{aligned}
$$

Consolidating the remaining terms yields the final expression for the likelihood function, given by

$$
\begin{aligned}
L(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} & \gamma_i^{\delta_{i1}+\delta_{i2}} \times \left[\lambda_{01}\left(Y_{i1}\right) \exp\{h_1(\boldsymbol{x}_i)\}\right]^{\delta_{i1}} \\
& \times \left[\lambda_{02}\left(Y_{i2}\right) \exp\{h_2(\boldsymbol{x}_i)\}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\lambda_{03}\left(Y_{i2}\right) \exp\{h_3(\boldsymbol{x}_i)\}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \exp\left\{-\gamma_i \left[\Lambda_{01}\left(Y_{i1}\right) \exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}\left(Y_{i1}\right) \exp\{h_2(\boldsymbol{x}_i)\}\right.\right. \\
& \qquad \left.\left. + \, \delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] \exp\{h_3(\boldsymbol{x}_i)\}\right]\right\}.
\end{aligned}
\tag{A.9}
$$

As no maximizer exists for this function over the space of absolutely continuous cumulative baseline hazards, we constrain the parameter space of cumulative baseline hazards, Λ_{01}, Λ_{02}, and Λ_{03}, to consist of piecewise constant CADLAG (right continuous with left-hand limits) functions, with jumps occurring at observed event times. The maximizers over this discrete space are termed non-parametric maximum likelihood estimates (NPMLEs) of Λ_{01}, Λ_{02}, and Λ_{03}. In this framework, we modify the likelihood function (A.9) by replacing $\lambda_{0g}(t)$ with $\Delta\Lambda_{0g}(t) = \Lambda_{0g}(t) - \Lambda_{0g}(t-)$, representing the jump size at t. With this substitution, and multiplication by the density function for our frailty term, we obtain

$$
\begin{aligned}
\tilde{L}\left(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}\right) = \prod_{i=1}^{n} & \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \times \gamma_i^{\frac{1}{\theta}-1} \times e^{-\frac{\gamma_i}{\theta}} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \times \left[\Delta\Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)}\right]^{\delta_{i1}} \\
& \times \left[\Delta\Lambda_{02}\left(Y_{i2}\right) e^{h_2(\boldsymbol{x}_i)}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\Delta\Lambda_{03}\left(Y_{i2}\right) e^{h_3(\boldsymbol{x}_i)}\right]^{\delta_{i1}\delta_{i2}} \\
& \times \exp\left\{-\gamma_i \left[\Lambda_{01}\left(Y_{i1}\right) e^{h_1(\boldsymbol{x}_i)} + \Lambda_{02}\left(Y_{i1}\right) e^{h_2(\boldsymbol{x}_i)}\right.\right. \\
& \qquad \left.\left. + \, \delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] e^{h_3(\boldsymbol{x}_i)}\right]\right\},
\end{aligned}
\tag{A.10}
$$

the 'complete' likelihood (A.10) defined on a step-wise constant space for $\Lambda_{0g}, g = 1, 2, 3,$. This will serve as the basis of our proposed neural expectation-maximization (EM) algorithm, detailed below.

B NEM-SCR Algorithm Details

In the following, we provide additional detail on our neural expectation-maximization for semi-competing risks (NEM-SCR) algorithm. Viewing the subject-specific frailties as missing data, the algorithm iterates between three steps, namely the expectation (E) step, the maximization (M) step, and the neural (N) step. In the E-step, we compute the conditional expectation of the log-likelihood (A.10) given the observed data, $\mathcal{D}$, and the current estimates of $\boldsymbol{\psi}$, denoted by $\boldsymbol{\psi}_c$, wherein the conditional expectation is with respect to the distribution of $\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c$. In the M-step, we calculate the NPMLEs of Λ_{0g} while assuming the known values of h_g. Subsequently, we substitute these estimates into the conditional expectation of (A.10). This process yields the conditional expectation of the log 'profile' likelihood, serving as the objective function for utilizing deep neural networks to derive estimates for the log risk functions and frailty variance.

B.1 Conditional Frailty Distributions

We detail the derivations of the conditional densities of $\boldsymbol{\gamma}$ and $\log(\boldsymbol{\gamma})$ given the data, two quantities needed in the proposed NEM-SCR algorithm. Denote by $\tilde{L}(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma})$, which was derived in Section A.2, the likelihood of the 'complete' data, and by $f(\gamma_i)$ the marginal density of γ_i. We assume that, independent of $\boldsymbol{x}_i$, each γ_i independently follows a Gamma distribution with a density function

$$f(\gamma_i) = \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \gamma_i^{\frac{1}{\theta}-1} \exp\left\{-\frac{\gamma_i}{\theta}\right\} \tag{B.1}$$

so that $\mathbb{E}[\gamma_i] = 1$ and $\mathrm{Var}(\gamma_i) = \theta$, and the marginal density of $\boldsymbol{\gamma}$ is the product over the n independent γ_i densities. Thus, for a fixed value of θ, the 'posterior' distribution of $\boldsymbol{\gamma}$ is

$$\begin{aligned} f(\boldsymbol{\gamma} \mid \mathcal{D}, \boldsymbol{\psi}) \propto f(\boldsymbol{\gamma}) \times \tilde{L}(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) &= \prod_{i=1}^{n} \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \gamma_i^{\frac{1}{\theta}-1} \exp\left\{-\frac{\gamma_i}{\theta}\right\} \times \gamma_i^{\delta_{i1}+\delta_{i2}} \\ &\times \left[\Delta\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\}\right]^{\delta_{i1}} \times \left[\Delta\Lambda_{02}(Y_{i2}) \exp\{h_2(\boldsymbol{x}_i)\}\right]^{(1-\delta_{i1})\delta_{i2}} \times \left[\Delta\Lambda_{03}(Y_{i2}) \exp\{h_3(\boldsymbol{x}_i)\}\right]^{\delta_{i1}\delta_{i2}} \\ &\times \exp\left\{-\gamma_i \left[\Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\} + \delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] \exp\{h_3(\boldsymbol{x}_i)\}\right]\right\}. \end{aligned}$$

Considering only those terms which involve γ_i, we can reduce the above expression to

$$\begin{aligned} f(\boldsymbol{\gamma} \mid \mathcal{D}, \boldsymbol{\psi}) \propto \prod_{i=1}^{n} \gamma_i^{\frac{1}{\theta}+\delta_{i1}+\delta_{i2}-1} \times \exp\Bigg\{-\gamma_i \Bigg[&\frac{1}{\theta} + \Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} \\ &+ \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\} + \delta_{i1}[\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})] \exp\{h_3(\boldsymbol{x}_i)\}\Bigg]\Bigg\}, \end{aligned} \tag{B.2}$$

which is recognized to be the 'kernel' of a Gamma distribution. Thus, conditional on the data, the γ_i's follow a Gamma$(\tilde{a},\ \tilde{b})$ distribution with

$$\tilde{a} = \frac{1}{\theta} + \delta_{i1} + \delta_{i2} \tag{B.3}$$

$$\begin{aligned} \tilde{b} = \frac{1}{\theta} &+ \Lambda_{01}(Y_{i1}) \exp\{h_1(\boldsymbol{x}_i)\} + \ \Lambda_{02}(Y_{i1}) \exp\{h_2(\boldsymbol{x}_i)\} \\ &+ \ \delta_{i1} \{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\} \exp\{h_3(\boldsymbol{x}_i)\}. \end{aligned} \tag{B.4}$$

Hence, the posterior means of the γ_i are given by $\mathbb{E}[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}] = \tilde{a}/\tilde{b}$. The posterior means of $\log(\gamma_i)$ can be derived as follows. Without loss of generality, let the rate parameter, $\tilde{b}$, equal 1, as its effect on the logarithm of γ_i is a negative linear shift by a factor of $\log(\tilde{b})$. The density of $\gamma_i \sim \text{Gamma}(\tilde{a}, 1)$ is given by

$$f(\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}) = \frac{1}{\Gamma(\tilde{a})} \gamma_i^{\tilde{a}-1} \exp\{-\gamma_i\} d\gamma_i = \frac{1}{\Gamma(\tilde{a})} \gamma_i^{\tilde{a}} \exp\{-\gamma_i\} \frac{d\gamma_i}{\gamma_i}.$$

Taking $\gamma_i = \exp\{\log(\gamma_i)\}$ and $d\gamma_i/\gamma_i = d\log(\gamma_i)$, we have

$$f(\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}) = \frac{1}{\Gamma(\tilde{a})} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i). \tag{B.5}$$

As (B.5) is a probability density function, it must integrate to unity. Since $\log(\gamma_i)$ has support in $\mathbb{R}$, we have

$$\Gamma(\tilde{a}) = \int_{\mathbb{R}} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i).$$

Differentiating under the integral with respect to $\tilde{a}$, we have that:

$$\begin{aligned}\frac{\partial}{\partial \tilde{a}} \left[\exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i)\right] &= \log(\gamma_i)\exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i) \\ &= \Gamma(\tilde{a})\log(\gamma_i) f(\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}).\end{aligned}$$

Dividing by $\Gamma(\tilde{a})$ and integrating over $\mathbb{R}$ with respect to $\log(\gamma_i)$ yields the posterior expectation, given by

$$\begin{aligned}\mathbb{E}[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}] &= -\log(\tilde{b}) + \int_{\mathbb{R}} \log(\gamma_i) f(\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}) \\ &= -\log(\tilde{b}) + \frac{1}{\Gamma(\tilde{a})} \int_{\mathbb{R}} \frac{\partial}{\partial \tilde{a}} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i) \\ &= -\log(\tilde{b}) + \frac{1}{\Gamma(\tilde{a})} \frac{\partial}{\partial \tilde{a}} \int_{\mathbb{R}} \exp\{\tilde{a}\log(\gamma_i) - \exp\{\log(\gamma_i)\}\} d\log(\gamma_i) \\ &= -\log(\tilde{b}) + \frac{1}{\Gamma(\tilde{a})} \frac{\partial}{\partial \tilde{a}} \Gamma(\tilde{a}) = -\log(\tilde{b}) + \frac{\partial}{\partial \tilde{a}} \log\left[\Gamma(\tilde{a})\right] \\ &= \digamma(\tilde{a}) - \log(\tilde{b}),\end{aligned}$$

where $\digamma(\tilde{a}) = \partial \log[\Gamma(\tilde{a})]/\partial \tilde{a}$ and $\Gamma(\cdot)$ is the gamma function.

