Supporting Information for "Penalized Estimation of Frailty-Based Illness-Death Models for Semi-Competing Risks"

Harrison T. Reeder
Massachusetts General Hospital Biostatistics
Department of Medicine, Harvard Medical School
and
Junwei Lu and Sebastien Haneuse
Department of Biostatistics, Harvard T.H. Chan School of Public Health

Introduction

In this appendix we present additional details and results beyond what could be presented in the main manuscript. To distinguish the two documents, alpha-numeric labels are used in this document while numeric labels are used in the main paper. Section A provides additional results from the data application. Section B describes the use of fitted illness-death models for individualized risk prediction, and presents an example using the data application. Section C provides the proof of Theorem 1. Section D defines likelihood-related functions for frailty-based parametric illness death models. Section E defines and proves technical lemmas used in the main result. Section F presents additional simulation details and results. Section G presents additional algorithmic details for optimization. Section H summarizes other spline-based hazard specifications.

A Additional Data Application Results

In this section, we present further additional results for the data application.

Table A.1: Characteristics of the study population, overall and by observed preeclampsia diagnosis and delivery outcome. Abbreviations: abnormal (Abn), white blood cell count (WBC), red blood cell count (RBC), red cell distribution width (RDW), mean corpuscular volume (MCV), gastroesophageal reflux disease (GERD).

	Total Births	Births with Preeclampsia	Births without Preeclampsia
Total	2127 (100%)	189 (100%)	1938 (100%)
Maternal Age ≥ 35	658 (30.9%)	61 (32.3%)	597 (30.8%)
Male Fetus	1074 (50.5%)	94 (49.7%)	980 (50.6%)
Current or Prior Cigarette Use	232 (10.9%)	29 (15.3%)	203 (10.5%)
Previous Preeclampsia	51 (2.4%)	10 (5.3%)	41 (2.1%)
Parity ≥ 1	1075 (50.5%)	58 (30.7%)	1017 (52.5%)
Public or No Insurance	761 (35.8%)	81 (42.9%)	680 (35.1%)
Other/Unknown Race/Ethnicity	431 (20.3%)	54 (28.6%)	377 (19.5%)
Hispanic Race/Ethnicity	194 (9.1%)	12 (6.3%)	182 (9.4%)
Asian Race/Ethnicity	209 (9.8%)	10 (5.3%)	199 (10.3%)
Black Race/Ethnicity	310 (14.6%)	29 (15.3%)	281 (14.5%)
BMI ≥ 30	550 (25.9%)	75 (39.7%)	475 (24.5%)
Pre-existing Diabetes	49 (2.3%)	14 (7.4%)	35 (1.8%)
Anemia	283 (13.3%)	31 (16.4%)	252 (13%)
Retention of Urine	49 (2.3%)	5 (2.6%)	44 (2.3%)
GERD	96 (4.5%)	11 (5.8%)	85 (4.4%)
Asthma	175 (8.2%)	22 (11.6%)	153 (7.9%)
Anxiety Disorder	166 (7.8%)	25 (13.2%)	141 (7.3%)
Mood Disorder	125 (5.9%)	15 (7.9%)	110 (5.7%)
Polycystic Ovarian Syndrome	30 (1.4%)	2 (1.1%)	28 (1.4%)
Hypothyroidism	140 (6.6%)	21 (11.1%)	119 (6.1%)
Leiomyoma	269 (12.6%)	14 (7.4%)	255 (13.2%)
Hepatitis Infection	27 (1.3%)	7 (3.7%)	20 (1%)
Herpesviral Infection	101 (4.7%)	8 (4.2%)	93 (4.8%)
Abn. WBC	462 (21.7%)	41 (21.7%)	421 (21.7%)
Abn. Urine WBC	95 (4.5%)	6 (3.2%)	89 (4.6%)
Abn. Urine Bilinogen	47 (2.2%)	5 (2.6%)	42 (2.2%)
Abn. Urine Specific Gravity	10 (0.5%)	1 (0.5%)	9 (0.5%)
Abn. RBC	322 (15.1%)	19 (10.1%)	303 (15.6%)
Abn. RDW	79 (3.7%)	8 (4.2%)	71 (3.7%)
Abn. Urine RBC	88 (4.1%)	8 (4.2%)	80 (4.1%)
Abn. Platelet Count	55 (2.6%)	4 (2.1%)	51 (2.6%)
Abn. MCV	153 (7.2%)	14 (7.4%)	139 (7.2%)
Abn. Hemoglobin	187 (8.8%)	15 (7.9%)	172 (8.9%)

A.1 Additional Estimation Results

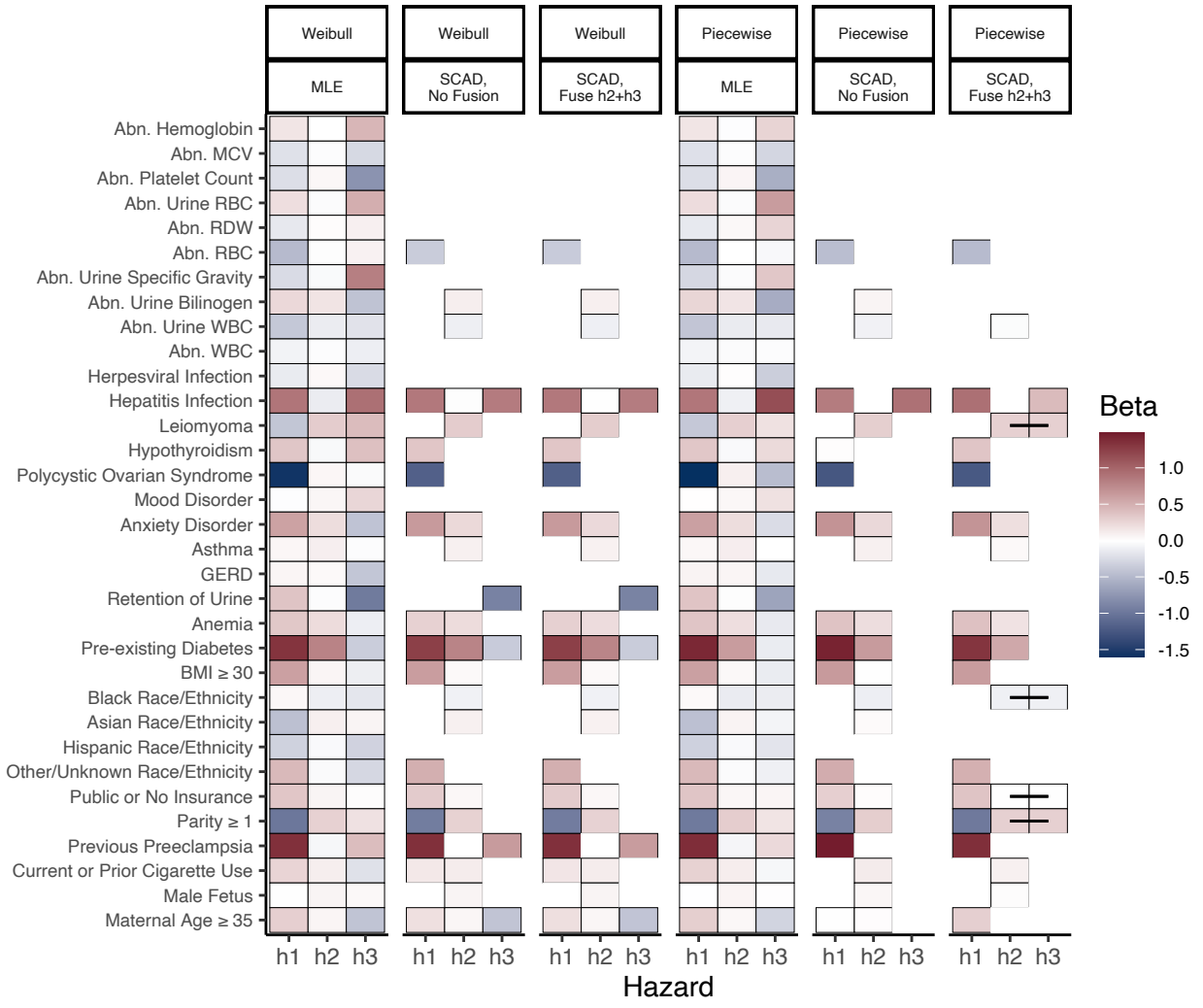

Figure A.1: Estimated coefficients, AIC-optimal SCAD-penalized estimators with and without ℓ_1 fusion between h_2 and h_3, and MLE under Markov specification. Fused coefficients connected with a black line.

Table A.2: Information criterion values for models presented in Figures 2 (BIC) and A.1 (AIC). Lower values reflect improved model fit.

BIC-Optimal Selection	λ_1	λ_2	BIC
Weibull MLE	0	0	-219.039
Weibull, SCAD, No Fusion	0.027	0	-762.464
Weibull, SCAD, Fuse $h_2 + h_3$	0.016	0.015	-784.840
Piecewise MLE	0	0	93.984
Piecewise, SCAD, No Fusion	0.017	0	-467.450
Piecewise, SCAD, Fuse $h_2 + h_3$	0.017	0.015	-487.490
AIC-Optimal Selection	λ_1	λ_2	AIC
Weibull MLE	0	0	-813.599
Weibull, SCAD, No Fusion	0.009	0	-907.093
Weibull, SCAD, Fuse $h_2 + h_3$	0.009	0.001	-907.164
Piecewise MLE	0	0	-568.524
Piecewise, SCAD, No Fusion	0.011	0	-668.483
Piecewise, SCAD, Fuse $h_2 + h_3$	0.010	0.010	-675.313

B Individualized Joint Outcome Risk Prediction

Fitted illness-death models can be used to generate clinically meaningful predictions of individualized joint risk, and the timing of the non-terminal and terminal events. In the preeclampsia setting, the model can predict across time how likely an individual is to be in one of four categories:

(1) still pregnant without preeclampsia

(2) already delivered without preeclampsia

(3) already delivered with preeclampsia, and

(4) still pregnant with preeclampsia.

These four probabilities comprise an individualized "risk profile," and are derived for the illness-death model in Putter et al. (2007) by integrating over regions of the joint density of the semi-competing outcomes (T_1, T_2). The resulting formulas can be concisely represented for both Markov and semi-Markov illness-death models, by defining

$$H_3(t \mid t_1, \mathbf{X}_3) = \begin{cases} H_3(t \mid \mathbf{X}_3) - H_3(t_1 \mid \mathbf{X}_3) & \text{Markov} \\ H_3(t - t_1 \mid \mathbf{X}_3) & \text{semi-Markov.} \end{cases}$$

Then for fixed frailty γ, these four probabilities (numbered as above) are:

$$\begin{aligned}
\pi^{(1)}(t \mid \mathbf{X}, \gamma) &= \exp\{-\gamma[H_1(t \mid \mathbf{X}_1) + H_2(t \mid \mathbf{X}_2)]\} \\
\pi^{(2)}(t \mid \mathbf{X}, \gamma) &= \int_0^t \gamma h_2(t_2 \mid \mathbf{X}_2) \exp\{-\gamma[H_1(t_2 \mid \mathbf{X}_1) + H_2(t_2 \mid \mathbf{X}_2)]\} dt_2 \\
\pi^{(3)}(t \mid \mathbf{X}, \gamma) &= \int_0^t \gamma h_1(t_1 \mid \mathbf{X}_1) \exp\{-\gamma[H_1(t_1 \mid \mathbf{X}_1) + H_2(t_1 \mid \mathbf{X}_2)]\} \\
&\quad \times (1 - \exp\{-\gamma H_3(t \mid t_1, \mathbf{X}_3)\}) dt_1 \\
\pi^{(4)}(t \mid \mathbf{X}, \gamma) &= \int_0^t \gamma h_1(t_1 \mid \mathbf{X}_1) \exp\{-\gamma[H_1(t_1 \mid \mathbf{X}_1) + H_2(t_1 \mid \mathbf{X}_2) + H_3(t \mid t_1, \mathbf{X}_3)\} dt_1,
\end{aligned}$$

and collectively denoted $\boldsymbol{\pi}(t \mid \mathbf{X}, \gamma) = \{\pi^{(1)}(t \mid \mathbf{X}, \gamma), \pi^{(2)}(t \mid \mathbf{X}, \gamma), \pi^{(3)}(t \mid \mathbf{X}, \gamma), \pi^{(4)}(t \mid \mathbf{X}, \gamma)\}$. These probabilities sum to 1, and the integrals can be computed numerically using standard software.

B.1 Sample Predictions from Data Application

Example individualized risk profiles are shown in Figure B.1, generated from the above AIC-selected Weibull model with fusion penalty (shown above in column three of Figure A.1).

The four panels of Figure B.1 correspond with sample individuals having covariates outlined in Table B.1. At each time point, the height of each colored area of the plot gives the probability that the individual will be in the corresponding outcome category at that time, stacking from top to bottom $\pi^{(1)}(t \mid \mathbf{X}, \gamma)$, $\pi^{(2)}(t \mid \mathbf{X}, \gamma)$, $\pi^{(3)}(t \mid \mathbf{X}, \gamma)$, and $\pi^{(4)}(t \mid \mathbf{X}, \gamma)$. For example, the model predicts that at 34 weeks of gestation, individual D has about a 10% chance of having developed preeclampsia and still being pregnant (blue), a 10% chance of having developed preeclampsia and already given birth (purple), a 2% chance of having given birth without preeclampsia (red), and a 78% chance of being pregnant without preeclampsia (grey).

From these detailed risk patterns, we can also read simple overall probabilities of developing preeclampsia by looking at the combined height of the blue and purple bars at the right end of the plot. For patients A-D, the predicted overall probability of developing preeclampsia by week 40 is about 2%, 22%, 52% and 47%, respectively.