B.2 E-Step

The E-step calculates the expected log-conditional likelihood of the 'complete' data given the observed data, $\mathcal{D}$, and the current estimates of our parameters, denoted by $\boldsymbol{\psi}_c$. The M-step later will maximize this expectation. However, as the maximizer of this objective function over the space of absolutely continuous cumulative baseline hazards does not exist [5], we restrict the parameter space of the cumulative baseline hazards to the one containing piecewise constant functions, with jumps occurring at observed event times. Maximizers over this discrete space are termed nonparametric maximum likelihood estimates of $\Lambda_{01}, \Lambda_{02}$, and Λ_{03}. Under this parameter space, $\lambda_{0g}(t)$ in Q is replaced by $\Delta\Lambda_{0g}(t)$, the jump size at t for the baseline hazards of each state transition [5], and $\Lambda_{0g}(t) = \sum_{s=0}^{t} \Delta\Lambda_{0g}(s)$. Note that $\Delta\Lambda_{0g}(s) = 0$ if s is not one of the observed event times corresponding to state transition g. Thus, we work with $\tilde{L}(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma})$, defined over the new parameter space, and derive

$$\begin{aligned}Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_c) &= \mathbb{E}\left[\log \tilde{L}(\boldsymbol{\psi}; \mathcal{D}, \boldsymbol{\gamma}) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] \\ &= \mathbb{E}\Bigg[\log\Bigg(\prod_{i=1}^{n} \gamma_i^{\delta_{i1}+\delta_{i2}} \frac{\theta^{-\frac{1}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)} \gamma_i^{\frac{1}{\theta}-1} \exp\left\{-\frac{\gamma_i}{\theta}\right\} \times [\Delta\Lambda_{01}(Y_{i1})\exp\{h_1(\boldsymbol{x}_i)\}]^{\delta_{i1}} \\ &\quad \times [\Delta\Lambda_{02}(Y_{i2})\exp\{h_2(\boldsymbol{x}_i)\}]^{(1-\delta_{i1})\delta_{i2}} \times [\Delta\Lambda_{03}(Y_{i2})\exp\{h_3(\boldsymbol{x}_i)\}]^{\delta_{i1}\delta_{i2}} \\ &\quad \times \exp\{-\gamma_i[\Lambda_{01}(Y_{i1})\exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1})\exp\{h_2(\boldsymbol{x}_i)\} \\ &\qquad + \delta_{i1}\{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\}\exp\{h_3(\boldsymbol{x}_i)\}]\}\Bigg) \mid \mathcal{D}, \boldsymbol{\psi}_c\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{i=1}^{n} \delta_{i1}\log(\gamma_i) + \delta_{i2}\log(\gamma_i) + \delta_{i1}\log[\Delta\Lambda_{01}(Y_{i1})] + \delta_{i1}h_1(\boldsymbol{x}_i) + (1-\delta_{i1})\delta_{i2}\log[\Delta\Lambda_{02}(Y_{i2})] \\ &\quad + (1-\delta_{i1})\delta_{i2}h_2(\boldsymbol{x}_i) + \delta_{i1}\delta_{i2}\log[\Delta\Lambda_{03}(Y_{i2})] + \delta_{i1}\delta_{i2}h_3(\boldsymbol{x}_i) - \gamma_i[\Lambda_{01}(Y_{i1})\exp\{h_1(\boldsymbol{x}_i)\} \\ &\quad + \Lambda_{02}(Y_{i1})\exp\{h_2(\boldsymbol{x}_i)\} + \delta_{i1}\{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\}\exp\{h_3(\boldsymbol{x}_i)\}] - \frac{1}{\theta}\log(\theta)\end{aligned}$$

$$+ \left(\frac{1}{\theta}-1\right)\log(\gamma_i) - \frac{1}{\theta}\gamma_i - \log\left[\Gamma\left(\frac{1}{\theta}\right)\right] \mid \mathcal{D}, \boldsymbol{\psi}_c\Big]$$

$$= \sum_{i=1}^{n} \delta_{i1}\mathbb{E}\left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] + \delta_{i2}\mathbb{E}\left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] + \delta_{i1}\log[\Delta\Lambda_{01}(Y_{i1})] + \delta_{i1}h_1(\boldsymbol{x_i})$$

$$+ (1-\delta_{i1})\delta_{i2}\log[\Delta\Lambda_{02}(Y_{i2})] + (1-\delta_{i1})\delta_{i2}h_2(\boldsymbol{x_i}) + \delta_{i1}\delta_{i2}\log[\Delta\Lambda_{03}(Y_{i2})] + \delta_{i1}\delta_{i2}h_3(\boldsymbol{x_i})$$

$$- \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\left[\Lambda_{01}(Y_{i1})\exp\{h_1(\boldsymbol{x}_i)\} + \Lambda_{02}(Y_{i1})\exp\{h_2(\boldsymbol{x}_i)\}\right.$$

$$\left. + \delta_{i1}\left\{\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1})\right\}\exp\{h_3(\boldsymbol{x}_i)\}\right]$$

$$- \frac{1}{\theta}\log(\theta) + \left(\frac{1}{\theta}-1\right)\mathbb{E}\left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] - \frac{1}{\theta}\mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right] - \log\left[\Gamma\left(\frac{1}{\theta}\right)\right]$$

$$= Q_1 + Q_2 + Q_3 + Q_4,$$

where Q_1, Q_2, Q_3, and Q_4 are additive pieces of 'Q', each involving non-overlapping unknown parameters:

$$Q_1 = \sum_{i=1}^{n} \delta_{i1}\mathbb{E}\left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] + \delta_{i1}\left\{\log\left[\Delta\Lambda_{01}\left(Y_{i1}\right)\right] + h_1(\boldsymbol{x_i})\right\}$$
$$- \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\Lambda_{01}\left(Y_{i1}\right)\exp\{h_1(\boldsymbol{x}_i)\}$$

$$Q_2 = \sum_{i=1}^{n} \delta_{i2}\mathbb{E}\left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] + (1-\delta_{i1})\,\delta_{i2}\left\{\log\left[\Delta\Lambda_{02}\left(Y_{i2}\right)\right] + h_2(\boldsymbol{x_i})\right\}$$
$$- \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\Lambda_{02}\left(Y_{i1}\right)\exp\{h_2(\boldsymbol{x}_i)\}$$

$$Q_3 = \sum_{i=1}^{n} \delta_{i1}\delta_{i2}\left\{\log\left[\Delta\Lambda_{03}\left(Y_{i2}\right)\right] + h_3(\boldsymbol{x_i})\right\}$$
$$- \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\delta_{i1}\left\{\Lambda_{03}\left(Y_{i2}\right) - \Lambda_{03}\left(Y_{i1}\right)\right\}\exp\{h_3(\boldsymbol{x}_i)\}$$

$$Q_4 = \sum_{i=1}^{n} -\frac{1}{\theta}\log(\theta) + \left(\frac{1}{\theta}-1\right)\mathbb{E}\left[\log(\gamma_i) \mid \mathcal{D}, \boldsymbol{\psi}_c\right] - \frac{1}{\theta}\mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right] - \log\left[\Gamma\left(\frac{1}{\theta}\right)\right].$$

B.3 M-Step

As shown in the previous section, our objective function, $Q(\boldsymbol{\psi} \mid \mathcal{D}, \boldsymbol{\psi}_c)$, can be written as the sum of pieces Q_1, Q_2, Q_3, and Q_4. Each of the first three involves only the baseline hazard and h functions for a state transition, and the last one involves only the frailty variance. Thus, the M-step updates for $\Lambda_{0g}, h_g, g = 1, 2, 3$, can be defined utilizing Q_1, Q_2, and Q_3, separately, and the frailty variance, θ, with Q_4. Finally, because both Λ_{0g}, h_g are nonparametric, we adopt a profiling approach to facilitate maximization. Specifically, for each $g = 1, 2, 3$, we maximize Q_g with respect to the jump sizes of Λ_{0g}, while fixing h_g, and obtain the nonparametric estimates as follows. First, for any fixed $t > 0$, differentiating Q_1 with respect to $\Delta\Lambda_{01}(t)$, the jump size at t, the score function for $\Delta\Lambda_{01}(t)$ is

$$\frac{\partial Q_1}{\partial\Delta\Lambda_{01}(t)} = \sum_{i=1}^{n} \frac{\delta_{i1}I(Y_{i1}=t)}{\Delta\Lambda_{01}(t)} - \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\left(Y_{i1} \geq t\right)\exp\{h_1(\boldsymbol{x}_i)\}.$$

Setting this equal to zero, the maximizer, $\widehat{\Delta\Lambda_{01}}(t)$ is obtained as

$$\widehat{\Delta\Lambda_{01}}(t) = \frac{\sum_{i=1}^{n} \delta_{i1} I\left(Y_{i1}=t\right)}{\sum_{i=1}^{n} \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\left(Y_{i1} \geq t\right)\exp\left\{h_{1,c}(\boldsymbol{x}_i)\right\}}. \tag{B.6}$$

Then, differentiating Q_2 with respect to $\Delta\Lambda_{02}(t)$, we have the score function

$$\frac{\partial Q_2}{\partial \Delta\Lambda_{02}(t)} = \sum_{i=1}^{n} \frac{(1-\delta_{i1})\,\delta_{i2} I\,(Y_{i2}=t)}{\Delta\Lambda_{02}(t)} - \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\,(Y_{i1}\geq t)\exp\{h_2(\boldsymbol{x}_i)\}.$$

By setting this to zero and solving for $\Delta\Lambda_{02}(t)$, we obtain the maximizer, $\widehat{\Delta\Lambda_{02}}(t)$, as

$$\widehat{\Delta\Lambda_{02}}(t) = \frac{\sum_{i=1}^{n}(1-\delta_{i1})\,\delta_{i2} I\,(Y_{i2}=t)}{\sum_{i=1}^{n}\mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\,(Y_{i1}\geq t)\exp\left\{h_{2,c}(\boldsymbol{x}_i)\right\}}. \tag{B.7}$$

Finally, differentiating Q_3 with respect to $\Delta\Lambda_{03}(t)$ yields a score function for $\Delta\Lambda_{03}(t)$:

$$\frac{\partial Q_3}{\partial \Delta\Lambda_{03}(t)} = \sum_{i=1}^{n} \frac{\delta_{i1}\delta_{i2} I\,(Y_{i2}=t)}{\Delta\Lambda_{03}\,(t)} - \mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\delta_{i1}\left[I(Y_{i2}\geq t) - I(Y_{i1}\geq t)\right]\exp\{h_3(\boldsymbol{x}_i)\}.$$