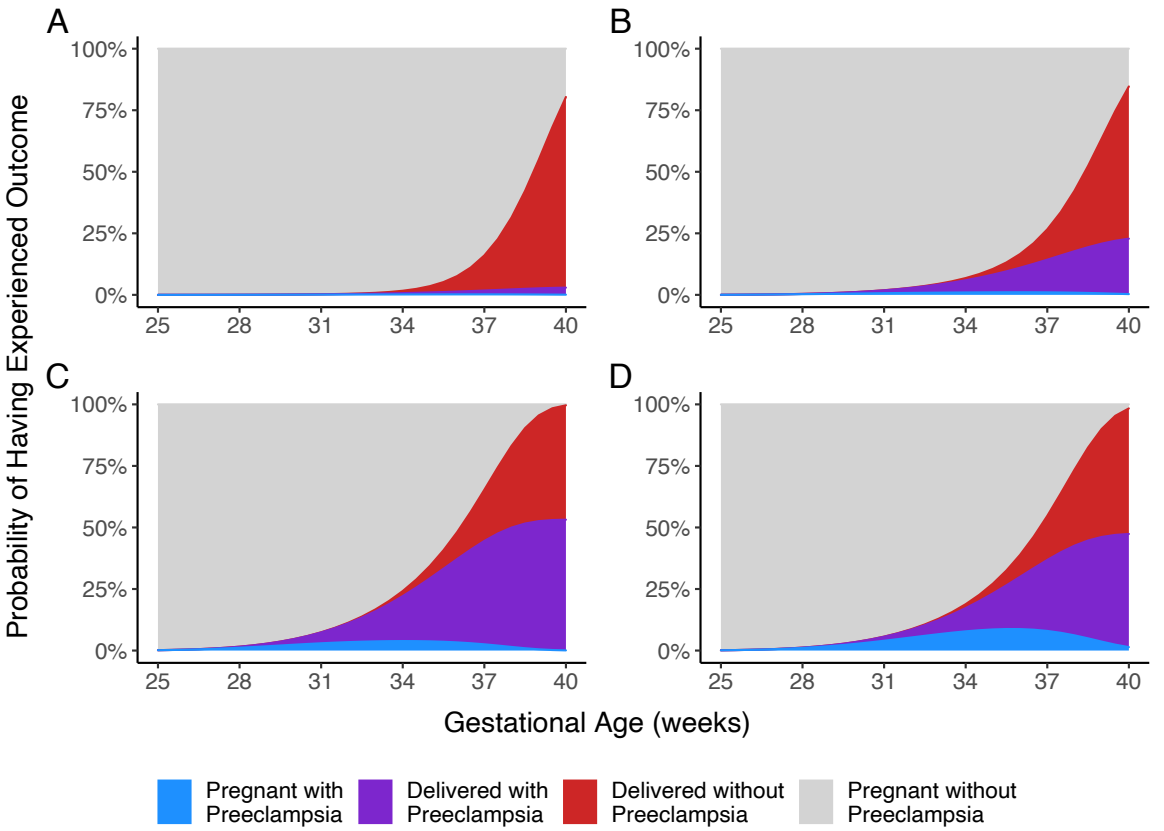

Figure B.1: Sample predicted risk profiles for four sample individuals. AIC-selected Weibull model with SCAD and fusion penalty under Markov specification. Frailty value fixed at $\gamma = 1$.

Table B.1: Characteristics of sample individuals used in Figure B.1.

	A	B	C	D
Abn. RBC	No	No	No	No
Abn. Urine Bilinogen	No	No	No	No
Abn. Urine WBC	Yes	No	No	No
Hepatitis Infection	No	No	No	No
Leiomyoma	No	No	No	No
Hypothyroidism	No	No	Yes	Yes
Polycystic Ovarian Syndrome	No	No	No	No
Anxiety Disorder	No	No	No	No
Asthma	No	No	Yes	No
Retention of Urine	No	No	No	Yes
Anemia	No	Yes	No	No
Pre-existing Diabetes	No	No	Yes	Yes
BMI ≥ 30	No	Yes	Yes	Yes
Race/Ethnicity	White	Other/Unknown	White	White
Public or No Insurance	No	No	No	Yes
Parity ≥ 1	Yes	No	Yes	No
Previous Preeclampsia	No	No	Yes	No
Current or Prior Cigarette Use	No	No	No	No
Male Fetus	Yes	No	No	Yes
Maternal Age ≥ 35	Yes	No	Yes	No

B.2 Use of Frailties in Individualized Risk Prediction

As described in the main text, the patient-specific frailty γ accounts for residual within-patient dependence between (T_1, T_2). Thus, it represents a potentially important component of variation in individualized risk predictions (Putter and Van Houwelingen, 2015). However, γ is latent and, therefore, cannot be observed for an individual to plug into $\boldsymbol{\pi}(t \mid \mathbf{X}, \gamma)$, and must be fixed to a chosen value as in Figure B.1 with $\gamma = 1$. In practice we might also compare predicted risk profiles for an individual across different values of γ, and in this way, the frailty can be viewed as a way of characterizing individual-level variability in the predicted risk profile. For example, Figure B.2 shows predicted risks across frailty values for each sample subject at the fixed timepoint of 37 weeks of gestation. Note that now, the x-axis does not represent time, but differing choices of frailty γ plugged into the above risk profile formulae. Intuitively, we see that the predicted probabilities of experiencing some combination of the outcomes by week 37 tend to be smaller for smaller frailty values, and

larger for larger frailty values. Though individuals' frailties are unobserved, these plots can be used to communicate to patients the variability of possible outcome probability estimates depending on unmeasured latent characteristics (Lee et al., 2020).

Alternatively, this prediction framework also enables risk profile estimates to be marginalized over the frailty distribution by simply integrating over $f_\gamma(\gamma \mid \sigma)$, yielding the marginal profile $\boldsymbol{\pi}(t \mid \mathbf{X}) = \int \boldsymbol{\pi}(t \mid \mathbf{X}, \gamma) f_\gamma(\gamma \mid \sigma) d\gamma$. Corresponding marginal risk profile formulas derived under a gamma frailty distribution are given here for reference:

$$\pi^{(1)}(t \mid \mathbf{X}) = \{1 + e^\sigma [H_1(t_2 \mid \mathbf{X}_1) + H_2(t_2 \mid \mathbf{X}_2)]\}^{-\exp(-\sigma)}$$

$$\pi^{(2)}(t \mid \mathbf{X}) = \int_0^t h_2(t_2 \mid \mathbf{X}_2) \{1 + e^\sigma [H_1(t_2 \mid \mathbf{X}_1) + H_2(t_2 \mid \mathbf{X}_2)]\}^{-\exp(-\sigma)-1} dt_2$$

$$\begin{aligned}\pi^{(3)}(t \mid \mathbf{X}) = &\int_0^t h_1(t_1 \mid \mathbf{X}_1) \{1 + e^\sigma [H_1(t_1 \mid \mathbf{X}_1) + H_2(t_1 \mid \mathbf{X}_2)]\}^{-\exp(-\sigma)-1} dt_1 \\ &- \int_0^t h_1(t_1 \mid \mathbf{X}_1) \{1 + e^\sigma [H_1(t_1 \mid \mathbf{X}_1) + H_2(t_1 \mid \mathbf{X}_2) + H_3(t \mid t_1, \mathbf{X}_3)]\}^{-\exp(-\sigma)-1} dt_1\end{aligned}$$

$$\pi^{(4)}(t \mid \mathbf{X}) = \int_0^t h_1(t_1 \mid \mathbf{X}_1) \{1 + e^\sigma [H_1(t_1 \mid \mathbf{X}_1) + H_2(t_1 \mid \mathbf{X}_2) + H_3(t \mid t_1, \mathbf{X}_3)]\}^{-\exp(-\sigma)-1} dt_1$$

Because in this data application the estimated frailty variance was small, there is little difference between the population-averaged risk profiles for the sample subject compared to Figure B.1, so the plot is omitted.

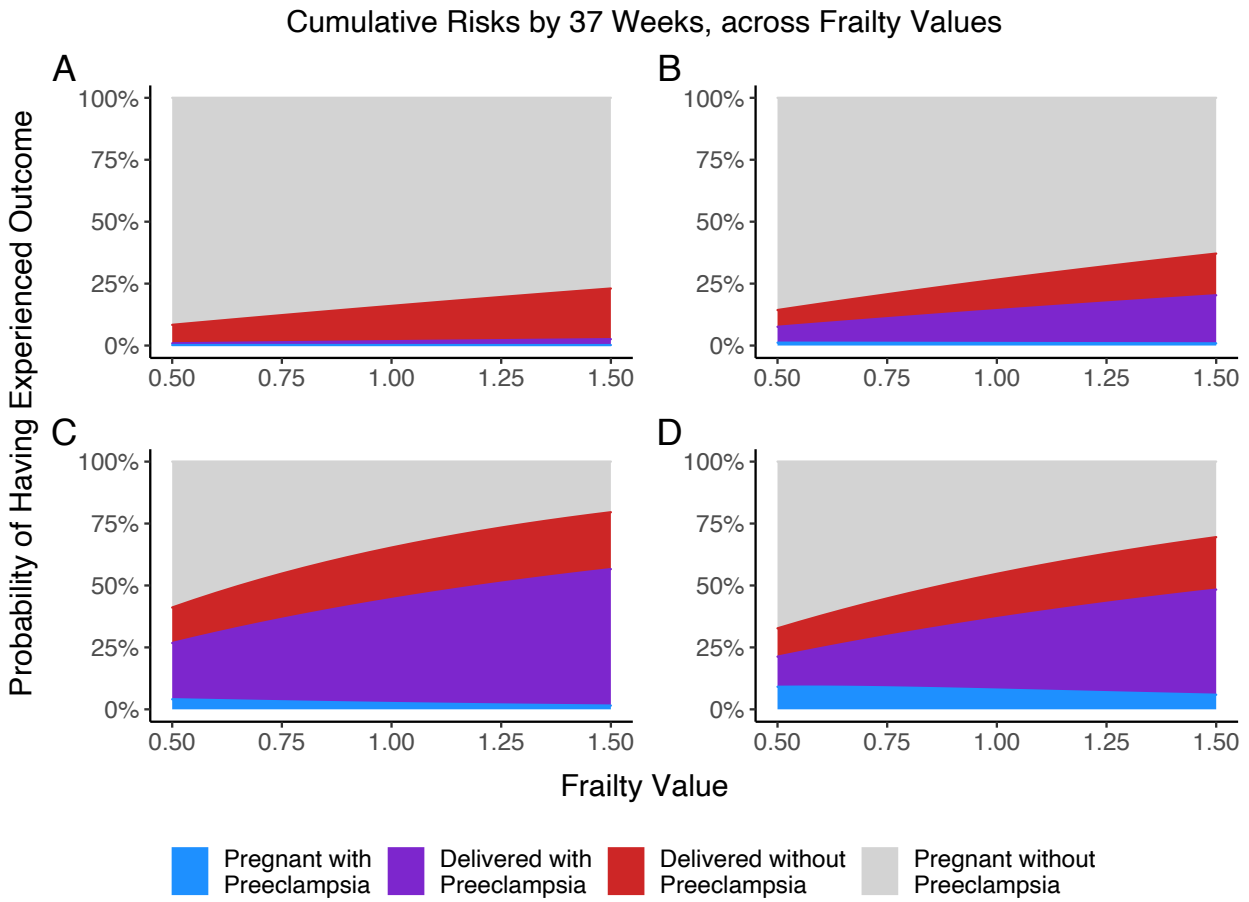

Figure B.2: Predicted risk at 37 weeks of gestation across values of frailty γ for four sample individuals. AIC-selected Weibull model with SCAD and fusion penalty under Markov specification.

C Proof of Main Result

In this section, we present a proof of Theorem 1. This proof follows the strategy of Theorem 1 of Loh and Wainwright (2015), with several differences. In particular, the restricted strong convexity (RSC) condition in Loh and Wainwright (2015) no longer applies in this setting, because under certain baseline hazard specifications such as the Weibull, the resulting baseline parameters $\boldsymbol{\phi}$ have heavy-tailed deviations of the gradient and Hessian that decay at a slower rate (see Lemmas E.2, E.3, and E.4). Therefore, we introduce and apply the alternative RSC condition given in (C.2), which in Lemma E.5 is verified to hold with high probability under the corresponding Assumptions listed in Section 4 of the main text.

Finally, we note that the same statistical rate can be immediately obtained for other choices of penalty function p_λ besides SCAD, as described by Loh and Wainwright (2015), Section 2.2. In particular, the theorem supposition of $3/\{4(\xi-1)\} < \rho$ for SCAD penalties relates the level of penalty non-convexity to the Hessian eigenvalue lower bound, and can be replaced by $3/\{4\xi\} < \rho$ for the MCP penalty of Zhang (2010), or omitted entirely for the lasso penalty. The theorem can be correspondingly adjusted below.

Proof of Theorem 1. To begin, define the linear transformation of the regression parameters $\boldsymbol{\beta}' = ((\boldsymbol{\beta}'_1)^\intercal, (\boldsymbol{\beta}'_2)^\intercal, (\boldsymbol{\beta}'_3)^\intercal) = (\boldsymbol{\beta}_1^\intercal, (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\intercal, (\boldsymbol{\beta}_3 - \boldsymbol{\beta}_1)^\intercal)$. Similarly define $\boldsymbol{\psi}' = ((\boldsymbol{\beta}')^\intercal, \boldsymbol{\phi}^\intercal, \sigma)^\intercal$. Define the contrast matrix $\mathbf{M}$ such that $\boldsymbol{\psi} = \mathbf{M}\boldsymbol{\psi}'$, and corresponding loss $\ell'(\boldsymbol{\psi}') = \ell(\mathbf{M}\boldsymbol{\psi}')$.