Equating this to zero and solving it, we have the maximizer, $\widehat{\Delta\Lambda_{03}}(t)$, as:

$$\widehat{\Delta\Lambda_{03}}(t) = \frac{\sum_{i=1}^{n}\delta_{i1}\delta_{i2} I\,(Y_{i2}=t)}{\sum_{i=1}^{n}\mathbb{E}\left[\gamma_i \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\delta_{i1}\left[I(Y_{i2}\geq t) - I(Y_{i1}\geq t)\right]\exp\left\{h_{3,c}(\boldsymbol{x}_i)\right\}}. \tag{B.8}$$

B.4 N-Step

Noting that $\widehat{\Lambda_{0g}}(t) = \sum_{s\leq t}\widehat{\Delta\Lambda_{0g}}(s)$, then plugging the $\widehat{\Delta\Lambda_{0g}}(t)$ into Q_g yields the expected log-profile likelihoods for h_g $(g=1,2,3)$. That is, with an added subscript P (for profile), we have that

$$\begin{aligned}
Q_{1,P} &= \sum_{i=1}^{n}\delta_{i1}\left\{h_1(\boldsymbol{x_i}) - \log[\sum_{j=1}^{n}\mathbb{E}\left[\gamma_j \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\,(Y_{j1}\geq Y_{i1})\exp\left\{h_1(\boldsymbol{x}_j)\right\}]\right\} \\
&\quad - |D_1| + \sum_{i=1}^{n}\delta_{i1}\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}_c], \\
Q_{2,P} &= \sum_{i=1}^{n}(1-\delta_{i1})\delta_{i2}\left\{h_2(\boldsymbol{x_i}) - \log[\sum_{j=1}^{n}\mathbb{E}\left[\gamma_j \mid \mathcal{D}, \boldsymbol{\psi}_c\right] I\,(Y_{j2}\geq Y_{i2})\exp\left\{h_2(\boldsymbol{x}_j)\right\}]\right\} \\
&\quad - |D_2| + \sum_{i=1}^{n}\delta_{i2}\mathbb{E}[\log(\gamma_i)|\mathcal{D}, \boldsymbol{\psi}_c], \\
Q_{3,P} &= \sum_{i=1}^{n}\delta_{i1}\delta_{i2}\left\{h_3(\boldsymbol{x_i}) - \log[\sum_{j=1}^{n}\mathbb{E}\left[\gamma_j \mid \mathcal{D}, \boldsymbol{\psi}_c\right]\delta_{j1} I\,[Y_{j2}\geq \max(Y_{i2}, Y_{j1})]\exp\left\{h_3(\boldsymbol{x}_j)\right\}]\right\} - |D_3|,
\end{aligned}$$

where $|D_1|, |D_2|, |D_3|$ are the numbers of the observed progressions, deaths prior to progression, and deaths following progression, respectively. Note that in each of $Q_{1,P}, Q_{2,P}, Q_{3,P}$, only the first term involves h_g that needs to be maximized with respect to. To avoid redundancy, in the main text, we use the notations $Q_{1,P}, Q_{2,P}, Q_{3,P}$ to specifically refer to the first term in each expression, as the remaining terms can be treated as constant in the context of maximization with respect to h_g. The detailed derivations follow those in a standard Cox model setting [3]. Algorithm 1 below summarizes this procedure.

Algorithm 1 A Neural Expectation-Maximization Algorithm for Semi-Competing Risks

Require: Observed Data: $\mathcal{D} = \{(Y_{i1}, \delta_{i1}, Y_{i2}, \delta_{i2}, \boldsymbol{x}_i)\}_{i=1}^{n}$, Initial Values: $\theta^{(0)}, \Lambda_{0g}^{(0)}, h_g(\boldsymbol{x}_i)$, $g = 1, 2, 3$, Network Parameters: $\Theta^{(0)} = \{\boldsymbol{W}_l^{(0)}, b_l^{(0)}\}$, Learning Rate: η, Batch Size: B, Max Epochs: E, Inner-Loop Epochs: E_θ, Tolerance: τ_1, Convergence Tolerance: τ_2

Set iteration $k \leftarrow 0$
repeat
  $k \leftarrow k + 1$
  E–Step (Latent Frailties):
  for $i = 1$ to n **do**
    Compute $\mathbb{E}[\gamma_i \mid \mathcal{D}, \Lambda_{0g}^{(k-1)}, \theta^{(k-1)}, h_g^{(k-1)}]$ and $\mathbb{E}[\log(\gamma_i) \mid \cdot]$
  end for
  M–Step (Baseline Hazards):
  for $g = 1$ to 3 **do**
    for each unique event time t **do**
      Update jump size $\Delta\Lambda_{0g}^{(k)}(t)$
    end for
    $\Lambda_{0g}^{(k)}(t) \leftarrow \sum_{s \le t} \Delta\Lambda_{0g}^{(k)}(s)$
  end for
  N–Step (Log-Risk Functions & θ):
  Initialize Adam states for Θ; freeze $\{\Lambda_{0g}^{(k)}, \mathbb{E}[\gamma_i \mid \cdot], \mathbb{E}[\log(\gamma_i) \mid \cdot]\}$
  for epoch $= 1$ to E **do**
    Shuffle and form mini-batches of size B
    for each batch $\mathcal{B}$ **do**
      Forward pass: compute $\hat{h}_g(\boldsymbol{x}_i)$ for $i \in \mathcal{B}$, $g = 1, 2, 3$
      for $g = 1$ to 3 **do**
        Compute profile loss $Q_{g,P}$
      end for
      $Q \leftarrow Q_1 + Q_2 + Q_3$
      Backward pass: compute $\nabla_\Theta L$
      Adam step: update $\Theta \leftarrow \mathrm{Adam}(\Theta, \nabla_\Theta Q)$
    end for
  end for
  Separate θ Update:
  for inner $= 1$ to E_θ **do**
    Compute Q_4
    Back propagate
    Step
    if $|\Delta Q_4| \le \tau_1$ **then**
      break
    end if
  end for
  $\theta^{(k)} \leftarrow$ current value of θ
until $\left\|\psi^{(k)} - \psi^{(k-1)}\right\| \le \tau_2$

Ensure: Final estimates $\{\Lambda_{0g}^{(k)},\ h_g^{(k)}(\cdot),\ \theta^{(k)}\}$

C Bivariate Brier Score

We provide additional details of the derivation of the bivariate Brier score, outlined in Section 4 in the main text. Namely, we show that the expectation of the bivariate Brier score is equal to the mean squared error of the predictor, $\pi_i(t)$, plus a constant. To proceed, we compute the expectation in additive pieces. In the first piece, we consider the region where at least the non-terminal event is observed by time t, and Y_{i1} is less than or equal to Y_{i2}, but the terminal event may or may not be observed. In the second piece, we consider the region where the terminal event is observed prior to the non-terminal event occurring. In the third piece, we consider the region where neither event has been observed by time t.

Piece 1: At least the non-terminal event is observed by time t, and Y_{i1} is less than or equal to Y_{i2}, but the terminal event may or may not be observed.

$$\begin{aligned}
&\mathbb{E}\left[\frac{\pi_i(t)^2 \times I\left(Y_{i1} \leq t,\ \delta_{i1} = 1,\ Y_{i1} \leq Y_{i2}\right)}{G_i(Y_{i1})}\right] = \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq C_i,\ T_{i1} \leq T_{i2}\right)}{G_i(T_{i1})}\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left\{\mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq C_i,\ T_{i1} \leq T_{i2}\right)}{G_i(Y_{i1})} \mid T_{i1}, T_{i2}\right]\right\} \\
&= \pi_i(t)^2 \times \mathbb{E}\left\{\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right)}{G_i(T_{i1})} \times \mathbb{E}\left[I\left(T_{i1} \leq C_i\right) \mid T_{i1}, T_{i2}\right]\right\} \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right)}{G_i(T_{i1})} \times \Pr\left(T_{i1} \leq C_i\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right)}{G_i(T_{i1})} \times G_i\left(T_{i1}\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[I\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right)\right] = \pi_i(t)^2 \times \Pr\left(T_{i1} \leq t,\ T_{i1} \leq T_{i2}\right). \qquad \text{(C.1)}
\end{aligned}$$

Piece 2: The terminal event is observed prior to the non-terminal event occurring.

$$\begin{aligned}
&\mathbb{E}\left[\frac{\pi_i(t)^2 \times I\left(Y_{i1} \leq t,\ Y_{i2} \leq t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \leq Y_{i2}\right)}{G_i(Y_{i2})}\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2},\ T_{i2} \leq C_i\right)}{G_i(Y_{i2})}\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left\{\mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2},\ T_{i2} \leq C_i\right)}{G_i(Y_{i2})} \mid T_{i1}, T_{i2}\right]\right\} \\
&= \pi_i(t)^2 \times \mathbb{E}\left\{\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right)}{G_i(Y_{i2})} \times \mathbb{E}\left[I\left(T_{i2} \leq C_i\right) \mid T_{i1}, T_{i2}\right]\right\} \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right)}{G_i(Y_{i2})} \times \Pr\left(T_{i2} \leq C_i\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[\frac{I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right)}{G_i(Y_{i2})} \times G_i\left(T_{i2}\right)\right] \\
&= \pi_i(t)^2 \times \mathbb{E}\left[I\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right)\right] = \pi_i(t)^2 \times \Pr\left(T_{i2} \leq t,\ T_{i1} > T_{i2}\right). \qquad \text{(C.2)}
\end{aligned}$$

Piece 3: Neither event has been observed by time t.

$$\mathbb{E}\left[\frac{[1-\pi_i(t)]^2 \times I\left(Y_{i1} > t,\ Y_{i2} > t\right)}{G_i(t)}\right] = \frac{[1-\pi_i(t)]^2}{G_i(t)} \times \mathbb{E}\left[I\left(T_{i1} > t,\ T_{i2} > t,\ C_i > t\right)\right]$$

$$= \frac{[1-\pi_i(t)]^2}{G_i(t)} \times \Pr\left(T_{i1} > t,\ T_{i2} > t,\ C_i > t\right)$$

$$= \frac{[1-\pi_i(t)]^2}{G_i(t)} \times \Pr\left(T_{i1} > t,\ T_{i2} > t\right) \times \Pr\left(C_i > t\right) \quad \text{since} \quad T_{i1}, T_{i2} \perp C_i$$

$$= \frac{[1-\pi_i(t)]^2}{G_i(t)} \times S_i(t) \times G_i\left(t\right) = [1-\pi_i(t)]^2 \times S_i(t). \tag{C.3}$$