Under this transformation the penalized objective function Q_λ given in (10) can be equivalently represented by $\ell'(\boldsymbol{\psi}') + \sum_{g=1}^{3}\sum_{j=1}^{d_g} p_\lambda\left(|\beta'_{gj}|\right)$.

By the chain rule, $\nabla\ell'(\boldsymbol{\psi}') = \mathbf{M}^\intercal\nabla\ell(\boldsymbol{\psi})$ and $\nabla^2\ell'(\boldsymbol{\psi}') = \mathbf{M}^\intercal\nabla^2\ell(\boldsymbol{\psi})\mathbf{M}$. Define the jth column of $\mathbf{M}$ as $\mathbf{M}_{\cdot j}$ and the induced matrix 1-norm as the maximum absolute column sum $\|\mathbf{M}\|_1 = \max_j \|\mathbf{M}_{\cdot j}\|_1$. For this particular transformation, it can be shown that $\|\mathbf{M}\|_1 = 3$. Then repeated application of the triangle inequality and Hölder's inequality illustrates that the maximal values are equivalent up to a known constant:

$$\|\nabla\ell'(\boldsymbol{\psi}')\|_\infty \le \|\mathbf{M}\|_1\|\nabla\ell(\boldsymbol{\psi})\|_\infty = 3\|\nabla\ell(\boldsymbol{\psi})\|_\infty,$$
$$\|\nabla^2\ell'(\boldsymbol{\psi}')\|_{\max} \le \|\mathbf{M}\|_1^2\|\nabla^2\ell(\boldsymbol{\psi})\|_{\max} = 9\|\nabla^2\ell(\boldsymbol{\psi})\|_{\max}.$$

As a result, it suffices to show the desired rate for the estimator

$$\widehat{\boldsymbol{\psi}} = \underset{\|\boldsymbol{\psi}\|_1 \le R_1}{\arg\min}\, \ell(\boldsymbol{\psi}) + \sum_{g=1}^{3}\sum_{j=1}^{d_g} p_\lambda\left(|\beta_{gj}|\right). \tag{C.1}$$

Defining $\boldsymbol{\nu} = \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*$, then by Assumption 2 and the side constraint on (C.1), our stationary point satisfies $\|\boldsymbol{\nu}\|_2 \le \|\boldsymbol{\nu}\|_1 \le R$. Then by Lemma E.5, under Assumptions 1, 3, and 4, there exists a positive constant c establishing the following RSC condition with high probability:

$$\langle\nabla\ell(\boldsymbol{\psi}^* + \boldsymbol{\nu}) - \nabla\ell(\boldsymbol{\psi}^*), \boldsymbol{\nu}\rangle \ge \rho\|\boldsymbol{\nu}\|_2^2 - c\sqrt{\frac{\log(dn)}{n}}\|\boldsymbol{\nu}\|_1^2. \tag{C.2}$$

Recall the defined penalty function $P_\lambda(\boldsymbol{\psi}) = \sum_{g=1}^{3}\sum_{j=1}^{k} p_\lambda(\beta_{gj})$. Setting p_λ to be the SCAD penalty function (6), then $P_\lambda(\boldsymbol{\psi}) + \|\boldsymbol{\psi}\|_2^2/\{2(\xi-1)\}$ is convex, so it follows that

$$\left\langle \nabla P_\lambda(\widehat{\boldsymbol{\psi}}), \boldsymbol{\psi}^* - \widehat{\boldsymbol{\psi}} \right\rangle \leq P_\lambda(\boldsymbol{\psi}^*) - P_\lambda(\widehat{\boldsymbol{\psi}}) + \frac{1}{2(\xi-1)}\|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*\|_2^2,$$

where $\nabla P_\lambda(\widehat{\boldsymbol{\psi}})$ is a subgradient of P_λ at $\widehat{\boldsymbol{\psi}}$. Combining this with (C.2), and the first order condition

$$\langle \nabla \ell(\widehat{\boldsymbol{\psi}}) + \nabla P_\lambda(\widehat{\boldsymbol{\psi}}), \boldsymbol{\psi} - \widehat{\boldsymbol{\psi}} \rangle \geq 0 \quad \text{for all feasible } \boldsymbol{\psi},$$

yields

$$\rho\|\boldsymbol{\nu}\|_2^2 - c\sqrt{\frac{\log(dn)}{n}}\|\boldsymbol{\nu}\|_1^2 \leq -\langle \nabla \ell(\boldsymbol{\psi}^*), \boldsymbol{\nu}\rangle + P_\lambda(\boldsymbol{\psi}^*) - P_\lambda(\widehat{\boldsymbol{\psi}}) + \frac{1}{2(\xi-1)}\|\boldsymbol{\nu}\|_2^2.$$

Rearranging and applying Hölder's inequality gives

$$\left\{\rho - \frac{1}{2(\xi-1)}\right\}\|\boldsymbol{\nu}\|_2^2 \leq P_\lambda(\boldsymbol{\psi}^*) - P_\lambda(\widehat{\boldsymbol{\psi}}) + \|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty \|\boldsymbol{\nu}\|_1 + c\sqrt{\frac{\log(dn)}{n}}\|\boldsymbol{\nu}\|_1^2.$$

Under the constraint that $\|\boldsymbol{\nu}\|_1 \leq R$, this results in

$$\left\{\rho - \frac{1}{2(\xi-1)}\right\}\|\boldsymbol{\nu}\|_2^2 \leq P_\lambda(\boldsymbol{\psi}^*) - P_\lambda(\widehat{\boldsymbol{\psi}}) + \left(\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty + Rc\sqrt{\frac{\log(dn)}{n}}\right)\|\boldsymbol{\nu}\|_1.$$

By Lemma E.3, $\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty = O_P\left(\sqrt{\log(d)/n}\right)$. So for suitable choices of R and $\lambda = c'\sqrt{\log(dn)/n}$ with c' sufficiently large, then with high probability

$$\left\{\rho - \frac{1}{2(\xi-1)}\right\}\|\boldsymbol{\nu}\|_2^2 \leq P_\lambda(\boldsymbol{\psi}^*) - P_\lambda(\widehat{\boldsymbol{\psi}}) + \frac{\lambda}{2}\|\boldsymbol{\nu}\|_1. \tag{C.3}$$

By the subadditive property of the SCAD penalty, $P_\lambda(\boldsymbol{\nu}) \leq P_\lambda(\boldsymbol{\psi}^*) + P_\lambda(\widehat{\boldsymbol{\psi}})$. Moreover, Lemma 4 of Loh and Wainwright (2015) shows that for the SCAD penalty, $\lambda|x| \leq p_\lambda(x) + x^2/\{2(\xi-1)\}$ for any $x \in \mathbb{R}$. Together, these results give that

$$\frac{\lambda}{2}\|\boldsymbol{\nu}\|_1 \leq \frac{1}{2}P_\lambda(\boldsymbol{\nu}) + \frac{1}{4(\xi-1)}\|\boldsymbol{\nu}\|_2^2 \leq \frac{P_\lambda(\boldsymbol{\psi}^*)}{2} + \frac{P_\lambda(\widehat{\boldsymbol{\psi}})}{2} + \frac{1}{4(\xi-1)}\|\boldsymbol{\nu}\|_2^2.$$

Combining with (C.3) and rearranging yields

$$0 \leq \left\{\rho - \frac{3}{4(\xi-1)}\right\}\|\boldsymbol{\nu}\|_2^2 \leq \frac{3}{2}P_\lambda(\boldsymbol{\psi}^*) - \frac{1}{2}P_\lambda(\widehat{\boldsymbol{\psi}}), \tag{C.4}$$

where the lower bound of 0 follows by assumption that $\rho > 3/\{4(\xi - 1)\}$. Assume without loss of generality that the unpenalized k baseline parameters of $\boldsymbol{\phi}$ and frailty log-variance parameter σ are non-zero, so $s \geq (k+1)$. Then define S to be the index set of the $k+1$ unpenalized elements, plus the $s-k-1$ largest elements of $\boldsymbol{\beta}$. Then Lemma 5 of Loh and Wainwright (2015) states that

$$3P_\lambda(\boldsymbol{\psi}^*) - P_\lambda(\widehat{\boldsymbol{\psi}}) \leq 3\lambda\|\boldsymbol{\nu}_S\|_1 - \lambda\|\boldsymbol{\nu}_{S^c}\|_1.$$

Substituting this into (C.4) gives

$$\left\{2\rho - \frac{3}{2(\xi-1)}\right\}\|\boldsymbol{\nu}\|_2^2 \leq 3\lambda\|\boldsymbol{\nu}_S\|_1 - \lambda\|\boldsymbol{\nu}_{S^c}\|_1 \leq 3\lambda\|\boldsymbol{\nu}_S\|_1 \leq 3\lambda\sqrt{s}\|\boldsymbol{\nu}\|_2,$$

which yields the final result

$$\|\boldsymbol{\nu}\|_2 \leq \frac{6\lambda\sqrt{s}}{4\rho - 3(\xi-1)^{-1}}.$$

The statistical rate follows by the choice of $\lambda = O\left(\sqrt{\log(dn)/n}\right)$. □

D Observed Data Likelihood Expressions for Gamma-Frailty Illness-Death Model

In this section, we derive expressions for the observed data likelihood, gradient, and Hessian functions for the gamma-frailty illness-death model. As in the main text, for brevity we adopt a semi-Markov specification for h_3 throughout, though only simple modifications are required for corresponding formulae under the Markov specification.

We denote ith subject's observed event times (y_{1i}, y_{2i}) and corresponding observed outcome indicators $(\delta_{1i}, \delta_{2i})$. Next, denote the sum of cumulative cause-specific hazards as

$$A_i = H_{01}(y_{1i})e^{\mathbf{X}_{1i}^\intercal \boldsymbol{\beta}_1} + H_{02}(y_{1i})e^{\mathbf{X}_{2i}^\intercal \boldsymbol{\beta}_2} + H_{03}(y_{2i} - y_{1i})e^{\mathbf{X}_{3i}^\intercal \boldsymbol{\beta}_3}.$$

Now, the negative log-likelihood loss can be succinctly written as

$$\ell(\boldsymbol{\psi}) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\delta_{1i}\log h_1(y_{1i}) + (1-\delta_{1i})\delta_{2i}\log h_2(y_{1i}) + \delta_{1i}\delta_{2i}\log h_3(y_{2i}-y_{1i}) \\ + \delta_{1i}\delta_{2i}\log(1+e^{\sigma}) - (e^{-\sigma}+\delta_{1i}+\delta_{2i})\log(1+e^{\sigma}A_i)\Big\}.$$

For detailed casewise derivation of the observed data likelihood, see Appendix B of Lee et al. (2015).

Finally, as in the main text we reduce repetition by defining unifying notation for the observed outcomes:

$$\widetilde{y}_{gi} = \begin{cases} y_{1i}, & g \in \{1,2\}, \\ y_{2i} - y_{1i}, & g = 3, \end{cases} \qquad \text{and} \qquad \widetilde{\delta}_{gi} = \begin{cases} \delta_{1i}, & g = 1, \\ (1-\delta_{1i})\delta_{2i}, & g = 2, \\ \delta_{1i}\delta_{2i}, & g = 3. \end{cases}$$

D.1 General Case

D.1.1 Gradient of Loss

Considering just the ith subject's contribution, for $g = 1, 2, 3$ and $j = 1, \ldots, k_g$ the gradient expressions are

$$\frac{\partial \ell_i(\boldsymbol{\psi})}{\partial \sigma} = \frac{\delta_{1i}\delta_{2i}e^{\sigma}}{1+e^{\sigma}} + \frac{\log(1+e^{\sigma}A_i)}{e^{\sigma}} - \frac{1+e^{\sigma}(\delta_{1i}+\delta_{2i})}{1+e^{\sigma}A_i}A_i, \tag{D.1}$$

$$\frac{\partial \ell_i(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}_g^\intercal} = \left\{\widetilde{\delta}_{gi} - \frac{1+e^{\sigma}(\delta_{1i}+\delta_{2i})}{1+e^{\sigma}A_i}H_{0g}(\widetilde{y}_{gi})e^{\mathbf{X}_{gi}^\intercal \boldsymbol{\beta}_g}\right\}\mathbf{X}_{gi}, \tag{D.2}$$

$$\frac{\partial \ell_i(\boldsymbol{\psi})}{\partial \phi_{gj}} = \widetilde{\delta}_{gi}\left\{\frac{\partial}{\partial \phi_{gj}}\log h_{0g}(\widetilde{y}_{gi})\right\} - \frac{\{1+e^{\sigma}(\delta_{1i}+\delta_{2i})\}e^{\mathbf{X}_{gi}^\intercal \boldsymbol{\beta}_g}}{1+e^{\sigma}A_i}\left\{\frac{\partial H_{0g}(\widetilde{y}_{gi})}{\partial \phi_{gj}}\right\}. \tag{D.3}$$