Combining (C.1) - (C.3), and summing over the n individuals, we can see that

$$\begin{aligned}
\mathbb{E}\left[\text{BBS}(t)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \le t,\ \delta_{i1} = 1,\ Y_{i1} \le Y_{i2}\right)}{G_i(Y_{i1})}\right. \\
&\quad \left. + \frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \le t,\ Y_{i2} \le t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \le Y_{i2}\right)}{G_i(Y_{i2})} + \frac{[1-\pi_i(t)]^2 \cdot I\left(Y_{i1} > t,\ Y_{i2} > t\right)}{G_i(t)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \le t,\ \delta_{i1} = 1,\ Y_{i1} \le Y_{i2}\right)}{G_i(Y_{i1})}\right] \\
&\quad + \mathbb{E}\left[\frac{\pi_i(t)^2 \cdot I\left(Y_{i1} \le t,\ Y_{i2} \le t,\ \delta_{i1} = 0,\ \delta_{i2} = 1,\ Y_{i1} \le Y_{i2}\right)}{G_i(Y_{i2})}\right] + \mathbb{E}\left[\frac{[1-\pi_i(t)]^2 \cdot I\left(Y_{i1} > t,\ Y_{i2} > t\right)}{G_i(t)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_i(t)^2 \cdot \Pr\left(T_{i1} \le t,\ T_{i1} \le T_{i2}\right) + \ \pi_i(t)^2 \cdot \Pr\left(T_{i2} \le t,\ T_{i1} > T_{i2}\right) + [1-\pi_i(t)]^2 \cdot S_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n} \pi_i(t)^2 \cdot [1 - S_i(t)] + [1-\pi_i(t)]^2 \cdot S_i(t) \\
&= \text{MSE}(t) + \frac{1}{n}\sum_{i=1}^{n} S_i\left(t\right) \cdot \left[1 - S_i\left(t\right)\right].
\end{aligned}$$

In expectation, the bivariate Brier score is equal to the mean squared error of the predictor plus a piece that is constant with respect to $\pi_i(t)$. This term represents the irreducible error incurred by approximating $S_i(t)$.

D Additional Simulation Results

D.1 Bivariate Brier Score

To examine the performance of the proposed bivariate Brier score, we generated 1,000 independent datasets of size $n = 1,000$ based on the illness-death model. Across all simulated datasets and simulation settings, we assumed Weibull baseline hazards with a shape parameter of 1.5 and a scale parameter of 0.2, and a population frailty variance of $\theta = 0.5$. We considered four simulation settings, varying whether the semi-competing outcomes depended on a uniform random covariate, and varying the administrative censoring rate at 0% and 50%. We calculated the integrated bivariate Brier score for 1-year survival over a grid of 100 evenly spaced time points, and compared the results from the model fit to a calculation which utilized the true model parameters. This comparison to the 'truth' gives the degree of irreducible error in the bivariate Brier score for each setting. Table D.1 shows that the results from the model fit were on par with those calculated using the true model parameters, giving an approximate lower bound for the bivariate Brier score.

Table D.1: Mean (SD) integrated bivariate Brier score (iBBS) under various data generation settings, averaged over 1,000 generated datasets for each setting to assess the degree of irreducible error in the iBBS.

Covariates Generated	Censoring Generated	True iBBS	Calculated iBBS
No	No	0.0187 (0.0068)	0.0199 (0.0073)
Yes	No	0.0181 (0.0067)	0.0205 (0.0077)
No	Yes	0.0206 (0.0067)	0.0219 (0.0072)
Yes	Yes	0.0195 (0.0066)	0.0221 (0.0075)

D.2 Sensitivity Analysis: Gamma Frailty

In a sensitivity analysis, we studied the robustness of our method to the assumed gamma distribution of the latent frailties, γ_i, as this is the only parametric assumption we make in the proposed method. Specifically, we reproduced the results of our main analysis, where we generated γ_i from a gamma distribution with mean 1 and variance θ and compared these results with settings where we instead generated γ_i from a log-normal distribution with mean 1 and variance θ, keeping all other aspects of the data generating mechanism fixed. For simplicity, we focused on one setting for the sample size ($n = 1,000$) and one setting for the censoring rate (50%), as these are the most realistic, and varied the true risk functions (linear versus nonlinear), the frailty variance ($\theta = 0.5$ versus 2.0) and the frailty distribution (gamma versus log-normal). The results of this analysis are given in Table D.2. Overall, these results show that our proposed approach performed well when the latent frailties truly followed a gamma distribution. When the latent frailties instead follow a log-normal distribution, we saw that our approach was robust in terms of its predictive performance (measured by iBBS and C-index), and its bias in estimating the baseline hazard functions for each state transition. However, we do note that the bias in the log-risk functions, h_g, grew noticeably with this misspecification, particularly when the frailty variance, θ, was large or the true risk function was highly nonlinear.

Table D.2: Sensitivity analysis comparing an assumed gamma versus log-normal frailty: average (SD) mean integrated squared errors for each log-risk function ($h_g(\boldsymbol{X}_i); g = 1, 2, 3$), integrated bivariate Brier score (iBBS), bivariate concordance (C) index, and average bias in estimating the baseline hazards, varying the risk functions and dependence (θ). All results are based on $n = 1,000$ observations and 50% censoring.

Simulation Settings				First Transition: $h_1(\boldsymbol{x}_i)$		Second Transition: $h_2(\boldsymbol{x}_i)$		Third Transition: $h_3(\boldsymbol{x}_i)$	
n	Risk Function	Censoring	θ	Gamma	Log-Normal	Gamma	Log-Normal	Gamma	Log-Normal
1,000	Linear	50%	0.5	0.73 (0.17)	2.14 (0.25)	0.88 (0.16)	2.91 (0.48)	0.43 (0.18)	2.16 (0.25)
1,000	Linear	50%	2.0	0.51 (0.17)	4.80 (0.45)	0.35 (0.11)	6.14 (0.62)	0.61 (0.18)	6.41 (1.04)
1,000	Nonlinear	50%	0.5	4.07 (0.52)	5.46 (0.51)	4.28 (0.55)	5.76 (0.55)	4.39 (0.58)	5.27 (0.47)
1,000	Nonlinear	50%	2.0	4.15 (0.48)	11.14 (1.07)	4.03 (0.58)	13.52 (1.37)	4.47 (0.54)	10.89 (1.33)

Simulation Settings				iBBS		C-Index		$\lambda_{0g}(t)$	
n	Risk Function	Censoring	θ	Gamma	Log-Normal	Gamma	Log-Normal	Gamma	Log-Normal
1,000	Linear	50%	0.5	0.03 (0.002)	0.03 (0.006)	0.65 (0.01)	0.64 (0.01)	0.95 (1.49)	0.95 (0.88)
1,000	Linear	50%	2.0	0.04 (0.006)	0.07 (0.006)	0.63 (0.01)	0.65 (0.01)	1.12 (2.48)	0.47 (1.08)
1,000	Nonlinear	50%	0.5	0.14 (0.004)	0.05 (0.006)	0.69 (0.01)	0.60 (0.02)	0.31 (0.27)	0.77 (0.53)
1,000	Nonlinear	50%	2.0	0.13 (0.010)	0.09 (0.005)	0.67 (0.01)	0.69 (0.01)	0.27 (0.45)	0.53 (0.55)

D.3 Sensitivity Analysis: Computation Time

We benchmarked how the computational cost of our proposed method scales with sample size and covariate dimension: $n \in \{100, 1000, 10000\}$, $p \in \{1, 10, 100, 1000\}$, averaging the wall time over 100 replicates. All runtime comparisons reported here were conducted using CPU-based implementations. Although our method is implemented in PyTorch, tensors were explicitly detached and transferred to CPU memory prior to downstream computation (e.g., via `.detach().cpu().numpy()`), and GPU acceleration was not used in these analyses. We then compared the wall time of our method with existing methods. All existing methods were likewise executed using their standard CPU implementations. Across all settings of (n, p), we generated the covariates following the same multivariate normal structure used in the primary simulation studies, with identical correlation structure and marginal distributions. Further, we generated the outcomes under the linear data-generating mechanism from the main simulations. These results are summarized in Table D.3, omitting settings where a method cannot be run (e.g., $p > n$ for the (semi-)parametric methods) or the runtime is prohibitively large. Overall, the wall time for our approach increased roughly linearly in n and sub-linearly in p. In the largest setting, training took about 20 seconds on a standard 14-inch MacBook Pro with an Apple M4 Pro chip (14-core CPU, 20-core GPU, 16-core Neural Engine), indicating the method is feasible without dedicated servers and adds only modest overhead relative to standard approaches given the added flexibility. We also note that our proposed method is computationally competitive with the standard existing methods and quite a bit faster than some of the more contemporary methods. Lastly, while GPU acceleration could further reduce runtime in larger-scale applications, it was not required for these analyses.

Table D.3: Summary of wall times (in seconds) for each (n, p) combination, averaged over 100 replicates.