D.1.2 Hessian of Loss

Considering just the ith subject's contribution, using general notation over $g \in \{1,2,3\}$, $r \in \{1,2,3\}$, $j = 1,\ldots,k_g$, and $l = 1,\ldots,k_r$, the Hessian is expressions are

$$\frac{\partial^2 \ell_i(\boldsymbol{\psi})}{\partial\sigma\partial\sigma} = \frac{A_i + e^\sigma A_i(2A_i - \delta_{1i} - \delta_{2i})}{(1+e^\sigma A_i)^2} - \frac{\delta_{1i}\delta_{2i}e^\sigma}{(1+e^\sigma)^2} - e^\sigma \log(1+e^\sigma A_i), \tag{D.4}$$

$$\begin{aligned}\frac{\partial^2 \ell_i(\boldsymbol{\psi})}{\partial\boldsymbol{\beta}_g\partial\boldsymbol{\beta}_r^\intercal} &= \frac{\{1+e^\sigma(\delta_{1i}+\delta_{2i})\}e^{\mathbf{X}_{gi}^\intercal\boldsymbol{\beta}_g}}{1+e^\sigma A_i} H_{0g}(\widetilde{y}_{gi}) \\ &\quad \times \left\{ \frac{e^\sigma H_{0r}(\widetilde{y}_{ri})e^{\mathbf{X}_{ri}^\intercal\boldsymbol{\beta}_r}}{1+e^\sigma A_i} - \mathbb{I}(g=r) \right\} \mathbf{X}_{ri}\mathbf{X}_{gi}^\intercal,\end{aligned} \tag{D.5}$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\psi})}{\partial\sigma\partial\boldsymbol{\beta}_r^\intercal} = \frac{e^\sigma H_{0r}(\widetilde{y}_{ri})e^{\mathbf{X}_{ri}^\intercal\boldsymbol{\beta}_r}}{1+e^\sigma A_i} \left\{ \frac{1+e^\sigma(\delta_{1i}+\delta_{2i})}{1+e^\sigma A_i} A_i - (\delta_{1i}+\delta_{2i}) \right\} \mathbf{X}_{ri}, \tag{D.6}$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\psi})}{\partial\sigma\partial\phi_{gj}} = \frac{e^\sigma e^{\mathbf{X}_{gi}^\intercal\boldsymbol{\beta}_g}}{1+e^\sigma A_i} \left\{ \frac{\partial H_{0g}(\widetilde{y}_{gi})}{\partial\phi_{gj}} \right\} \left\{ \frac{1+e^\sigma(\delta_{1i}+\delta_{2i})}{1+e^\sigma A_i} A_i - (\delta_{1i}+\delta_{2i}) \right\}, \tag{D.7}$$

$$\begin{aligned}\frac{\partial^2 \ell_i(\boldsymbol{\psi})}{\partial\phi_{gj}\partial\boldsymbol{\beta}_r^\intercal} &= \frac{(1+e^\sigma(\delta_{1i}+\delta_{2i}))e^{\mathbf{X}_{gi}^\intercal\boldsymbol{\beta}_g}}{1+e^\sigma A_i} \left\{ \frac{\partial H_{0g}(\widetilde{y}_{gi})}{\partial\phi_{gj}} \right\} \\ &\quad \times \left\{ \frac{e^\sigma H_{0r}(\widetilde{y}_{ri})e^{\mathbf{X}_{ri}^\intercal\boldsymbol{\beta}_r}}{1+e^\sigma A_i} - \mathbb{I}(g=r) \right\} \mathbf{X}_{ri},\end{aligned} \tag{D.8}$$

$$\begin{aligned}\frac{\partial^2 \ell_i(\boldsymbol{\psi})}{\partial\phi_{gj}\partial\phi_{rl}} &= \frac{\{1+e^\sigma(\delta_{1i}+\delta_{2i})\}e^{\mathbf{X}_{gi}^\intercal\boldsymbol{\beta}_g}}{1+e^\sigma A_i} \\ &\quad \times \left[\frac{e^\sigma e^{\mathbf{X}_{ri}^\intercal\boldsymbol{\beta}_r}}{1+e^\sigma A_i} \left\{ \frac{\partial H_{0g}(\widetilde{y}_{gi})}{\partial\phi_{gj}} \frac{\partial H_{0r}(\widetilde{y}_{ri})}{\partial\phi_{rl}} \right\} - \left\{ \frac{\partial^2 H_{0g}(\widetilde{y}_{gi})}{\partial\phi_{gj}\partial\phi_{rl}} \right\} \right] \\ &\quad + \widetilde{\delta}_{gi} \left\{ \frac{\partial^2}{\partial\phi_{gj}\partial\phi_{rl}} \log h_{0g}(\widetilde{y}_{gi}) \right\}.\end{aligned} \tag{D.9}$$

D.2 Piecewise Constant Baseline Hazard

Recall that for the gth transition baseline hazard, the piecewise constant specification requires a user-defined set of knots $0 = t_g^{(1)} < \cdots < t_g^{(k_g)} < t_g^{(k_g+1)} = \infty$ defining the intervals over which the hazard is constant. Using general notation over $g \in \{1,2,3\}$, $r \in \{1,2,3\}$, $j = 1,\ldots,k_g$, and $l = 1,\ldots,k_r$, the cause-specific log-baseline hazard and its first two

derivatives are

$$\log h_{0g}(t) = \sum_{j=1}^{k_g} \phi_{gj} \mathbb{I}(t^{(j)} \leq t < t^{(j+1)}),$$

$$\frac{\partial}{\partial \phi_{gj}} \log h_{0g}(t) = \mathbb{I}(t^{(j)} \leq t < t^{(j+1)}), \tag{D.10}$$

$$\frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} \log h_{0g}(t) = 0. \tag{D.11}$$

For the gth transition, define $B_{gj}(t) = (\min(t, t_g^{(j+1)}) - t_g^{(j)}) \mathbb{I}(t \geq t_g^{(j)})$ to represent the amount of time spent in the jth interval. Then the cumulative cause-specific hazard is

$$H_{0g}(t) = \sum_{j=1}^{k_1} e^{\phi_{gj}} B_{gj}(t), \tag{D.12}$$

$$\frac{\partial}{\partial \phi_{gj}} H_{0g}(t) = e^{\phi_{gj}} B_{gj}(t), \tag{D.13}$$

$$\frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} H_{0g}(t) = e^{\phi_{gj}} B_{gj}(t) \mathbb{I}(g = r, j = l). \tag{D.14}$$

D.3 Weibull Baseline Hazard

Using general notation over $g \in \{1, 2, 3\}$, $r \in \{1, 2, 3\}$, $j = 1, \ldots, k_g$, and $l = 1, \ldots, k_r$, the cause-specific log-baseline hazard and its first two derivatives are

$$\log h_{0g}(t) = \phi_{g1} + \phi_{g2} + (e^{\phi_{g1}} - 1) \log t,$$

$$\frac{\partial}{\partial \phi_{gj}} \log h_{0g}(t) = 1 + (e^{\phi_{g1}} \log t) \mathbb{I}(j = 1), \tag{D.15}$$

$$\frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} \log h_{0g}(t) = (e^{\phi_{g1}} \log t) \mathbb{I}(g = r, j = l = 1). \tag{D.16}$$

The cause-specific cumulative hazard and its first two derivatives are then

$$H_{0g}(t) = e^{\phi_{g2}} t^{\exp(\phi_{g1})}, \tag{D.17}$$

$$\frac{\partial}{\partial \phi_{gj}} H_{0g}(t) = e^{\phi_{g2}} t^{\exp(\phi_{g1})} (e^{\phi_{g1}} \log t)^{\mathbb{I}(j=1)}, \tag{D.18}$$

$$\begin{aligned} \frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} H_{0g}(t) = {} & e^{\phi_{g2}} t^{\exp(\phi_{g1})} (e^{\phi_{g1}} \log t)^{\{\mathbb{I}(j=1) + \mathbb{I}(l=1)\}} \mathbb{I}(g = r) \\ & + \{e^{\phi_{g2}} e^{\phi_{g1}} t^{\exp(\phi_{g1})} \log t\} \mathbb{I}(g = r, j = l = 1). \end{aligned} \tag{D.19}$$

E Technical Lemmas

This section presents supporting details underlying the theoretical results. To summarize,

- Under Assumption 1, Lemma E.1 verifies that the piecewise constant baseline hazard specification satisfies Assumption 4.
- Under Assumption 1, Lemma E.2 verifies that the Weibull baseline hazard specification satisfies Assumption 4.
- Under Assumption 1 and Assumption 4, Lemma E.3 confirms a probabilistic bound on the largest gradient element.
- Under Assumption 1 and Assumption 4, Lemma E.4 confirms a probabilistic bound on the largest deviation of a Hessian element from its mean.
- Under Assumption 3 and the above Hessian bound, a Restricted Strong Convexity condition follows with high probability.

Lemma E.1. *Under Assumption 1, the piecewise constant baseline hazard specification satisfies Assumption 4.*

Proof. Assumption 4a follows by inspection of the baseline cumulative hazard function (D.12), and its derivatives (D.13), and (D.14), which are piecewise linear and therefore bounded on a closed interval. Assumption 4c follows trivially, as any second derivatives of the log baseline hazard function (D.11) are the zero function. Finally, note that any first derivative of the log baseline hazard function (D.10) is just an indicator function, so $\widetilde{\Delta}_{gi}\left\{\partial \log h_{0g}(\widetilde{Y}_{gi})/\partial\phi_{gj}\right\}$ is itself a Bernoulli random variable. Therefore, it must have finite variance, and Assumption 4b is established. □

Lemma E.2. *Under Assumption 1, the Weibull baseline hazard specification satisfies Assumption 4.*

Proof. We start with a proof of Assumption 4a. Note that $\{\boldsymbol{\psi} : \|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|_2 \leq R\}$ is a compact subset of $\mathbb{R}^{d+k+1}$, and $[0,t]$ is a closed interval. Then because the baseline cumulative hazard function (D.17) and its derivatives (D.18) and (D.19) are continuous in t and $\boldsymbol{\psi}$, by the Extreme Value Theorem the functions are bounded over the given space.

To prove Assumption 4b, note that for each so-called 'scale' parameter ϕ_{g2}, the corresponding derivative of the log baseline hazard (D.15) is one, so

$$\operatorname{Var}\left\{\widetilde{\Delta}_{gi}\frac{\partial}{\partial\phi_{g2}}\log h_{0g}(\widetilde{Y}_{gi})\right\} = \operatorname{Var}\left(\widetilde{\Delta}_{gi}\right),$$

which is finite, as $\widetilde{\Delta}_{gi}$ is a binary random variable.

However, for each 'shape' parameter ϕ_{g1}, (D.15) is an unbounded function as $t \to 0$. Using the law of total variance, then

$$\begin{aligned}\operatorname{Var}\left(\widetilde{\Delta}_{gi}\log\widetilde{Y}_{gi}\right) =& \mathbb{E}_{\widetilde{\Delta}_{gi}}\left\{\operatorname{Var}\left(\widetilde{\Delta}_{gi}\log\widetilde{Y}_{gi} \mid \widetilde{\Delta}_{gi}\right)\right\} + \operatorname{Var}_{\widetilde{\Delta}_{gi}}\left\{\mathbb{E}\left(\widetilde{\Delta}_{gi}\log\widetilde{Y}_{gi} \mid \widetilde{\Delta}_{gi}\right)\right\} \\ =& \Pr\left(\widetilde{\Delta}_{gi}=1\right)\operatorname{Var}\left(\log\widetilde{Y}_{gi} \mid \widetilde{\Delta}_{gi}=1\right) \\ &+ \Pr\left(\widetilde{\Delta}_{gi}=1\right)\Pr\left(\widetilde{\Delta}_{gi}=0\right)\mathbb{E}\left(\log\widetilde{Y}_{gi} \mid \widetilde{\Delta}_{gi}=1\right)^2.\end{aligned}$$

Thus, to show $\operatorname{Var}\left(\widetilde{\Delta}_{gi}\log\widetilde{Y}_{gi}\right)$ is finite it suffices to show finiteness of $E\left(\log\widetilde{Y}_{gi} \mid \widetilde{\Delta}_{gi}=1\right)$ and $E\left(\log^2\widetilde{Y}_{gi} \mid \widetilde{\Delta}_{gi}=1\right)$.

Because the three transition submodels are analogous, without loss of generality we will focus on showing finite variance in the case of $g=1$. To start, assume no censoring, and no covariates. Then the marginal distribution of Y_{1i} depends on the correlation of T_{1i} and T_{2i} induced by γ_i. Under the assumption that $\phi_{11}=\phi_{21}$, Jiang and Haneuse (2015) show that

$$\Pr(\Delta_{1i}=1) = \left(\frac{e^{\phi_{12}}}{e^{\phi_{12}}+e^{\phi_{22}}}\right),$$

and derive the conditional distribution of Y_{1i} as

$$\begin{aligned}f_{Y_{1i}<\infty}(y_{1i} \mid \Delta_{1i}=1) =& \left(\frac{e^{\phi_{12}}+e^{\phi_{22}}}{e^{\phi_{12}}}\right)\frac{e^{\phi_{12}+\phi_{11}}y_{1i}^{e^{\phi_{11}}-1}}{\{1+e^{\sigma}(e^{\phi_{12}}+e^{\phi_{22}})y_{1i}^{e^{\phi_{11}}}\}^{e^{-\sigma}+1}} \\ =& \frac{\left(e^{\phi_{12}}+e^{\phi_{22}}\right)e^{\phi_{11}}y_{1i}^{e^{\phi_{11}}-1}}{\{1+e^{\sigma}(e^{\phi_{12}}+e^{\phi_{22}})y_{1i}^{e^{\phi_{11}}}\}^{e^{-\sigma}+1}}.\end{aligned}$$

By these formulas, it can be shown that

$$\begin{aligned}\mathbb{E}\left(\log Y_{1i} \mid \Delta_{1i}=1\right) &= \int_0^\infty \log(y_{1i}) f_{Y_{1i}<\infty}(y_{1i} \mid \Delta_{1i}=1)dy_{1i} < \infty, \\ \mathbb{E}\left(\log^2 Y_{1i} \mid \Delta_{1i}=1\right) &= \int_0^\infty \log^2(y_{1i}) f_{Y_{1i}<\infty}(y_{1i} \mid \Delta_{1i}=1)dy_{1i} < \infty,\end{aligned}$$

and thus we conclude that $\operatorname{Var}(\Delta_{1i}\log Y_{1i})$ is finite. As long as the random censoring distribution satisfies $\operatorname{Var}(\log C_i) < \infty$, our conclusion remains true if we incorporate censoring. It also remains true in the presence of covariates, by adding $\mathbf{X}_{i1}^\intercal\boldsymbol{\beta}_1$ and $\mathbf{X}_{i2}^\intercal\boldsymbol{\beta}_2$ to ϕ_{12} and ϕ_{22} respectively. Finally, our conclusion holds if we allow the shape parameters to differ, though the closed form expressions become more complicated.

The proof of Assumption 4c follows directly. When $j=1$ and $l=1$ the form of the second derivative of the log baseline hazard (D.16) decomposes into $w_{jl}^{gr}(\boldsymbol{\psi}) = e^{\phi_{g1}}$ and $z_{jl}^{gr}(t) = \log t$, so $\operatorname{Var}\left\{\widetilde{\Delta}_{gi}z_{jl}^{gr}\left(\widetilde{Y}_{gi}\right)\right\} = \operatorname{Var}\left(\widetilde{\Delta}_{gi}\log\widetilde{Y}_{gi}\right)$, which is finite by the previous result. When either $j \neq 1$ or $l \neq 1$, (D.16) is the zero function and the condition follows trivially.

□

Lemma E.3. *Under Assumption 1 and Assumption 4, then there exist positive constants c_1, c_2 such that with probability $1-\epsilon$ the gradient of the negative log-likelihood loss satisfies*

$$\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty \leq \sqrt{\frac{\log\{4(d+1)/\epsilon\}}{2c_1 n}} + \sqrt{\frac{2kc_2}{n\epsilon}}.$$

Proof. By Assumption 1, the elements of $\mathbf{X}_i$ are bounded on $[-\tau_X, \tau_X]$, and $0 < Y_{1i} \leq Y_{2i} \leq \tau_Y$. Then by Assumption 4a, for fixed parameter $\boldsymbol{\psi}^*$ the gradient functions (D.1) and (D.2) corresponding to σ and $\boldsymbol{\beta}$ are bounded over this domain. Therefore, choose positive constant c_1 such that $c_1 \geq \max\{\|\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\psi}^*)\|_\infty, |\nabla_\sigma \ell_i(\boldsymbol{\psi}^*)|\}$, where $\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\psi}^*)$ is the ith subject's contribution to the gradient component corresponding to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\psi}^*$, and $\nabla_\sigma \ell_i(\boldsymbol{\psi}^*)$ is analogously defined.

Now, using the property that random variables bounded by $[-c_1, c_1]$ are sub-Gaussian with variance proxy c_1^2, we may apply Hoeffding's inequality to each gradient component and take a union bound over all $(d+1)$ elements, yielding

$$\Pr\left[\max\{\|\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\psi}^*)\|_\infty, |\nabla_\sigma \ell(\boldsymbol{\psi}^*)|\} > t\right] \leq 2(d+1)\exp\left(-\frac{nt^2}{2c_1^2}\right). \tag{E.1}$$

However, because the gradient contributions from $\boldsymbol{\phi}$ may instead be heavy tailed, we introduce a moment inequality approach to bound these random variables. The form of each such gradient element as given in (D.3) has two terms: the first term has finite variance by Assumption 4b, while the second term is bounded by Assumption 1 and Assumption 4a. Because bounded random variables have finite variance, then all elements of $\nabla_{\boldsymbol{\phi}}(\boldsymbol{\psi}^*)$ have finite variance.

Choose positive constant $c_2 \geq \max_j\{\mathrm{Var}([\nabla_{\boldsymbol{\phi}} \ell_i(\boldsymbol{\psi}^*)]_j)\}$ to be an upper bound on the variance of the ith subject's contributions to all gradient elements in $\boldsymbol{\phi}$ evaluated at $\boldsymbol{\psi}^*$. Then using Chebyshev's inequality, and taking a union bound over all k elements, yields

$$\Pr(\|\nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\psi}^*)\|_\infty > t) \leq \frac{kc_2}{nt^2}. \tag{E.2}$$

Having addressed each major component of the gradient $\nabla \ell(\boldsymbol{\psi}^*)$, then combining (E.1) and (E.2) using a union bound, then the maximum over all of the gradient elements is bounded by

$$\begin{aligned}\Pr(\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty > t) \leq\ & \Pr\left(\max\{\|\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\psi}^*)\|_\infty, |\nabla_\sigma \ell(\boldsymbol{\psi}^*)|\} > t\right) + \Pr(\|\nabla_{\boldsymbol{\phi}} \ell(\boldsymbol{\psi}^*)\|_\infty > t) \\ &\leq 2(d+1)\exp\left(-\frac{nt^2}{2c_1^2}\right) + \frac{kc_2}{nt^2}.\end{aligned}$$

Inverting this result, we have that with probability $1-\epsilon$,

$$\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty \leq \sqrt{\frac{\log(4(d+1)/\epsilon)}{2c_1 n}} + \sqrt{\frac{2kc_2}{n\epsilon}}.$$

Note that the first term implies a $\sqrt{\log(d)/n}$ rate, while the second implies a $1/\sqrt{n}$ rate, yielding the desired overall result

$$\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty = O_P\left(\sqrt{\frac{\log d}{n}}\right).$$

□

Lemma E.4. *For any scalar $u \in [0,1]$ and any $(d+k+1)$-vector $\boldsymbol{\nu}$ satisfying $\|\boldsymbol{\nu}\|_2 \leq R$, consider the ith subject's Hessian contribution evaluated at $\boldsymbol{\psi}^* + u\boldsymbol{\nu}$. Define the matrix of elementwise deviations from its expectation as*

$$G^{\boldsymbol{\nu}}(u) = \nabla^2 \ell(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) - \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}).$$

Then for a grid of points $u_m = m/n$ for $m = 1, \ldots, n$, there exist positive constants $c_3, c_4, c_5, c_6 < \infty$ such that

$$\Pr\big(\max_{1\leq m\leq n} \|G^{\boldsymbol{\nu}}(u_m)\|_{\max} > t\big) \leq 2n[(d+k+1)^2 - k^2]\exp\left(-\frac{nt^2}{2c_3^2}\right) + 2nk^2\exp\left(-\frac{nt^2}{2c_4^2}\right) + \frac{4k^2c_5^2c_6}{nt^2}.$$

Proof. This result is similar in spirit to Lemma E.3, in controlling the maximum deviation of a collection of random variables from their means. However, now our approach needs to also account for a grid of parameter values $\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}$, where $u_m = m/n$ for $m = 1, \ldots, n$.

Let $\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}\ell_i(\boldsymbol{\psi})$ be the submatrix of the ith subject's Hessian contribution corresponding with the second derivatives of $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\psi}$. Denote

$$G^{\boldsymbol{\nu}}_{\boldsymbol{\beta}\boldsymbol{\beta}}(u) = \nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}\ell(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) - \mathbb{E}[\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\psi}^* + u\boldsymbol{\nu})],$$

and define all other submatrices similarly.

Step 1: Elements corresponding to partial derivatives of $\boldsymbol{\beta}$ and σ

By Assumption 1, the elements of $\mathbf{X}_i$ are bounded on $[-\tau_X, \tau_X]$, and $0 < Y_{1i} \leq Y_{2i} \leq \tau_Y$. Moreover, $\boldsymbol{\psi}^* + u\boldsymbol{\nu}$ lies in an ℓ_2-ball of radius R around $\boldsymbol{\psi}^*$. Then by Assumption 4a, the Hessian functions (D.4), (D.5), (D.6), (D.7), and (D.8) corresponding with partial derivatives of σ and $\boldsymbol{\beta}$ are bounded over this domain. Choose positive upper bound c_3 on these $(d+k+1)^2 - k^2$ elements, such that

$$c_3 \geq \max_{\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|_2\leq R} \big\{\|\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}\ell_i(\boldsymbol{\psi})\|_{\max}, \|\nabla^2_{\boldsymbol{\beta}\sigma}\ell_i(\boldsymbol{\psi})\|_{\max}, \|\nabla^2_{\boldsymbol{\beta}\boldsymbol{\phi}}\ell_i(\boldsymbol{\psi})\|_{\max}, \|\nabla^2_{\boldsymbol{\phi}\sigma}\ell_i(\boldsymbol{\psi})\|_{\max}, |\nabla^2_{\sigma\sigma}\ell_i(\boldsymbol{\psi})|\big\}.$$

This means that each element is a random variable bounded by $[-c_3, c_3]$, so is sub-Gaussian with variance proxy c_3^2. Then using Hoeffding's inequality and taking a union bound over

the Hessian components and over the n points of the u_m grid yields

$$\begin{aligned}\Pr\big(\max_{1\le m\le n}\big\{&\|G^{\boldsymbol{\nu}}_{\boldsymbol{\beta\beta}}(u_m)\|_{\max},\|G^{\boldsymbol{\nu}}_{\boldsymbol{\beta\sigma}}(u_m)\|_{\max},\\ &\|G^{\boldsymbol{\nu}}_{\boldsymbol{\beta\phi}}(u_m)\|_{\max},\|G^{\boldsymbol{\nu}}_{\boldsymbol{\phi\sigma}}(u_m)\|_{\max},|G^{\boldsymbol{\nu}}_{\sigma\sigma}(u_m)|\big\}>t\big)\\ &\le 2n[(d+k+1)^2-k^2]\exp\left(-\frac{nt^2}{2c_3^2}\right).\end{aligned} \tag{E.3}$$

Step 2: Elements corresponding to second derivatives of $\boldsymbol{\phi}$

Importantly, the remaining elements of the Hessian which correspond to the second derivatives of baseline hazard parameters $\boldsymbol{\phi}$ may be unbounded on this domain. However, under Assumption 1, Assumption 4a, and Assumption 4c, then by (D.9) the random Hessian element corresponding to the partial derivatives of ϕ_{gj} and ϕ_{rl} takes the form

$$\frac{\partial^2 l_i(\boldsymbol{\psi})}{\partial\phi_{gj}\partial\phi_{rl}} = B^{gr}_{jl}(\boldsymbol{\psi},\mathbf{X}_i,\mathbf{Y}_i,\boldsymbol{\Delta}_i)+\widetilde{\Delta}_{gi}\left(w^{gr}_{jl}(\boldsymbol{\psi})z^{gr}_{jl}(\widetilde{Y})\right), \tag{E.4}$$

where each B^{gr}_{jl} is a function bounded on the domain $\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|_2\le R$, $0\le Y_{1i}\le Y_{2i}\le\tau_Y$, $\boldsymbol{\Delta}_i\in\{0,1\}^2$, and $\|\mathbf{X}_i\|_\infty\le\tau_X$. So, the goal is to control each of term of (E.4), and then combine the results.

Towards bounding the first term, choose positive constant c_4 that upper bounds all B^{gr}_{jl} over the inputs:

$$c_4\ge\max_{g,r,j,l}\left\{\max_{\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|_2\le R}\left[\max_{\substack{0\le Y_{1i}\le Y_{2i}\le\tau_Y,\\ \boldsymbol{\Delta}_i\in\{0,1\}^2}}\left(\max_{\|\mathbf{X}_i\|_\infty\le\tau_X}B^{gr}_{jl}(\boldsymbol{\psi},\mathbf{X}_i,\mathbf{Y}_i,\boldsymbol{\Delta}_i)\right)\right]\right\}.$$

Therefore, each B^{gr}_{jl} is sub-Gaussian with with variance proxy c_4^2, so applying Hoeffding's inequality and a union bound over $m=1,\dots,n$ yields

$$\begin{aligned}&\Pr\left(\max_{1\le m\le n}\frac{1}{n}\sum_{i=1}^n\left|B^{gr}_{jl}(\boldsymbol{\psi}^*+u_m\boldsymbol{\nu},\mathbf{X}_i,\mathbf{Y}_i,\boldsymbol{\Delta}_i)-\mathbb{E}[B^{gr}_{jl}(\boldsymbol{\psi}^*+u_m\boldsymbol{\nu},\mathbf{X}_i,\mathbf{Y}_i,\boldsymbol{\Delta}_i)]\right|>t\right)\\ &\le 2n\exp\left(-\frac{nt^2}{2c_4^2}\right).\end{aligned} \tag{E.5}$$

To control the second term, note that each w^{gr}_{jl} is continuous, so by the Extreme Value Theorem is bounded on $\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|_2\le R$. Choose positive constant

$$c_5\ge\max_{g,r,j,l}\left\{\max_{\|\boldsymbol{\psi}-\boldsymbol{\psi}^*\|_2\le R}|w^{gr}_{jl}(\boldsymbol{\psi})|\right\}.$$

Next, by Assumption 4c, each $\mathrm{Var}\left(\widetilde{\Delta}_{gi}z^{gr}_{jl}(\widetilde{Y}_i)\right)$ is finite, so choose

$$c_6\ge\max_{g,r,j,l}\left\{\mathrm{Var}\left(\widetilde{\Delta}_{gi}z^{gr}_{jl}(\widetilde{Y}_i)\right)\right\}.$$

Then by bounding $|w_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu})|$ over $m = 1, \ldots, n$ by c_5, and using Chebyshev's inequality on $\widetilde{\Delta}_{gi} z_{jl}^{gr}(\widetilde{Y}_i)$, we have

$$\begin{aligned}
&\Pr\left(\max_{1\leq m\leq n} \frac{1}{n}\sum_{i=1}^{n}\left|\widetilde{\Delta}_{gi} w_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}) z_{jl}^{gr}(\widetilde{Y}_i) - \mathbb{E}\left[\widetilde{\Delta}_{gi} w_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}) z_{jl}^{gr}(\widetilde{Y}_i)\right]\right| > t\right) \\
&\leq \Pr\left(\left\{\max_{1\leq m\leq n} |w_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu})|\right\}\frac{1}{n}\sum_{i=1}^{n}|\widetilde{\Delta}_{gi} z_{jl}^{gr}(\widetilde{Y}) - \mathbb{E}[\widetilde{\Delta}_{gi} z_{jl}^{gr}(\widetilde{Y}_i)]| > t\right) \qquad \text{(E.6)} \\
&\leq \Pr\left(\frac{c_5}{n}\sum_{i=1}^{n}|\widetilde{\Delta}_{gi} z_{jl}^{gr}(\widetilde{Y}_i) - \mathbb{E}[\widetilde{\Delta}_{gi} z_{jl}^{gr}(\widetilde{Y}_i)]| > t\right) \leq \frac{c_5^2 c_6}{nt^2}.
\end{aligned}$$