Proposed Approach								
n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	0.02	0.02	0.03	0.02	0.03	0.57	100
100	10	0.02	0.02	0.03	0.02	0.03	0.17	100
100	100	0.02	0.04	0.13	0.05	0.09	3.99	100
100	1,000	0.06	0.07	0.11	0.08	0.10	0.65	100
1,000	1	0.49	0.79	1.06	1.06	1.31	1.86	100
1,000	10	0.50	0.81	1.08	1.04	1.27	2.21	100
1,000	100	0.47	0.69	0.95	0.89	1.14	2.41	100
1,000	1,000	1.23	1.41	1.71	1.62	1.96	3.33	100
10,000	1	5.41	6.38	6.93	6.62	7.21	9.79	100
10,000	10	6.16	6.85	7.64	7.35	8.20	12.40	100
10,000	100	6.66	7.84	8.75	8.45	9.39	12.25	100
10,000	1,000	16.84	18.72	19.98	19.48	20.93	25.44	100
Xu (2010)								
n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	0.05	0.07	0.09	0.08	0.12	0.15	100
100	10	0.27	0.33	0.39	0.37	0.41	0.70	100
1,000	1	0.26	0.30	0.34	0.34	0.36	0.63	100
1,000	10	1.46	1.56	1.81	1.87	1.94	3.40	100
1,000	100	88.86	117.39	148.00	138.20	160.64	352.46	100
10,000	1	2.06	2.59	3.24	2.94	3.92	5.30	100
10,000	10	12.49	14.68	19.50	18.29	25.06	29.31	100
10,000	100	648.72	1132.97	1321.76	1346.13	1561.24	1743.00	100
Lee (2015)								
n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	0.10	0.12	0.17	0.16	0.21	0.28	100
100	10	0.10	0.12	0.16	0.15	0.18	0.25	100
100	100	0.19	0.22	0.32	0.30	0.43	0.52	100
100	1,000	2.45	2.97	3.70	3.18	5.07	5.17	100
1,000	1	0.61	0.64	0.78	0.71	0.87	1.47	100
1,000	10	0.65	0.68	0.93	0.85	0.93	1.59	100
1,000	100	1.14	1.18	1.49	1.27	1.60	2.39	100
1,000	1,000	8.86	9.83	10.61	10.15	10.93	14.31	100
10,000	1	5.92	6.10	8.63	7.55	12.62	13.04	100
10,000	10	6.30	8.26	10.23	8.50	13.59	14.06	100
10,000	100	11.73	12.25	13.48	12.51	14.94	15.95	100
10,000	1,000	59.01	71.82	85.38	82.76	105.11	107.59	100
Lee (2017)								
n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	0.05	0.09	0.13	0.13	0.17	0.21	100
100	10	0.12	0.16	0.19	0.19	0.22	0.32	100
100	100	1.48	1.52	2.18	1.67	3.20	3.29	100
100	1,000	84.03	107.65	127.79	119.68	157.33	159.34	100
1,000	1	0.40	0.41	0.46	0.42	0.43	1.06	100
1,000	10	1.00	1.03	1.09	1.04	1.05	2.54	100
1,000	100	14.77	15.26	16.39	15.54	15.75	32.20	100
1,000	1,000	936.72	1025.59	1104.67	1096.69	1164.58	1623.50	100

10,000	1	10.18	10.41	14.61	15.71	18.78	19.27	100
10,000	10	16.40	19.40	28.97	33.12	34.12	34.86	100
10,000	100	181.38	217.64	277.03	321.77	326.83	337.06	100
10,000	1,000	9492.12	10256.15	12951.71	12094.95	15856.51	16351.42	100

Gorfine (2020)

n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	0.13	0.16	0.19	0.18	0.21	0.51	100
100	10	1.20	1.29	1.67	1.38	1.62	8.44	100
1,000	1	8.46	9.65	11.47	10.83	12.61	26.14	100
1,000	10	95.07	107.76	121.68	117.80	131.12	234.44	100
1,000	100	20131.32	22732.97	25164.28	24467.50	26849.99	38980.33	100
10,000	1	1100.00	1338.07	1591.24	1511.90	1762.02	3437.36	100
10,000	10	13071.19	15327.17	17546.62	16892.73	19025.05	37132.49	100

Katz (2022)

n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	10	19.72	26.72	36.69	31.00	46.67	63.01	100
100	100	27.09	34.72	44.81	38.97	54.43	78.28	100
100	1,000	460.74	560.28	684.37	617.02	697.85	1516.64	100
1,000	10	50.37	60.61	70.00	67.09	72.29	121.66	100
1,000	100	365.09	396.79	457.19	425.80	484.07	944.32	100
1,000	1,000	30032.24	50899.95	63211.32	65687.53	76722.11	90226.82	100
10,000	10	1750.71	2496.33	2930.40	2741.99	3311.81	6134.17	100
10,000	100	11230.41	17303.98	24675.89	25240.58	29378.86	48944.73	100

Lee (2021)

n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	104.30	116.37	132.15	118.96	155.25	249.11	100
100	10	343.54	414.40	458.75	420.26	540.05	856.34	100
1,000	1	936.66	1027.33	1251.10	1045.04	1461.75	1945.39	100
1,000	10	2693.38	3062.60.45	3588.61	3494.45	3555.48	6697.34	100
10,000	1	8557.45	11175.03	12097.27	11621.45	12970.02	21713.81	100
10,000	10	22585.55	24578.09	26138.20	25724.73	27219.38	40697.74	100

Ishwaran (2022)

n	*p*	Min.	Q1	Mean	Median	Q3	Max.	N Eval.
100	1	0.12	0.13	0.16	0.16	0.18	0.26	100
100	10	0.16	0.19	0.25	0.26	0.27	0.46	100
100	100	0.36	0.37	0.49	0.50	0.55	0.84	100
100	1,000	1.46	1.50	2.01	2.09	2.34	3.53	100
1,000	1	5.46	6.58	7.40	7.48	8.17	9.73	100
1,000	10	8.21	9.00	10.80	11.38	11.75	16.99	100
1,000	100	14.80	15.67	19.58	20.84	21.40	31.64	100
1,000	1,000	51.97	53.43	68.16	71.23	75.42	116.54	100
10,000	1	288.46	292.23	351.67	356.61	368.24	574.25	100
10,000	10	387.49	390.43	469.78	489.85	501.41	736.48	100
10,000	100	649.00	653.74	805.57	834.87	848.62	1487.67	100
10,000	1,000	1836.74	1849.18	2248.28	2390.14	2454.35	2749.12	100

D.4 Sensitivity Analysis: Varying Additional Parameters

In a sensitivity analysis, we expanded the settings of our main simulation to additionally vary $\theta \in \{0.5, 2.0\}$, to contrast settings of moderate versus strong dependence between the state transitions, and to vary the censoring rate $\in \{25\%, 50\%\}$ to better understand how censoring impacts the various methods.

In terms of estimating θ, the proposed method had the lowest bias in 12 of the 16 expanded settings, specifically in the eight settings where the true risk functions were nonlinear and in four of the eight linear settings (Table D.4). Among the four linear settings where the proposed method did not have the lowest bias, the results were comparable to the methods of [4] and [2], which have the best performances. We also note that the estimated θ are slightly closer to the truth for smaller values of θ. Table D.5 then compared each approach in terms of the MISE for the predicted log-risk functions. Across all methods, the MISE increased slightly with the frailty variance and censoring rate. In addition, the variability decreased with increasing sample size. Despite being a highly nonlinear approximation, the proposed method performed comparably to the better performing methods, across all state transitions, when the true underlying function of the predictors was linear. In all nonlinear settings, our approach had the lowest MISEs, suggesting that our method outperforms the (semi-)parametric approaches when the functional form of the predictors is truly nonlinear. We also studied the estimation of the baseline hazard functions across all state transitions in Table D.6. These results showed that the proposed method had the lowest bias in 10 of the 16 cases, including all eight with nonlinear covariate effects. Moreover, the proposed method had a bias comparable to the better performing methods in the remaining cases, even though our method is a nonparametric approximation. Lastly, across all methods and settings, we saw that there is more variability as the censoring rate increases.

Finally, to assess predictive accuracy, we calculated the integrated bivariate Brier score at one year over

a sequence of 100 evenly spaced time points and compared the results of our method with the existing methods. As shown, the proposed method had the lowest integrated bivariate Brier score (lower = better) in 13 of the 16 settings (Table D.7). Similarly, the proposed method had the highest bivariate C-index (higher = better) in 14 of 16 simulation settings, and a comparable bivariate C-index in the two settings where it was not strictly the highest (Table D.8).

Table D.4: Estimated mean frailty variance and empirical standard errors (parentheses), averaged over 500 replicates for each simulation setting. Bold values denote the method with the lowest bias.

Simulation Settings				Methods					
n	Risk Function	Censoring Rate	θ	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Proposed
1,000	Linear	25%	0.5	0.43 (0.09)	0.93 (0.05)	0.54 (0.26)	0.76 (0.26)	0.65 (0.12)	**0.51 (0.03)**
5,000	Linear	25%	0.5	0.42 (0.04)	0.92 (0.04)	0.55 (0.26)	0.48 (0.11)	0.68 (0.12)	**0.50 (0.02)**
1,000	Linear	25%	2.0	1.45 (0.18)	0.79 (0.06)	0.57 (0.26)	**2.04 (0.57)**	1.40 (0.19)	2.06 (0.03)
5,000	Linear	25%	2.0	1.44 (0.08)	0.78 (0.06)	0.57 (0.25)	**1.97 (0.21)**	1.48 (0.08)	2.08 (0.01)
1,000	Nonlinear	25%	0.5	0.70 (0.05)	0.52 (0.17)	0.57 (0.26)	1.74 (0.39)	2.59 (3.12)	**0.49 (0.01)**
5,000	Nonlinear	25%	0.5	0.71 (0.02)	0.70 (0.05)	0.55 (0.25)	1.48 (0.17)	2.21 (1.67)	**0.46 (0.01)**
1,000	Nonlinear	25%	2.0	0.83 (0.26)	0.73 (0.28)	0.55 (0.27)	3.12 (0.62)	4.93 (5.79)	**2.14 (0.11)**
5,000	Nonlinear	25%	2.0	0.81 (0.11)	0.73 (0.26)	0.56 (0.26)	2.83 (0.26)	4.52 (3.96)	**2.11 (0.01)**
1,000	Linear	50%	0.5	0.37 (0.10)	0.42 (0.13)	0.58 (0.27)	1.12 (0.41)	1.25 (0.21)	**0.57 (0.02)**
5,000	Linear	50%	0.5	0.36 (0.04)	0.40 (0.12)	**0.55 (0.26)**	0.56 (0.15)	1.31 (0.10)	0.58 (0.01)
1,000	Linear	50%	2.0	1.15 (0.18)	0.65 (0.23)	0.58 (0.26)	2.88 (0.83)	2.54 (0.51)	**2.13 (0.02)**
5,000	Linear	50%	2.0	1.13 (0.08)	0.64 (0.22)	0.56 (0.26)	**2.01 (0.29)**	2.68 (0.16)	2.10 (0.01)
1,000	Nonlinear	50%	0.5	1.76 (1.00)	0.54 (0.20)	0.60 (0.27)	2.53 (0.76)	3.65 (2.50)	**0.52 (0.01)**
5,000	Nonlinear	50%	0.5	0.34 (0.06)	0.57 (0.19)	0.57 (0.26)	1.62 (0.32)	3.37 (1.99)	**0.54 (0.01)**
1,000	Nonlinear	50%	2.0	0.99 (0.39)	0.67 (0.26)	0.56 (0.26)	3.59 (0.85)	4.94 (4.29)	**2.09 (0.15)**
5,000	Nonlinear	50%	2.0	0.59 (0.10)	0.68 (0.26)	0.59 (0.27)	2.95 (0.34)	4.14 (2.58)	**2.03 (0.01)**

Table D.5: Average (SD) mean integrated squared errors (MISE) for the simulated log-risk surfaces, $h_g(\boldsymbol{X}_i)$, for each state transition hazard; $g = 1, 2, 3$, averaged over 500 replicates for each simulation setting. Bold values denote the method which has the lowest average MISE for each setting and state transition.

Simulation Settings				Methods					
n	Risk Function	Censoring Rate	θ	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Proposed
First Transition: $h_1(\boldsymbol{X}_i)$									
1,000	Linear	25%	0.5	0.15 (0.05)	0.41 (0.09)	0.26 (0.09)	**0.13 (0.05)**	2.97 (0.29)	0.36 (0.11)
5,000	Linear	25%	0.5	0.12 (0.02)	0.32 (0.04)	0.18 (0.04)	**0.11 (0.02)**	5.70 (0.33)	0.28 (0.06)
1,000	Linear	25%	2.0	0.10 (0.04)	0.58 (0.10)	**0.26 (0.09)**	0.56 (0.09)	5.70 (0.65)	0.33 (0.11)
5,000	Linear	25%	2.0	0.04 (0.02)	0.47 (0.06)	**0.16 (0.03)**	0.54 (0.04)	5.38 (0.49)	0.22 (0.04)
1,000	Nonlinear	25%	0.5	4.27 (0.39)	4.47 (0.40)	4.37 (0.38)	4.49 (0.41)	8.04 (0.98)	**4.01 (0.58)**
5,000	Nonlinear	25%	0.5	4.26 (0.17)	4.41 (0.18)	4.27 (0.17)	4.48 (0.18)	8.01 (0.68)	**2.94 (0.23)**
1,000	Nonlinear	25%	2.0	4.31 (0.41)	4.54 (0.41)	4.43 (0.40)	4.56 (0.41)	7.99 (1.14)	**3.82 (0.59)**
5,000	Nonlinear	25%	2.0	4.31 (0.18)	4.49 (0.18)	4.33 (0.17)	4.58 (0.18)	8.06 (0.66)	**2.88 (0.20)**
1,000	Linear	50%	0.5	**0.46 (0.10)**	0.70 (0.11)	0.58 (0.15)	2.47 (0.26)	5.63 (0.63)	0.73 (0.17)
5,000	Linear	50%	0.5	0.43 (0.04)	0.62 (0.05)	0.50 (0.07)	**0.09 (0.02)**	5.40 (0.24)	0.78 (0.08)
1,000	Linear	50%	2.0	0.46 (0.12)	0.89 (0.12)	0.71 (0.16)	**0.45 (0.10)**	5.59 (0.76)	0.88 (0.16)
5,000	Linear	50%	2.0	**0.41 (0.06)**	0.80 (0.08)	0.63 (0.07)	0.42 (0.08)	5.09 (0.31)	0.79 (0.07)
1,000	Nonlinear	50%	0.5	5.14 (0.68)	4.66 (0.42)	4.78 (0.43)	4.40 (0.40)	8.10 (1.02)	**4.07 (0.52)**
5,000	Nonlinear	50%	0.5	4.58 (0.19)	4.60 (0.19)	4.68 (0.19)	4.36 (0.17)	8.06 (0.67)	**3.57 (0.18)**
1,000	Nonlinear	50%	2.0	4.94 (0.63)	4.64 (0.42)	4.66 (0.42)	4.50 (0.40)	8.12 (1.04)	**4.15 (0.48)**
5,000	Nonlinear	50%	2.0	4.50 (0.19)	4.59 (0.19)	4.58 (0.18)	4.51 (0.18)	8.10 (0.67)	**3.31 (0.19)**
Second Transition: $h_2(\boldsymbol{X}_i)$									
1,000	Linear	25%	0.5	**0.04 (0.02)**	0.17 (0.05)	0.08 (0.04)	0.12 (0.04)	2.92 (0.23)	0.17 (0.08)
5,000	Linear	25%	0.5	**0.01 (0.004)**	0.12 (0.02)	0.02 (0.01)	0.11 (0.02)	5.69 (0.31)	0.09 (0.03)
1,000	Linear	25%	2.0	**0.08 (0.03)**	0.45 (0.08)	0.13 (0.05)	0.54 (0.08)	5.59 (0.50)	0.29 (0.11)
5,000	Linear	25%	2.0	**0.03 (0.01)**	0.36 (0.05)	0.05 (0.02)	0.54 (0.03)	5.36 (0.46)	0.17 (0.04)
1,000	Nonlinear	25%	0.5	4.24 (0.38)	4.42 (0.38)	4.21 (0.36)	4.47 (0.39)	7.92 (0.93)	**4.04 (0.60)**
5,000	Nonlinear	25%	0.5	4.24 (0.17)	4.38 (0.17)	4.16 (0.16)	4.48 (0.17)	7.93 (0.45)	**3.16 (0.27)**
1,000	Nonlinear	25%	2.0	4.28 (0.40)	4.47 (0.40)	4.25 (0.37)	4.55 (0.40)	7.75 (0.98)	**3.77 (0.57)**
5,000	Nonlinear	25%	2.0	4.29 (0.17)	4.45 (0.18)	4.20 (0.16)	4.58 (0.18)	7.78 (0.45)	**2.95 (0.23)**
1,000	Linear	50%	0.5	**0.05 (0.02)**	0.16 (0.05)	0.10 (0.05)	2.45 (0.25)	5.53 (0.46)	0.26 (0.14)
5,000	Linear	50%	0.5	**0.02 (0.01)**	0.10 (0.02)	0.04 (0.02)	0.08 (0.02)	5.39 (0.19)	0.16 (0.05)
1,000	Linear	50%	2.0	**0.12 (0.05)**	0.39 (0.08)	0.16 (0.06)	0.43 (0.08)	5.39 (0.54)	0.35 (0.11)
5,000	Linear	50%	2.0	**0.07 (0.02)**	0.31 (0.04)	**0.07 (0.03)**	0.41 (0.07)	5.02 (0.23)	0.21 (0.04)
1,000	Nonlinear	50%	0.5	5.01 (0.81)	4.35 (0.37)	4.31 (0.37)	4.37 (0.38)	7.93 (0.94)	**4.28 (0.55)**
5,000	Nonlinear	50%	0.5	4.27 (0.17)	4.31 (0.17)	4.21 (0.17)	4.36 (0.17)	7.92 (0.45)	**3.42 (0.32)**
1,000	Nonlinear	50%	2.0	4.84 (0.70)	4.44 (0.39)	4.32 (0.38)	4.49 (0.39)	7.90 (0.96)	**4.03 (0.58)**
5,000	Nonlinear	50%	2.0	4.33 (0.17)	4.41 (0.17)	4.26 (0.17)	4.51 (0.18)	7.91 (0.46)	**3.16 (0.25)**
Third Transition: $h_3(\boldsymbol{X}_i)$									
1,000	Linear	25%	0.5	**0.15 (0.05)**	0.41 (0.09)	0.16 (0.07)	0.13 (0.05)	2.97 (0.29)	0.33 (0.12)
5,000	Linear	25%	0.5	0.12 (0.02)	0.32 (0.04)	**0.05 (0.02)**	0.11 (0.02)	5.70 (0.33)	0.15 (0.05)
1,000	Linear	25%	2.0	**0.10 (0.04)**	0.58 (0.10)	0.29 (0.12)	0.56 (0.09)	5.70 (0.65)	0.40 (0.14)
5,000	Linear	25%	2.0	**0.04 (0.02)**	0.47 (0.06)	0.13 (0.04)	0.54 (0.04)	5.38 (0.50)	0.25 (0.06)
1,000	Nonlinear	25%	0.5	4.27 (0.39)	4.47 (0.40)	4.52 (0.43)	4.49 (0.41)	8.04 (0.98)	**4.19 (0.49)**
5,000	Nonlinear	25%	0.5	4.26 (0.17)	4.41 (0.18)	4.35 (0.18)	4.48 (0.18)	8.01 (0.68)	**3.27 (0.25)**
1,000	Nonlinear	25%	2.0	4.31 (0.41)	4.54 (0.41)	4.64 (0.45)	4.56 (0.41)	7.99 (1.14)	**4.05 (0.56)**
5,000	Nonlinear	25%	2.0	4.31 (0.18)	4.49 (0.18)	4.45 (0.19)	4.58 (0.18)	8.06 (0.66)	**3.17 (0.24)**
1,000	Linear	50%	0.5	0.46 (0.10)	0.70 (0.11)	**0.23 (0.10)**	2.47 (0.26)	5.63 (0.63)	0.43 (0.18)
5,000	Linear	50%	0.5	0.43 (0.05)	0.61 (0.06)	**0.08 (0.03)**	0.09 (0.02)	5.40 (0.24)	0.31 (0.12)
1,000	Linear	50%	2.0	0.46 (0.12)	0.89 (0.12)	**0.32 (0.13)**	0.45 (0.10)	5.59 (0.76)	0.61 (0.18)
5,000	Linear	50%	2.0	0.41 (0.06)	0.80 (0.08)	**0.14 (0.05)**	0.42 (0.08)	5.09 (0.32)	0.51 (0.09)
1,000	Nonlinear	50%	0.5	5.14 (0.68)	4.66 (0.42)	4.70 (0.45)	4.40 (0.40)	8.10 (1.02)	**4.39 (0.58)**
5,000	Nonlinear	50%	0.5	4.57 (0.19)	4.60 (0.19)	4.40 (0.19)	4.36 (0.17)	8.06 (0.67)	**3.42 (0.27)**
1,000	Nonlinear	50%	2.0	4.94 (0.63)	4.64 (0.42)	4.68 (0.46)	4.50 (0.40)	8.12 (1.05)	**4.47 (0.54)**
5,000	Nonlinear	50%	2.0	4.50 (0.19)	4.59 (0.19)	4.45 (0.20)	4.51 (0.18)	8.10 (0.67)	**3.56 (0.28)**

Table D.6: Average bias (SD) for estimating the baseline hazards, $\lambda_{0g}(t); g = 1, 2, 3$, evaluated at $t = 1$ and averaged over all g baselines. Bold values denote the method which has the lowest average bias.