To bring these pieces together, we denote the Hessian component $G_{\boldsymbol{\phi\phi}}^{\boldsymbol{\nu}}(u_m)$ such that

$$\begin{aligned}
\|G_{\boldsymbol{\phi\phi}}^{\boldsymbol{\nu}}(u_m)\|_{\max} = &\max_{g,r,j,l}\frac{1}{n}\sum_{i=1}^{n}\Big\{|B_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}, \mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\Delta}_i) - \mathbb{E}[B_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}, \mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\Delta}_i)]| \\
&+ \left|\widetilde{\Delta}_{gi}\left(w_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}) z_{jl}^{gr}(\widetilde{Y}_i)\right) - \mathbb{E}\left[\widetilde{\Delta}_{gi}\left(w_{jl}^{gr}(\boldsymbol{\psi}^* + u_m\boldsymbol{\nu}) z_{jl}^{gr}(\widetilde{Y}_i)\right)\right]\right|\Big\}.
\end{aligned}$$

Then combining (E.5) and (E.6), and taking a union bound over all k^2 elements of this Hessian submatrix yields

$$\Pr\left(\max_{1\leq m\leq n}\|G_{\boldsymbol{\phi\phi}}^{\boldsymbol{\nu}}(u_m)\|_{\max} > t\right) \leq 2k^2\left[\frac{4c_5^2 c_6}{nt^2} + 2n\exp\left(-\frac{nt^2}{2c_4^2}\right)\right]. \qquad \text{(E.7)}$$

Step 3: Combine result for overall bound on $\max_{1\leq m\leq n}\|G^{\boldsymbol{\nu}}(u_m)\|_{\max}$.

From the definition of $G^{\boldsymbol{\nu}}$, we have that

$$\begin{aligned}
\|G^{\boldsymbol{\nu}}(u_m)\|_{\max} = \max\big\{&\|G_{\boldsymbol{\phi\phi}}^{\boldsymbol{\nu}}(u_m)\|_{\max}, \|G_{\boldsymbol{\beta\beta}}^{\boldsymbol{\nu}}(u_m)\|_{\max}, \|G_{\boldsymbol{\beta\sigma}}^{\boldsymbol{\nu}}(u_m)\|_{\max}, \\
&\|G_{\boldsymbol{\beta\phi}}^{\boldsymbol{\nu}}(u_m)\|_{\max}, \|G_{\boldsymbol{\phi\sigma}}^{\boldsymbol{\nu}}(u_m)\|_{\max}, |G_{\sigma\sigma}^{\boldsymbol{\nu}}(u_m)|\big\}.
\end{aligned}$$

So combining the results of (E.3) and (E.7) via a union bound yields the desired result.

Equivalently, we may invert the result such that with probability $1-\epsilon$,

$$\begin{aligned}
\max_{1\leq m\leq n}\|G^{\boldsymbol{\nu}}(u_m)\|_{\max} \leq &\sqrt{\frac{2c_3^2\log(6n[(d+k+1)^2 - k^2]/\epsilon)}{n}} \\
&+ \sqrt{\frac{2c_4^2\log(6nk^2/\epsilon)}{n}} + \sqrt{\frac{12k^2c_5^2c_6}{n\epsilon}}.
\end{aligned}$$

□

Lemma E.5. *Under Assumptions 1, 2 and 3, and a model satisfying Assumption 4, then with probability $1-\epsilon$, for some positive constant c the following holds:*

$$\langle\nabla\ell(\boldsymbol{\psi}^* + \boldsymbol{\nu}) - \nabla\ell(\boldsymbol{\psi}^*), \boldsymbol{\nu}\rangle \geq \rho\|\boldsymbol{\nu}\|_2^2 - c\sqrt{\frac{\log(dn)}{n}}\|\boldsymbol{\nu}\|_1^2 \quad \forall\|\boldsymbol{\nu}\|_2 \leq R.$$

Proof. Using the integral definition of the mean-value theorem generalized to vector-valued functions, for some $u \in [0,1]$ we have

$$\langle \nabla \ell(\boldsymbol{\psi}^* + \boldsymbol{\nu}) - \nabla \ell(\boldsymbol{\psi}^*), \boldsymbol{\nu} \rangle = \boldsymbol{\nu}^\intercal \left(\int_0^1 \nabla^2 \ell(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) du \right) \boldsymbol{\nu},$$

where the integral is understood to be elementwise. Recall the definition $\Sigma(\boldsymbol{\psi}) = \mathbb{E}[\nabla^2 \ell(\boldsymbol{\psi})]$, and decompose

$$\begin{aligned} \boldsymbol{\nu}^\intercal \left(\int_0^1 \nabla^2 \ell(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) du \right) \boldsymbol{\nu} =& \boldsymbol{\nu}^\intercal \left(\int_0^1 \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) du \right) \boldsymbol{\nu} \\ &+ \boldsymbol{\nu}^\intercal \left(\int_0^1 \left[\nabla^2 \ell(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) - \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) \right] du \right) \boldsymbol{\nu}. \end{aligned} \tag{E.8}$$

To yield our final RSC condition, we bound each of these two righthand terms.

Step 1: Bound minimum eigenvalue of the population Hessian in a ball

Starting with the first term on the righthand side of (E.8), and recalling that $\lambda_{\min}(\cdot)$ denotes the matrix minimum eigenvalue,

$$\begin{aligned} \boldsymbol{\nu}^\intercal \left(\int_0^1 \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) du \right) \boldsymbol{\nu} &= \int_0^1 \boldsymbol{\nu}^\intercal \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) \boldsymbol{\nu} du \\ &\geq \int_0^1 \lambda_{\min} \left(\Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) \right) \|\boldsymbol{\nu}\|_2^2 du. \end{aligned} \tag{E.9}$$

Now, under Assumption 3, the population Hessian matrix minimum eigenvalue is bounded away from 0 by ρ over the ball $\|\boldsymbol{\nu}\|_2 \leq R$. For any $u \in [0,1]$ then $\boldsymbol{\psi}^* + u\boldsymbol{\nu}$ will be within that ball, yielding the bound

$$\boldsymbol{\nu}^\intercal \left(\int_0^1 \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) du \right) \boldsymbol{\nu} \geq \int_0^1 \rho \|\boldsymbol{\nu}\|_2^2 du = \rho \|\boldsymbol{\nu}\|_2^2 > 0.$$

Step 2: Bound elementwise maximum deviation between sample and population Hessian

To simplify notation, define the matrix-valued deviation function $G^{\boldsymbol{\nu}}(u) = \nabla^2 \ell(\boldsymbol{\psi}^* + u\boldsymbol{\nu}) - \Sigma(\boldsymbol{\psi}^* + u\boldsymbol{\nu})$ as in Lemma E.4. For an arbitrary matrix A, denote A_{jl} its (j,l)th element, and $\|A\|_{\max} = \max_{j,l} |A_{jl}|$. Then repeatedly using the triangle inequality and Hölder's inequality yields

$$\left| \boldsymbol{\nu}^\intercal \left(\int_0^1 G^{\boldsymbol{\nu}}(u) du \right) \boldsymbol{\nu} \right| \leq \|\boldsymbol{\nu}\|_1^2 \int_0^1 \|G^{\boldsymbol{\nu}}(u)\|_{\max} du.$$

Our goal is then to find an upper bound for $\|G(u)\|_{\max}$ over u, as

$$\int_0^1 \|G^{\boldsymbol{\nu}}(u)\|_{\max} du \leq \sup_{u \in [0,1]} \|G(u)\|_{\max}.$$

Lemma E.3 shows an approach to bounding the maximum of a collection of bounded random variables, but now we must bound the maximum both over the elements and over the interval $u \in [0,1]$. For this we use an ϵ-net argument. Define a grid of points $u_m = m/n$ for $m = 1, \ldots, n$, then for any u there exists m such that $|u - u_m| \leq \frac{1}{n}$. Then

$$\begin{aligned} \|G^{\boldsymbol{\nu}}(u)\|_{\max} &\leq \sup_{u\in[0,1]} \max_{j,l} |[G^{\boldsymbol{\nu}}(u)]_{jl}| \\ &\leq \max_{1\leq m\leq n} \max_{j,l} |[G^{\boldsymbol{\nu}}(u_m)]_{jl}| + \sup_{|u-u_m|\leq 1/n} \max_{j,l} |[G^{\boldsymbol{\nu}}(u)]_{jl} - [G^{\boldsymbol{\nu}}(u_m)]_{jl}|. \end{aligned} \tag{E.10}$$

By Lemma E.4, the first term on the RHS of (E.10) is bounded with probability $1-\epsilon$ by

$$\begin{aligned} \max_{1\leq m\leq n} \|G^{\boldsymbol{\nu}}(u_m)\|_{\max} \leq \sqrt{\frac{2c_3^2 \log(6n[(d+k+1)^2 - k^2]/\epsilon)}{n}} \\ + \sqrt{\frac{2c_4^2 \log(6nk^2/\epsilon)}{n}} + \sqrt{\frac{12k^2 c_5^2 c_6}{n\epsilon}}. \end{aligned} \tag{E.11}$$

To bound the final term of (E.10), each of the (j,l)th elements $G^{\boldsymbol{\nu}}_{jl}(u)$ is continuous in u over the closed interval $[0,1]$ and is therefore locally Lipschitz on the interval. Define $c' < \infty$ such that for all j, l, $G^{\boldsymbol{\nu}}_{jl}$ is c'-Lipschitz smooth on the interval $[0,1]$. Then

$$\sup_{|u-u_m|\leq 1/n} \max_{j,l} |G^{\boldsymbol{\nu}}_{jl}(u) - G^{\boldsymbol{\nu}}_{jl}(u_m)| \leq \sup_{|u-u_m|\leq 1/n} c'|u - u_m| \leq \frac{c'}{n}.$$

Then putting this result and (E.11) into (E.10), we have that with probability $1-\epsilon$,

$$\begin{aligned} \sup_{u\in[0,1]} \max_{j,l} |G^{\boldsymbol{\nu}}_{jl}(u)| \leq \sqrt{\frac{2c_3^2 \log(6n[(d+k+1)^2 - k^2]/\epsilon)}{n}} \\ + \sqrt{\frac{2c_4^2 \log(6nk^2/\epsilon)}{n}} + \sqrt{\frac{12k^2 c_5^2 c_6}{n\epsilon}} + \frac{c'}{n}. \end{aligned} \tag{E.12}$$

Step 3: Plug bounds into final expression

By plugging (E.9) and (E.12) into (E.8), we conclude that with probability $1-\epsilon$ there exists some positive constant c such that

$$\langle \nabla \ell(\boldsymbol{\psi}^* + \boldsymbol{\nu}) - \nabla \ell(\boldsymbol{\psi}^*), \boldsymbol{\nu} \rangle \geq \rho \|\boldsymbol{\nu}\|_2^2 - c\sqrt{\frac{\log(dn)}{n}} \|\boldsymbol{\nu}\|_1^2 \quad \forall \|\boldsymbol{\nu}\|_2 \leq R.$$

□

F Additional Simulation Details

In this section, we present additional details on simulation settings, and simulation results. We specifically point out that the non-zero elements of $\boldsymbol{\beta}^*$ shown below are always within the first 15 elements of $\mathbf{X}$, so due to the AR(0.25) serial auto-correlation specification of the covariates $\mathbf{X}$, the correlation between covariates with non-zero effects ranges from 0.25 for neighboring covariates to 0.25^{-14} between X_1 and X_{15}.

Table F.1: Summary of simulation settings run for each sample size $n = 300, 500, 1000$.

	Simulation Settings							
	1	2	3	4	5	6	7	8
True Baseline Parameters								
Moderate Non-Terminal Event Rate $\begin{pmatrix}\boldsymbol{\phi}_1^*\\ \boldsymbol{\phi}_2^*\\ \boldsymbol{\phi}_3^*\end{pmatrix} = \begin{pmatrix}(0.005, 0.015, 0.050, 0.0125)^\intercal\\ (0.010, 0.040, 0.075, 0.0500)^\intercal\\ (0.010, 0.040, 0.075, 0.0750)^\intercal\end{pmatrix}$	X	X	X	X				
Low Non-Terminal Event Rate $\begin{pmatrix}\boldsymbol{\phi}_1^*\\ \boldsymbol{\phi}_2^*\\ \boldsymbol{\phi}_3^*\end{pmatrix} = \begin{pmatrix}(0.035, 0.025, 0.020, 0.025)^\intercal\\ (0.005, 0.010, 0.025, 0.018)^\intercal\\ (0.008, 0.015, 0.024, 0.024)^\intercal\end{pmatrix}$					X	X	X	X
True Regression Parameters								
Shared Support $\begin{pmatrix}\boldsymbol{\beta}_1^*\\ \boldsymbol{\beta}_2^*\\ \boldsymbol{\beta}_3^*\end{pmatrix} = \begin{pmatrix}(0.3, -0.4, 0.5, 0.2, -0.4, 0.3, -0.4, 0.5, 0.2, -0.4, \mathbf{0})^\intercal\\ (0.8, -1.0, 0.6, 0.3, -0.5, 0.8, -1.0, 0.6, 0.3, -0.5, \mathbf{0})^\intercal\\ (0.6, -0.7, 0.7, 0.4, -0.3, 0.6, -0.7, 0.7, 0.4, -0.3, \mathbf{0})^\intercal\end{pmatrix}$	X	X			X	X		
Partially Shared Support $\begin{pmatrix}\boldsymbol{\beta}_1^*\\ \boldsymbol{\beta}_2^*\\ \boldsymbol{\beta}_3^*\end{pmatrix} = \begin{pmatrix}(0.3, -0.4, 0.5, 0.2, -0.4, 0.3, -0.4, 0.5, 0.2, -0.4, \mathbf{0})^\intercal\\ (\mathbf{0}_5, 0.6, -0.7, 0.5, 0.2, -0.4, 0.3, -0.4, 0.5, 0.2, -0.4, \mathbf{0})^\intercal\\ (\mathbf{0}_5, 0.6, -0.7, 0.7, 0.4, -0.3, 0.6, -0.7, 0.7, 0.4, -0.3, \mathbf{0})^\intercal\end{pmatrix}$			X	X			X	X
Dimensionality Regime								
Low Dimension ($d_g = 25$)	X		X		X		X	
High Dimension ($d_g = 350$)		X		X		X		X

F.1 Failures of MLE and Forward Selection Optimization

The MLE and forward selection comparator methods use the `optim` function built into `R`, which flags failure to converge. In Table F.2 we summarize the proportion of MLE and forward selection simulations that failed, by simulation setting. In all subsequent tables, we report results from the subset of successful iterations.