Simulation Settings				Methods					
n	Risk Function	Censoring Rate	θ	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Proposed
1,000	Linear	25%	0.5	0.28 (0.50)	**0.20 (0.30)**	10.46 (3.52)	0.96 (0.19)	4.36 (2.71)	0.27 (0.44)
5,000	Linear	25%	0.5	**0.15 (0.21)**	0.19 (0.29)	10.52 (3.61)	1.05 (0.89)	4.42 (3.12)	0.17 (0.25)
1,000	Linear	25%	2.0	**0.36 (0.53)**	1.62 (1.59)	7.37 (3.50)	2.37 (0.10)	4.20 (2.45)	0.57 (0.58)
5,000	Linear	25%	2.0	**0.32 (0.44)**	1.62 (1.59)	7.39 (3.49)	2.62 (2.09)	4.24 (3.17)	0.47 (0.39)
1,000	Nonlinear	25%	0.5	1.00 (0.91)	1.08 (0.94)	11.14 (4.16)	1.56 (0.12)	4.62 (4.86)	**0.25 (0.47)**
5,000	Nonlinear	25%	0.5	0.92 (0.72)	1.07 (0.87)	11.11 (4.20)	2.41 (0.84)	4.38 (2.37)	**0.37 (0.31)**
1,000	Nonlinear	25%	2.0	1.72 (1.31)	1.94 (1.41)	8.79 (4.07)	2.37 (0.09)	4.21 (2.46)	**0.22 (0.46)**
5,000	Nonlinear	25%	2.0	1.70 (1.20)	1.91 (1.36)	8.82 (4.03)	3.30 (1.50)	4.25 (2.44)	**0.12 (0.08)**
1,000	Linear	50%	0.5	1.40 (2.38)	**0.94 (1.39)**	10.83 (3.28)	1.17 (0.98)	4.69 (2.32)	0.95 (1.49)
5,000	Linear	50%	0.5	1.00 (1.44)	0.84 (1.19)	10.92 (3.32)	1.07 (0.90)	4.64 (2.30)	**0.46 (0.57)**
1,000	Linear	50%	2.0	**1.02 (1.31)**	1.15 (1.55)	9.02 (2.88)	2.82 (2.00)	4.58 (2.23)	1.12 (2.48)
5,000	Linear	50%	2.0	0.78 (0.68)	1.12 (1.54)	9.11 (2.94)	2.65 (2.14)	4.60 (2.38)	**0.57 (0.69)**
1,000	Nonlinear	50%	0.5	2.39 (1.54)	1.45 (1.17)	12.14 (3.32)	3.31 (2.37)	4.66 (2.36)	**0.31 (0.27)**
5,000	Nonlinear	50%	0.5	1.15 (0.71)	1.39 (0.97)	12.15 (3.17)	2.44 (0.85)	4.73 (2.27)	**0.17 (0.20)**
1,000	Nonlinear	50%	2.0	2.51 (1.58)	1.74 (1.45)	10.00 (3.66)	3.71 (1.64)	4.46 (2.31)	**0.27 (0.45)**
5,000	Nonlinear	50%	2.0	1.79 (1.40)	1.70 (1.38)	10.08 (3.66)	3.34 (1.47)	4.57 (2.42)	**0.13 (0.15)**

Table D.7: Average (SD) one year integrated bivarate Brier scores (iBBS). Bold values denote the method which has the lowest (lower values = higher predictive accuracy) average iBBS for each simulation setting.

Simulation Settings				Methods					
n	Risk Function	Censoring Rate	θ	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Proposed
1,000	Linear	25%	0.5	0.09 (0.005)	0.09 (0.006)	0.10 (0.006)	0.07 (0.01)	0.47 (0.02)	**0.05 (0.003)**
5,000	Linear	25%	0.5	0.09 (0.002)	0.09 (0.004)	0.09 (0.005)	0.06 (0.004)	0.50 (0.02)	**0.05 (0.001)**
1,000	Linear	25%	2.0	0.23 (0.009)	0.21 (0.010)	0.23 (0.010)	0.21 (0.05)	0.51 (0.02)	**0.16 (0.006)**
5,000	Linear	25%	2.0	**0.16 (0.003)**	0.22 (0.007)	0.20 (0.006)	0.26 (0.01)	0.51 (0.02)	**0.16 (0.003)**
1,000	Nonlinear	25%	0.5	0.15 (0,009)	0.14 (0.009)	0.14 (0.010)	0.18 (0.03)	0.52 (0.13)	**0.07 (0.005)**
5,000	Nonlinear	25%	0.5	0.15 (0.004)	0.13 (0.007)	0.16 (0.005)	0.17 (0.01)	0.51 (0.10)	**0.06 (0.003)**
1,000	Nonlinear	25%	2.0	0.24 (0.010)	0.18 (0.007)	0.21 (0.008)	0.33 (0.03)	0.55 (0.11)	**0.16 (0.006)**
5,000	Nonlinear	25%	2.0	0.18 (0.003)	0.18 (0.003)	0.19 (0.003)	0.32 (0.02)	0.52 (0.09)	**0.10 (0.005)**
1,000	Linear	50%	0.5	0.06 (0.009)	**0.03 (0.005)**	0.05 (0.006)	0.11 (0.02)	0.58 (0.03)	**0.03 (0.005)**
5,000	Linear	50%	0.5	**0.03 (0.001)**	**0.03 (0.001)**	**0.03 (0.001)**	0.06 (0.01)	0.58 (0.02)	**0.03 (0.002)**
1,000	Linear	50%	2.0	**0.03 (0.001)**	0.10 (0.007)	0.07 (0.003)	0.31 (0.05)	0.61 (0.03)	0.04 (0.006)
5,000	Linear	50%	2.0	0.10 (0.003)	0.10 (0.003)	0.10 (0.003)	0.27 (0.02)	0.61 (0.02)	**0.02 (0.001)**
1,000	Nonlinear	50%	0.5	0.15 (0.067)	**0.05 (0.006)**	0.08 (0.009)	0.22 (0.05)	0.62 (0.09)	0.14 (0.004)
5,000	Nonlinear	50%	0.5	**0.05 (0.003)**	**0.05 (0.003)**	**0.05 (0.003)**	0.17 (0.02)	0.54 (0.06)	0.09 (0.004)
1,000	Nonlinear	50%	2.0	0.19 (0.016)	**0.13 (0.007)**	0.17 (0.011)	0.36 (0.04)	0.61 (0.10)	**0.13 (0.010)**
5,000	Nonlinear	50%	2.0	0.13 (0.003)	0.13 (0.004)	0.13 (0.004)	0.33 (0.03)	0.58 (0.07)	**0.09 (0.005)**

Table D.8: Average (SD) bivarate concordance (C) index. Bold values denote the method which has the highest (higher values = higher predictive accuracy) average bivarate C-index for each simulation setting.

Simulation Settings				Methods					
n	Risk Function	Censoring Rate	θ	Xu (2010)	Lee (2015)	Lee (2017)	Gorfine (2020)	Kats (2022)	Proposed
1,000	Linear	25%	0.5	**0.67 (0.01)**	0.66 (0.01)	0.53 (0.02)	0.64 (0.02)	0.50 (0.02)	**0.67 (0.01)**
5,000	Linear	25%	0.5	0.63 (0.01)	**0.68 (0.01)**	0.56 (0.02)	0.66 (0.01)	0.34 (0.02)	0.67 (0.01)
1,000	Linear	25%	2.0	0.63 (0.01)	0.63 (0.02)	0.51 (0.02)	0.61 (0.06)	0.41 (0.02)	**0.64 (0.01)**
5,000	Linear	25%	2.0	0.63 (0.01)	0.62 (0.01)	0.51 (0.02)	0.54 (0.01)	0.42 (0.02)	**0.64 (0.01)**
1,000	Nonlinear	25%	0.5	0.61 (0.01)	0.61 (0.01)	0.52 (0.02)	0.53 (0.02)	0.43 (0.03)	**0.72 (0.03)**
5,000	Nonlinear	25%	0.5	0.64 (0.01)	0.62 (0.02)	0.54 (0.01)	0.53 (0.01)	0.45 (0.02)	**0.71 (0.01)**
1,000	Nonlinear	25%	2.0	0.59 (0.02)	0.59 (0.02)	0.53 (0.02)	0.51 (0.02)	0.46 (0.03)	**0.69 (0.01)**
5,000	Nonlinear	25%	2.0	0.63 (0.01)	0.60 (0.01)	0.53 (0.01)	0.51 (0.01)	0.46 (0.02)	**0.68 (0.01)**
1,000	Linear	50%	0.5	0.64 (0.01)	0.64 (0.01)	0.56 (0.03)	0.50 (0.02)	0.38 (0.02)	**0.65 (0.01)**
5,000	Linear	50%	0.5	**0.67 (0.01)**	0.65 (0.02)	0.56 (0.02)	0.64 (0.01)	0.39 (0.01)	0.65 (0.01)
1,000	Linear	50%	2.0	0.62 (0.02)	0.62 (0.02)	0.53 (0.02)	0.54 (0.02)	0.44 (0.02)	**0.63 (0.01)**
5,000	Linear	50%	2.0	0.59 (0.01)	**0.63 (0.01)**	0.54 (0.02)	0.55 (0.01)	0.44 (0.01)	**0.63 (0.02)**
1,000	Nonlinear	50%	0.5	0.58 (0.02)	0.62 (0.02)	0.60 (0.02)	0.53 (0.02)	0.45 (0.02)	**0.69 (0.01)**
5,000	Nonlinear	50%	0.5	0.59 (0.01)	0.61 (0.01)	0.60 (0.02)	0.54 (0.01)	0.44 (0.01)	**0.70 (0.01)**
1,000	Nonlinear	50%	2.0	0.62 (0.02)	0.60 (0.02)	0.62 (0.02)	0.52 (0.02)	0.46 (0.02)	**0.67 (0.01)**
5,000	Nonlinear	50%	2.0	0.61 (0.01)	0.61 (0.02)	0.62 (0.01)	0.52 (0.01)	0.45 (0.01)	**0.67 (0.01)**

E Additional Data Analysis Results

Table E.1: Log hazard ratio estimates and standard errors for three transition types in classical semi-competing risk models.