Table F.2: Optimization failure proportions for maximum likelihood and forward selection comparator models, by specification and simulation setting. Maximum likelihood estimates only available for low-dimensional setting.

	$n = 300$		$n = 500$		$n = 1000$	
Weibull	MLE	Forward	MLE	Forward	MLE	Forward
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.00	0.00	0.00	0.00	0.00	0.00
High-Dimension	—	0.67	—	0.09	—	0.00
Partially Non-Overlapping Support						
Low-Dimension	0.00	0.00	0.00	0.00	0.00	0.00
High-Dimension	—	0.70	—	0.10	—	0.00
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.01	0.10	0.00	0.06	0.00	0.03
High-Dimension	—	0.90	—	0.47	—	0.07
Partially Non-Overlapping Support						
Low-Dimension	0.00	0.00	0.00	0.00	0.00	0.00
High-Dimension	—	0.83	—	0.23	—	0.00
	$n = 300$		$n = 500$		$n = 1000$	
Piecewise Constant	MLE	Forward	MLE	Forward	MLE	Forward
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.00	0.00	0.00	0.00	0.00	0.00
High-Dimension	—	0.02	—	0.00	—	0.00
Partially Non-Overlapping Support						
Low-Dimension	0.00	0.00	0.00	0.00	0.00	0.00
High-Dimension	—	0.01	—	0.00	—	0.00
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.01	0.09	0.00	0.06	0.00	0.03
High-Dimension	—	0.55	—	0.24	—	0.06
Partially Non-Overlapping Support						
Low-Dimension	0.00	0.01	0.00	0.00	0.00	0.00
High-Dimension	—	0.26	—	0.02	—	0.00

F.2 Estimation Error Results

F.2.1 Main Results ($n = 500, 1000$), Piecewise Constant Model

Table F.3: Mean ℓ_2 estimation error of $\widehat{\boldsymbol{\beta}}$, piecewise constant baseline hazard specification. Maximum likelihood estimates only available for low-dimensional setting.

$n = 500$	Oracle	MLE	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.74	1.31	1.48	2.06	1.37	1.51	0.98
High-Dimension	0.74	—	2.83	2.76	2.22	2.22	1.22
Partially Non-Overlapping Support							
Low-Dimension	0.71	1.27	1.32	1.89	1.25	1.61	1.16
High-Dimension	0.73	—	2.71	2.49	2.18	2.28	1.43
Low Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.92	1.78	1.88	2.23	1.91	1.47	1.18
High-Dimension	0.89	—	3.87	2.56	2.21	2.34	1.30
Partially Non-Overlapping Support							
Low-Dimension	0.83	1.55	1.54	2.04	1.49	1.71	1.25
High-Dimension	0.82	—	3.44	2.43	2.18	2.32	1.58
$n = 1000$	Oracle	MLE	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.50	0.82	0.76	1.53	0.73	1.39	0.75
High-Dimension	0.49	—	1.22	2.41	1.48	1.83	0.80
Partially Non-Overlapping Support							
Low-Dimension	0.48	0.80	0.71	1.27	0.71	1.35	0.83
High-Dimension	0.48	—	1.15	2.18	1.22	1.84	0.98
Low Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.58	0.97	1.24	2.03	1.12	1.34	0.82
High-Dimension	0.59	—	1.85	2.39	2.09	1.78	0.89
Partially Non-Overlapping Support							
Low-Dimension	0.55	0.92	0.85	1.48	0.85	1.42	0.94
High-Dimension	0.55	—	1.42	2.22	1.76	1.93	1.04

F.2.2 Small Sample Result ($n = 300$), Weibull and Piecewise Constant Models

Table F.4: Mean ℓ_2 estimation error of $\widehat{\boldsymbol{\beta}}$, Weibull and piecewise constant baseline hazard specifications, sample size $n = 300$. Maximum likelihood estimates only available for low-dimensional setting.

Weibull $n = 300$	Oracle	MLE	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate							
Shared Support							
Low-Dimension	1.05	2.24	1.94	2.53	1.90	1.64	1.23
High-Dimension	1.05	—	6.02	2.96	22.04	2.88	22.04
Partially Non-Overlapping Support							
Low-Dimension	1.02	2.12	1.86	2.37	1.79	2.02	1.47
High-Dimension	1.00	—	6.98	2.53	20.47	2.52	20.47
Low Non-Terminal Event Rate							
Shared Support							
Low-Dimension	1.33	5.49	2.11	2.37	2.15	1.78	1.53
High-Dimension	1.33	—	5.67	2.75	20.90	2.71	20.87
Partially Non-Overlapping Support							
Low-Dimension	1.15	3.18	1.99	2.28	2.04	2.04	1.56
High-Dimension	1.14	—	5.96	2.50	20.93	2.52	20.93
Piecewise Constant $n = 300$	Oracle	MLE	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate							
Shared Support							
Low-Dimension	1.02	2.09	1.94	2.52	1.88	1.67	1.25
High-Dimension	1.02	—	6.72	2.97	25.92	2.86	24.97
Partially Non-Overlapping Support							
Low-Dimension	0.99	2.01	1.84	2.37	1.75	2.03	1.47
High-Dimension	0.98	—	7.11	2.53	26.37	2.52	26.05
Low Non-Terminal Event Rate							
Shared Support							
Low-Dimension	1.32	4.65	2.11	2.35	2.14	1.76	1.52
High-Dimension	1.31	—	7.42	2.77	21.46	2.71	19.28
Partially Non-Overlapping Support							
Low-Dimension	1.16	2.88	2.00	2.28	2.01	2.04	1.54
High-Dimension	1.16	—	7.78	2.51	24.34	2.52	23.57

F.3 Sign Inconsistency Results

F.3.1 Main Results ($n = 500, 1000$), Piecewise Constant Model

Table F.5: Mean count of sign-inconsistent $\widehat{\boldsymbol{\beta}}$ estimates, piecewise constant baseline hazard specification.

$n = 500$	Oracle	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.12	11.52	15.21	10.46	11.55	3.56
High-Dimension	0.13	35.43	26.20	30.09	21.51	20.79
Partially Non-Overlapping Support						
Low-Dimension	0.13	10.70	15.47	10.10	15.21	8.19
High-Dimension	0.15	33.50	27.06	35.51	22.86	26.78
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.28	16.02	19.45	16.98	13.05	5.28
High-Dimension	0.26	40.10	26.66	25.24	23.64	14.43
Partially Non-Overlapping Support						
Low-Dimension	0.23	12.42	17.34	12.17	16.76	9.27
High-Dimension	0.20	37.24	26.66	29.50	24.50	23.75
$n = 1000$	Oracle	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.01	4.05	13.58	3.96	8.33	1.66
High-Dimension	0.00	15.59	20.81	18.45	19.05	6.38
Partially Non-Overlapping Support						
Low-Dimension	0.02	4.10	13.40	4.27	12.14	3.94
High-Dimension	0.03	15.24	19.87	16.66	21.40	8.24
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.05	8.81	16.64	8.54	8.76	1.65
High-Dimension	0.06	22.83	24.67	22.04	19.69	6.01
Partially Non-Overlapping Support						
Low-Dimension	0.03	5.14	14.84	5.63	13.53	4.79
High-Dimension	0.02	17.00	22.64	22.43	22.94	8.49

F.3.2 Small Sample Result ($n = 300$), Weibull and Piecewise Constant Models

Table F.6: Mean count of sign-inconsistent $\widehat{\boldsymbol{\beta}}$ estimates, Weibull and piecewise constant baseline hazard specifications, sample size $n = 300$.

Weibull $n = 300$	Oracle	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.48	17.11	21.26	16.21	14.17	6.93
High-Dimension	0.48	50.56	28.73	209.17	26.88	209.17
Partially Non-Overlapping Support						
Low-Dimension	0.50	16.96	23.95	15.80	19.35	12.75
High-Dimension	0.55	50.21	28.72	208.05	28.49	208.05
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.91	19.07	21.01	19.84	16.38	10.24
High-Dimension	0.93	44.83	27.66	217.74	26.89	217.48
Partially Non-Overlapping Support						
Low-Dimension	0.66	18.31	22.05	17.72	19.45	13.61
High-Dimension	0.68	48.14	28.23	231.49	28.63	231.49
Piecewise Constant $n = 300$	Oracle	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.47	17.12	21.15	16.06	13.84	7.14
High-Dimension	0.49	53.07	28.79	177.87	27.06	173.04
Partially Non-Overlapping Support						
Low-Dimension	0.49	16.85	23.90	15.51	19.37	12.63
High-Dimension	0.54	53.45	28.65	184.61	28.72	183.47
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.88	19.05	21.14	19.63	16.75	10.27
High-Dimension	0.90	49.42	28.18	174.62	27.17	158.78
Partially Non-Overlapping Support						
Low-Dimension	0.67	18.17	22.16	17.85	20.02	13.35
High-Dimension	0.67	52.44	28.60	191.24	28.74	186.23

F.4 False Inclusion Results

F.4.1 Main Results ($n = 500, 1000$), Weibull and Piecewise Constant Models

Table F.7: Mean count of false inclusions, Weibull model specification.

$n = 500$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.77	6.18	1.70	11.66	0.79
High-Dimension	23.68	3.52	18.93	18.23	16.13
Partially Non-Overlapping Support					
Low-Dimension	0.82	5.97	1.97	12.24	1.53
High-Dimension	23.59	0.86	23.95	6.18	16.89
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.94	3.97	1.82	11.41	0.89
High-Dimension	23.15	5.51	4.28	11.97	10.61
Partially Non-Overlapping Support					
Low-Dimension	0.82	4.13	1.99	11.06	1.67
High-Dimension	22.49	2.10	9.97	4.64	12.62
$n = 1000$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.54	11.37	1.13	8.45	0.60
High-Dimension	11.79	7.32	10.41	19.43	6.47
Partially Non-Overlapping Support					
Low-Dimension	0.56	12.39	1.05	11.99	0.82
High-Dimension	11.57	6.93	12.16	17.41	5.02
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.55	5.25	2.08	8.28	0.41
High-Dimension	13.06	6.48	3.67	18.06	5.57
Partially Non-Overlapping Support					
Low-Dimension	0.43	10.79	1.67	11.83	1.06
High-Dimension	11.54	5.65	8.91	15.09	4.05

Table F.8: Mean count of false inclusions, piecewise constant model specification.

$n = 500$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.74	6.28	1.61	11.30	0.81
High-Dimension	23.12	3.60	13.24	17.62	18.15
Partially Non-Overlapping Support					
Low-Dimension	0.86	5.87	1.81	12.48	1.45
High-Dimension	22.43	0.89	20.48	6.36	19.38
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.94	4.13	1.72	11.44	0.85
High-Dimension	23.96	6.06	4.68	11.73	9.92
Partially Non-Overlapping Support					
Low-Dimension	0.84	4.46	2.12	11.48	1.59
High-Dimension	23.69	2.23	10.14	5.18	13.05
$n = 1000$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.53	11.44	1.00	8.29	0.61
High-Dimension	11.54	7.21	9.27	18.70	5.53
Partially Non-Overlapping Support					
Low-Dimension	0.56	11.90	0.94	11.56	0.84
High-Dimension	11.28	7.01	8.96	17.65	4.75
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	0.53	5.57	2.05	8.59	0.51
High-Dimension	13.18	6.78	3.67	18.73	4.64
Partially Non-Overlapping Support					
Low-Dimension	0.47	11.38	1.55	12.21	0.93
High-Dimension	11.77	6.47	9.83	16.45	3.89

F.4.2 Small Sample Result ($n = 300$), Weibull and Piecewise Constant Models

Table F.9: Mean count of false inclusions, Weibull and piecewise constant baseline hazard specifications, sample size $n = 300$.