Characteristic	**Xu (2010)**	**Lee (2015)**	**Lee (2017)**
Progression			
Age at Diagnosis (yrs.)	0.01 (0.002)	-0.04 (0.002)	0.02 (0.02)
Female (vs. Male)	-0.19 (0.05)	-0.03 (0.03)	0.07 (0.08)
Unknown Sex (vs. Male)	0.18 (0.80)	0.90 (0.71)	-0.26 (0.19)
Other Race (vs. White/Caucasian)	0.45 (0.20)	0.42 (0.14)	-0.55 (0.13)
Asian (vs. White/Caucasian)	0.05 (0.16)	0.35 (0.28)	0.11 (0.62)
Black/African American (vs. White/Caucasian)	0.15 (0.18)	-0.01 (0.22)	0.06 (0.13)
Unknown Race (vs. White/Caucasian)	-0.43 (0.18)	0.09 (0.16)	0.18 (0.15)
Hispanic (vs. Non-Hispanic)	0.36 (0.21)	0.02 (0.42)	-0.15 (0.17)
Unknown Ethnicity (vs. Non-Hispanic)	0.21 (0.07)	-0.03 (0.05)	-0.51 (0.13)
COPD[1]	0.33 (0.07)	0.22 (0.14)	-0.50 (0.24)
Unknown COPD Status[1]	-0.85 (0.10)	-0.22 (0.12)	0.17 (0.14)
Asthma	0.10 (0.11)	-0.29 (0.19)	0.21 (0.27)
Unknown Asthma Status	1.43 (0.10)	0.65 (0.29)	-0.57 (0.13)
Late Stage (3B-4, vs. Early Stage 1-3A)	0.52 (0.07)	0.08 (0.14)	-0.33 (0.12)
Radiation (vs. Chemotherapy)	-0.17 (0.12)	0.32 (0.18)	-0.40 (0.07)
Surgery (vs. Chemotherapy)	-1.08 (0.08)	-0.14 (0.04)	0.35 (0.16)
Other First-Line Treatment (vs. Chemotherapy)	0.36 (0.56)	-0.004 (0.00)	-0.16 (0.27)
Unknown First-Line Treatment (vs. Chemotherapy)	-1.70 (0.16)	-0.98 (0.60)	0.58 (0.14)
Former Smoker (vs. Never Smoker)	0.22 (0.08)	0.06 (0.02)	-0.23 (0.19)
Current Smoker (vs. Never Smoker)	0.40 (0.10)	-0.23 (0.08)	-0.30 (0.34)
Smoker, Status Unknown (vs. Never Smoker)	0.14 (0.22)	0.01 (0.07)	-0.13 (0.20)
Pack-Years of Smoking	-0.003 (0.001)	0.02 (0.004)	0.03 (0.03)
EGFR Mutation	-0.19 (0.11)	0.25 (0.13)	0.06 (0.67)
KRAS Mutation	0.30 (0.10)	0.90 (0.11)	-0.43 (0.24)
No Genetic Testing	-0.49 (0.07)	-0.18 (0.15)	0.47 (0.28)
Death			
Age at Diagnosis (yrs.)	0.03 (0.002)	0.03 (0.00)	-0.01 (0.01)
Female (vs. Male)	-0.39 (0.05)	0.27 (0.18)	0.79 (0.11)
Unknown Sex (vs. Male)	-0.63 (1.48)	-0.02 (1.10)	-0.06 (0.38)
Other Race (vs. White/Caucasian)	0.32 (0.22)	0.16 (0.20)	0.31 (0.17)
Asian (vs. White/Caucasian)	-0.18 (0.19)	0.07 (0.27)	0.69 (0.17)
Black/African American (vs. White/Caucasian)	0.14 (0.17)	0.08 (0.001)	0.50 (0.28)
Unknown Race (vs. White/Caucasian)	-0.60 (0.20)	0.02 (0.23)	0.31 (0.12)
Hispanic (vs. Non-Hispanic)	0.30 (0.25)	0.23 (0.30)	0.31 (0.33)
Unknown Ethnicity (vs. Non-Hispanic)	0.23 (0.09)	-0.27 (0.07)	0.25 (0.21)
COPD[1]	-0.12 (0.06)	-0.57 (0.37)	0.15 (0.32)
Unknown COPD Status[1]	0.19 (0.06)	-0.05 (0.03)	0.30 (0.33)
Asthma	-0.16 (0.10)	0.29 (0.23)	0.48 (0.19)
Unknown Asthma Status	-0.41 (0.07)	-0.19 (0.06)	0.67 (0.18)
Late Stage (3B-4, vs. Early Stage 1-3A)	1.56 (0.07)	0.97 (0.31)	-0.13 (0.20)
Radiation (vs. Chemotherapy)	-0.01 (0.11)	0.04 (0.15)	-0.17 (0.14)
Surgery (vs. Chemotherapy)	-1.11 (0.07)	-0.16 (0.05)	0.66 (0.23)
Other First-Line Treatment (vs. Chemotherapy)	-0.61 (0.95)	-1.19 (2.31)	0.31 (0.56)
Unknown First-Line Treatment (vs. Chemotherapy)	0.39 (0.07)	1.56 (0.15)	-0.61 (0.58)

Former Smoker (vs. Never Smoker)	0.23 (0.08)	-0.06 (0.03)	0.10 (0.12)
Current Smoker (vs. Never Smoker)	0.50 (0.09)	-0.12 (0.13)	-0.23 (0.09)
Smoker, Status Unknown (vs. Never Smoker)	0.23 (0.23)	-0.09 (0.11)	0.12 (0.19)
Pack-Years of Smoking	0.003 (0.001)	0.07 (0.01)	0.02 (0.002)
EGFR Mutation	-0.42 (0.16)	0.04 (0.27)	0.68 (0.21)
KRAS Mutation	0.10 (0.14)	0.01 (0.10)	0.50 (0.16)
No Genetic Testing	0.28 (0.08)	0.12 (0.10)	0.14 (0.41)
Progression → Death			
Age at Diagnosis (yrs.)	0.04 (0.003)	0.03 (0.001)	0.02 (0.01)
Female (vs. Male)	-0.32 (0.07)	-0.37 (0.10)	-0.13 (0.12)
Unknown Sex (vs. Male)	-0.58 (1.29)	0.68 (0.73)	0.33 (0.28)
Other Race (vs. White/Caucasian)	-0.16 (0.27)	-0.24 (0.12)	-0.004 (0.10)
Asian (vs. White/Caucasian)	-0.23 (0.23)	-0.27 (0.19)	-0.02 (0.09)
Black/African American (vs. White/Caucasian)	0.19 (0.23)	0.04 (0.08)	-0.41 (0.10)
Unknown Race (vs. White/Caucasian)	-0.27 (0.26)	0.06 (0.49)	0.21 (0.20)
Hispanic (vs. Non-Hispanic)	0.36 (0.21)	0.02 (0.42)	-0.15 (0.17)
Unknown Ethnicity (vs. Non-Hispanic)	0.21 (0.07)	-0.03 (0.05)	-0.51 (0.13)
COPD[1]	0.004 (0.10)	0.18 (0.09)	0.49 (0.06)
Unknown COPD Status[1]	0.33 (0.14)	-0.02 (0.07)	0.22 (0.07)
Asthma	-0.11 (0.16)	-0.07 (0.08)	0.22 (0.11)
Unknown Asthma Status	-0.58 (0.15)	-0.34 (0.11)	0.14 (0.05)
Late Stage (3B-4, vs. Early Stage 1-3A)	0.71 (0.09)	0.45 (0.33)	-0.68 (0.13)
Radiation (vs. Chemotherapy)	0.07 (0.14)	0.41 (0.45)	-0.26 (0.12)
Surgery (vs. Chemotherapy)	-1.16 (0.09)	-0.44 (0.10)	0.69 (0.18)
Other First-Line Treatment (vs. Chemotherapy)	0.36 (0.56)	-0.004 (0.00)	-0.16 (0.27)
Unknown First-Line Treatment (vs. Chemotherapy)	0.01 (0.21)	0.98 (0.39)	-0.65 (0.24)
Former Smoker (vs. Never Smoker)	0.23 (0.11)	0.07 (0.11)	0.46 (0.19)
Current Smoker (vs. Never Smoker)	0.57 (0.13)	0.57 (0.09)	0.80 (0.14)
Smoker, Status Unknown (vs. Never Smoker)	-0.30 (0.30)	-0.23 (0.38)	0.09 (0.28)
Pack-Years of Smoking	0.01 (0.001)	0.01 (0.003)	-0.06 (0.004)
EGFR Mutation	-0.04 (0.15)	-0.37 (0.02)	-0.21 (0.12)
KRAS Mutation	-0.02 (0.12)	-0.01 (0.35)	0.14 (0.13)
No Genetic Testing	0.22 (0.09)	0.12 (0.03)	0.14 (0.12)

[1]COPD: Chronic Obstructive Pulmonary Disease

F Summary of Comparisons Between Semi-Competing Risk Models

Table F.1: Comparison of key modeling assumptions and methodological characteristics, including model structure, flexibility, computational features, and data requirements, between the proposed approach and classical semi-competing risk models.

Assumption	**NEM-SCR**	**Cox PH**	**AFT**	**Implication**
Nonparametric covariate risk functions	✓	✗	✗	NEM-SCR can learn complex risk surfaces; classical models rely on fixed linear predictors.
Does *not* require proportional hazards over time	✓	✗	✓	Cox enforces PH; NEM-SCR and AFT accommodate nonproportional hazards.
Does *not* assume a parametric baseline hazards	✓	✓	✗	NEM-SCR and Cox have flexible baseline hazards; AFT requires a parametric distribution.
Handles high-dimensional covariates without manual feature engineering	✓	✗	✗	Neural networks naturally scale to many predictors; classical models require low-dimensional features.
Does *not* assume a parametric frailty distribution	✗	✗	✗	All make standard assumption of a gamma or log-normal frailty.
No unverifiable assumptions	✓	✓	✓	Note that some methods not considered assume a latent distribution on the lower wedge.
Retains hazard-level interpretability for each transition	✓	✓	✗	NEM-SCR and Cox provide interpretable hazards; AFT focuses on an accelerated time scale.
Requires only standard independent right-censoring assumption	✓	✓	✓	No additional censoring assumptions beyond the usual framework.

PH: Proportional Hazards; AFT: Accelerated Failure Time

References

[1] David R Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, 1972.

[2] Malka Gorfine, Nir Keret, Asaf Ben Arie, David Zucker, and Li Hsu. Marginalized frailty-based illness-death model: application to the uk-biobank survival data. Journal of the American Statistical Association, 116(535):1155–1167, 2021.

[3] Søren Johansen. An extension of cox's regression model. International Statistical Review/Revue Internationale de Statistique, pages 165–174, 1983.

[4] Kyu Ha Lee, Virginie Rondeau, and Sebastien Haneuse. Accelerated failure time models for semi-competing risks data in the presence of complex censoring. Biometrics, 73(4):1401–1412, 2017.

[5] Yi Li and Xihong Lin. Covariate measurement errors in frailty models for clustered survival data. Biometrika, 87(4):849–866, 2000.