Weibull $n = 300$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	1.22	2.76	1.94	12.70	1.27
High-Dimension	32.84	1.14	194.58	2.67	194.58
Partially Non-Overlapping Support					
Low-Dimension	1.08	1.20	2.29	6.69	1.75
High-Dimension	32.56	1.03	193.72	0.98	193.72
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	1.14	3.38	1.87	10.47	1.32
High-Dimension	25.76	4.15	200.79	5.56	200.57
Partially Non-Overlapping Support					
Low-Dimension	1.18	2.08	2.34	6.15	2.14
High-Dimension	29.12	1.71	216.09	1.08	216.09
Piecewise Constant $n = 300$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	1.20	2.89	2.20	12.13	1.21
High-Dimension	35.64	1.09	162.64	3.42	158.58
Partially Non-Overlapping Support					
Low-Dimension	1.12	1.31	2.08	6.81	1.73
High-Dimension	35.82	1.17	169.88	1.10	169.27
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	1.18	3.66	1.76	10.72	1.27
High-Dimension	30.79	4.17	156.88	5.90	142.89
Partially Non-Overlapping Support					
Low-Dimension	1.22	2.05	2.45	6.42	2.17
High-Dimension	33.62	1.56	174.97	1.04	170.26

F.5 False Exclusion Results

F.5.1 Main Results ($n = 500, 1000$), Weibull and Piecewise Constant Models

Table F.10: Mean count of false exclusions, Weibull model specification.

$n = 500$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	10.74	8.85	8.58	0.24	2.66
High-Dimension	12.12	22.87	16.29	3.66	2.76
Partially Non-Overlapping Support					
Low-Dimension	9.90	9.40	8.21	2.80	6.37
High-Dimension	10.96	26.35	14.57	17.18	7.46
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	15.29	15.34	15.32	1.35	4.56
High-Dimension	16.33	20.57	20.65	11.43	3.81
Partially Non-Overlapping Support					
Low-Dimension	11.88	12.97	10.43	5.24	7.80
High-Dimension	13.71	23.98	19.54	19.89	10.26
$n = 1000$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	3.49	2.15	2.86	0.04	0.91
High-Dimension	4.04	13.60	9.51	0.39	0.74
Partially Non-Overlapping Support					
Low-Dimension	3.56	1.42	3.22	0.59	3.18
High-Dimension	4.01	12.93	7.14	3.87	3.42
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	9.16	11.33	6.94	0.15	1.45
High-Dimension	10.67	17.81	18.49	1.02	1.37
Partially Non-Overlapping Support					
Low-Dimension	4.86	3.75	4.08	1.43	3.91
High-Dimension	5.42	16.32	13.27	6.89	4.66

Table F.11: Mean count of false exclusions, piecewise constant model specification.

$n = 500$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	10.77	8.89	8.85	0.25	2.75
High-Dimension	12.30	22.60	16.85	3.89	2.64
Partially Non-Overlapping Support					
Low-Dimension	9.84	9.59	8.29	2.72	6.74
High-Dimension	11.07	26.18	15.02	16.50	7.40
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	15.08	15.30	15.25	1.61	4.42
High-Dimension	16.11	20.60	20.56	11.91	4.52
Partially Non-Overlapping Support					
Low-Dimension	11.57	12.87	10.05	5.24	7.69
High-Dimension	13.54	24.43	19.36	19.31	10.70
$n = 1000$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	3.52	2.14	2.95	0.04	1.05
High-Dimension	4.05	13.60	9.18	0.36	0.85
Partially Non-Overlapping Support					
Low-Dimension	3.54	1.50	3.33	0.57	3.10
High-Dimension	3.96	12.86	7.71	3.75	3.49
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	8.28	11.05	6.49	0.17	1.14
High-Dimension	9.65	17.90	18.37	0.96	1.36
Partially Non-Overlapping Support					
Low-Dimension	4.67	3.45	4.08	1.32	3.86
High-Dimension	5.23	16.18	12.61	6.49	4.61

F.5.2 Small Sample Result ($n = 300$), Weibull and Piecewise Constant Models

Table F.12: Mean count of false exclusions, Weibull and piecewise constant baseline hazard specifications, sample size $n = 300$.

Weibull $n = 300$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	15.88	18.48	14.25	1.47	5.66
High-Dimension	17.69	27.59	14.30	24.21	14.30
Partially Non-Overlapping Support					
Low-Dimension	15.87	22.75	13.50	12.65	11.00
High-Dimension	17.64	27.69	14.06	27.51	14.06
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	17.90	17.59	17.92	5.89	8.90
High-Dimension	19.00	23.51	16.59	21.32	16.55
Partially Non-Overlapping Support					
Low-Dimension	17.12	19.96	15.34	13.28	11.46
High-Dimension	18.98	26.51	15.06	27.55	15.06
Piecewise Constant $n = 300$	Forward	Lasso	SCAD	Lasso + Fusion	SCAD + Fusion
Moderate Non-Terminal Event Rate					
Shared Support					
Low-Dimension	15.91	18.23	13.83	1.71	5.92
High-Dimension	17.39	27.70	14.95	23.64	14.19
Partially Non-Overlapping Support					
Low-Dimension	15.72	22.58	13.42	12.55	10.90
High-Dimension	17.59	27.49	14.48	27.61	13.95
Low Non-Terminal Event Rate					
Shared Support					
Low-Dimension	17.83	17.43	17.82	6.02	8.99
High-Dimension	18.56	24.00	17.45	21.27	15.64
Partially Non-Overlapping Support					
Low-Dimension	16.94	20.11	15.38	13.57	11.18
High-Dimension	18.78	27.04	16.02	27.70	15.74

G Additional Algorithmic Details

G.1 Tuning the Nesterov smoothing parameter μ

The Nesterov smoothing parameter μ defined in (8) is important for optimization under the proposed structured fusion penalty. If μ is too large then the approximation may be too loose to induce fusion, while if μ is too small the approximation will be insufficiently smooth, and the optimization algorithm will exhibit poor performance.

In our proposed optimization algorithm, we therefore avoid pre-specifying a single μ value, and instead adopt a simple algorithmic approach following Hahn et al. (2020) called 'progressive smoothing'. For each fixed set of regularization parameters (λ_1, λ_2), we first iterate proximal gradient descent to convergence with μ large (e.g., 10^{-2} in our applications), and then decrease μ and further iterate to convergence, and so on until μ is sufficiently small. (In our applications, we decrease μ over the sequence $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$). This approach allows for tight approximations of the fusion penalty to be achieved, while remaining computationally efficient because optimizing for each new μ typically requires only a handful of proximal gradient descent iterations from the preceding μ. This progressive smoothing procedure is fully implemented in our `SemiCompRisksPen` package.

H Spline-Based Baseline Hazard Specifications

In this section, we summarize several spline-based specifications for the baseline hazards in the illness-death model, to which the proposed estimation framework can be readily extended.

H.1 Polynomial B-Spline on Log-Hazard Scale

One choice is to specify each log baseline hazard as a polynomial B-spline function of degree b_g. This approach specifies

$$\log h_{0g}(t) = \sum_{j=1}^{k_g} \phi_{gj} B_{gj}(t)$$

where $k_g \geq b_g + 1$ is the number of desired baseline parameters for the gth hazard, and $B_{gj}(\cdot)$ is the jth B-spline basis function (de Boor, 2001). These basis functions are defined according to the placement of $k_g + b_g - 1$ knots, such that linear combinations of these basis functions can express a range of shapes constrained at those knots. Usually $b_g = 3$ such that the resulting spline function must be continuously twice-differentiable. B-splines are only well-defined over the interval spanned by the knots, so the boundary knots are typically fixed at 0 and the maximum observed time in the gth hazard, with inner knots typically placed at appropriate deciles. While extremely flexible, this specification requires

numerical integration to compute the cumulative hazard, making it slower in practice than other specifications.

Using general notation over $g \in \{1,2,3\}$, $r \in \{1,2,3\}$, $j = 1,\ldots,k_g$, and $l = 1,\ldots,k_r$, the first two derivatives of the cause-specific log-baseline are

$$\begin{aligned}\frac{\partial}{\partial \phi_{gj}} \log h_{0g}(t) =& B_{gj}(t) \\ \frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} \log h_{0g}(t) =& 0\end{aligned}$$

The cause-specific cumulative hazard and its first two derivatives do not have closed form, but can be written as

$$\begin{aligned}H_{0g}(t) &= \int_0^t \exp\left(\sum_{j=1}^{k_g} \phi_{gj} B_{gj}(s)\right) ds \\ \frac{\partial}{\partial \phi_{gj}} H_{0g}(t) &= \int_0^t B_{gj}(s) \exp\left(\sum_{j'=1}^{k_g} \phi_{gj'} B_{gj'}(s)\right) ds \\ \frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} H_{0g}(t) &= \int_0^t B_{gj}(s) B_{gl}(s) \exp\left(\sum_{j'=1}^{k_g} \phi_{gj'} B_{gj'}(s)\right) \mathbb{I}(g=r) ds\end{aligned}$$

H.2 Restricted Cubic Spline on Log-Cumulative Hazard Scale

A final spline-based approach follows Royston and Parmar (2002) in specifying the log-cumulative baseline hazard as a natural (or 'restricted') cubic spline function. Whereas B-splines are only defined over range spanned by the chosen knots, natural cubic splines extend linearly beyond the boundary knots. Set $z = \log(t)$, and to generate k_g total parameters, consider a set of k_g knots $0 \leq z_g^{(1)} < \cdots < z_g^{(k_g)}$. Define $v_{g1}(z) = 1$, $v_{g2}(z) = z$, and for $j = 3,\ldots,k_g$ define natural cubic spline basis functions as

$$v_{gj}(z) = (z - z_g^{(j)})_+^3 - \zeta_j (z - z_g^{(1)})_+^3 - (1 - \zeta_j)(z - z_g^{(k_g)})_+^3,$$

where $\zeta_{gj} = (z_g^{(k_g)} - z_g^{(j)})/(z_g^{(k_g)} - z_g^{(1)})$ and $(z)_+ = \max(0, z)$. Then this model specifies the baseline hazard as

$$\log H_{0g}(t) = \sum_{j=1}^{k_g} \phi_{gj} v_{gj}(z).$$

By specifying the spline model of the log-cumulative hazard, evaluation does not require numerical integration. However, while the log-cumulative hazard is constrained to be monotonically increasing, there is no such constraint inherently on the natural cubic spline. While in principle this might require putting formal constraints on $\boldsymbol{\phi}_g$ during optimization, in practice standard unconstrained methods suffice as long as the starting point is feasible and there

is a modest amount of data (Herndon and Harrell, 1990; Royston and Parmar, 2002). The Weibull model is a special case of the Royston-Parmar model, fixing no internal knots (e.g., setting $k_g = 2$).

The cumulative cause-specific hazard and its first two derivatives are

$$\begin{aligned}
H_{0g}(t) =& \exp\left(\sum_{j=1}^{k_g} \phi_{gj} v_{gj}(z)\right) \\
\frac{\partial}{\partial \phi_{gj}} H_{0g}(t) =& v_{gj}(z) \exp\left(\sum_{j=1}^{k_g} \phi_{gj} v_{gj}(z)\right) \\
\frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} H_{0g}(t) =& v_{gj}(z) v_{gl}(z) \exp\left(\sum_{j=1}^{k_g} \phi_{gj} v_{gj}(z)\right) \mathbb{I}(g = r)
\end{aligned}$$

Now, define $v'_{g1}(z) = 0$, and $v'_{g2}(z) = 1$, and for $j = 3, \ldots, k_g$,

$$v'_{gj}(z) = 3(z - z_g^{(j)})_+^2 - 3\zeta_j (z - z_g^{(1)})_+^2 - 3(1 - \zeta_j)(z - z_g^{(k_g)})_+^2.$$

Using general notation over $g \in \{1, 2, 3\}$, $r \in \{1, 2, 3\}$, $j = 1, \ldots, k_g$, and $l = 1, \ldots, k_r$, the cause-specific log-baseline hazard and its first two derivatives are

$$\begin{aligned}
\log h_{0g}(t) =& \log\left(\sum_{j=1}^{k_g} \phi_{gj} v'_{gj}(z)\right) + \sum_{j=1}^{k_g} \phi_{gj} v_{gj}(z) - \log t \\
\frac{\partial}{\partial \phi_{gj}} \log h_{0g}(t) =& \frac{v'_{gj}(z)}{\sum_{j'=1}^{k_g} \phi_{gj'} v'_{gj'}(z)} + v_{gj}(z) \\
\frac{\partial^2}{\partial \phi_{gj} \partial \phi_{rl}} \log h_{0g}(t) =& -\frac{v'_{gj}(z) v'_{gl}(z)}{\left(\sum_{j'=1}^{k_g} \phi_{gj'} v'_{gj'}(z)\right)^2} \mathbb{I}(g = r)
\end{aligned}$$

References

de Boor, C. (2001). *A Practical Guide to Splines: Revised Edition*, volume 27 of *Applied Mathematical Sciences*. Springer, New York.

Hahn, G., Lutz, S. M., Laha, N., and Lange, C. (2020). A framework to efficiently smooth L1 penalties for linear regression. *bioRxiv:2020.09.17.301788* .

Herndon, J. E. and Harrell, F. E. (1990). The restricted cubic spline hazard model. *Communications in Statistics - Theory and Methods* **19,** 639–663.

Jiang, F. and Haneuse, S. (2015). Simulation of semicompeting risk survival data and estimation based on multistate frailty model. Harvard University Biostatistics Working Paper Series 188, Harvard University.

Lee, C., Lee, S. J., and Haneuse, S. (2020). Time-to-event analysis when the event is defined on a finite time interval. *Statistical Methods in Medical Research* **29,** 1573–1591.

Lee, K. H., Haneuse, S., Schrag, D., and Dominici, F. (2015). Bayesian semiparametric analysis of semicompeting risks data: investigating hospital readmission after a pancreatic cancer diagnosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64,** 253–273.

Loh, P. L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16,** 559–616.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* **26,** 2389–2430.

Putter, H. and Van Houwelingen, H. C. (2015). Frailties in multi-state models: Are they identifiable? Do we need them? *Statistical Methods in Medical Research* **24,** 675–692.

Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* **21,** 2175–2197.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38,** 894–942.