Supplementary Material for: "Penalized likelihood estimation of the proportional hazards model for survival data with interval censoring"

Jun Ma *[1], Dominique-Laurent Couturier[2,3], Stephane Heritier[4], and Ian Marschner[5]

[1]Department of Mathematics Statistics, Macquarie University, Australia

[2]Cancer Research UK - Cambridge Institute, University of Cambridge, UK

[3]MRC Biostatistics Unit, University of Cambridge, UK

[4]School of Public Health and Preventive Medicine, Monash University, Australia

[5]NHMRC Clinical Trials Centre, The University of Sydney, Australia

S1 Components of score and Hessian matrix

Let x_{ij} be element j of vector $\mathbf{x}_i$. The first derivatives of Φ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are, for $j = 1, \ldots, p$ and $u = 1, \ldots, m$,

$$\begin{aligned}
\frac{\partial \Phi(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_j} =& \sum_{i=1}^{n} x_{ij} \left(\delta_i - \delta_i H_i(t_i) - \delta_i^R H_i(t_i) + \delta_i^L \frac{S_i(t_i) H_i(t_i)}{1 - S_i(t_i)} \right. \\
& \left. - \delta_i^I \frac{S_i(t_i^L) H_i(t_i^L) - S_i(t_i^R) H_i(t_i^R)}{S_i(t_i^L) - S_i(t_i^R)} \right), \\
\frac{\partial \Phi(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_u} =& \sum_{i=1}^{n} \left(\delta_i \frac{\psi_u(t_i)}{h_0(t_i)} - \delta_i \Psi_u(t_i) e^{\mathbf{x}_i \boldsymbol{\beta}} - \delta_i^R \Psi_u(t_i) e^{\mathbf{x}_i \boldsymbol{\beta}} + \delta_i^L \frac{S_i(t_i) \Psi_u(t_i)}{1 - S_i(t_i)} e^{\mathbf{x}_i \boldsymbol{\beta}} \right. \\
& \left. - \delta_i^I \frac{S_i(t_i^L) \Psi_u(t_i^L) - S_i(t_i^R) \Psi_u(t_i^R)}{S_i(t_i^L) - S_i(t_i^R)} e^{\mathbf{x}_i \boldsymbol{\beta}} \right) - \lambda \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_u},
\end{aligned}$$

*To whom correspondence should be addressed: jun.ma@mq.edu.au

where $\Psi_u(t) = \int_0^t \psi_u(\xi)d\xi$, the cumulative of basis function $\psi_u(t)$. Elements of the Hessian matrix are:

$$\begin{aligned}
\frac{\partial^2\Phi(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta_j\partial\beta_t} &= -\sum_{i=1}^{n} x_{ij}x_{it}\left(\delta_i H(t_i) + \delta_i^R H(t_i) + \delta_i^L \frac{S_i(t_i)H_i(t_i)(H_i(t_i)+S_i(t_i)-1)}{(1-S_i(t_i))^2}\right.\\
&\quad \left. + \delta_i^I \frac{S_i(t_i^L)S_i(t_i^R)(-H_i(t_i^L)+H_i(t_i^R))^2}{(S_i(t_i^L)-S_i(t_i^R))^2} + \delta_i^I \frac{-S_i(t_i^R)H_i(t_i^R)+S_i(t_i^L)H_i(t_i^L)}{S_i(t_i^L)-S_i(t_i^R)}\right)\\
\frac{\partial^2\Phi(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta_j\partial\theta_u} &= -\sum_{i=1}^{n} x_{ij}e^{\mathbf{x}_i\boldsymbol{\beta}}\left(\delta_i\Psi_u(t_i) + \delta_i^R\Psi_u(t_i) + \delta_i^L\frac{S_i(t_i)}{1-S_i(t_i)}\left[\frac{H_i(t_i)}{1-S_i(t_i)}-1\right]\right.\\
&\quad + \delta_i^I \frac{S_i(t_i^L)S_i(t_i^R)(-H_i(t_i^L)+H_i(t_i^R))(-\Psi_u(t_i^L)+\Psi_u(t_i^R))}{(S_i(t_i^L)-S_i(t_i^R))^2}\\
&\quad \left. + \delta_i^I \frac{S_i(t_i^L)\Psi_u(t_i^L) - S_i(t_i^R)\Psi_u(t_i^R)}{S_i(t_i^L)-S_i(t_i^R)}\right)\\
\frac{\partial^2\Phi(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\theta_u\partial\theta_v} &= -\sum_{i=1}^{n}\left(\delta_i\frac{\psi_u(t_i)\psi_v(t_i)}{h_0^2(t_i)} + e^{2\mathbf{x}_i\boldsymbol{\beta}}\left[\delta_i^L\frac{S_i(t_i)}{(1-S_i(t_i))^2}\Psi_u(t_i)\Psi_v(t_i)\right.\right.\\
&\quad \left.\left. + \delta_i^I\frac{S_i(t_i^L)S_i(t_i^R)}{(S_i(t_i^L)-S_i(t_i^R))^2}(\Psi_u(t_i^R)-\Psi_u(t_i^L))(\Psi_v(t_i^R)-\Psi_v(t_i^L))\right]\right)
\end{aligned}$$

S2 Proof of asymptotic results

Let $C^r[a,b]$ be the set for functions that have r continuous derivatives over interval $[a,b]$; see the main paper for the definition of a and b. Let the space for $\boldsymbol{\beta}$ be given by $B = \{\boldsymbol{\beta} : |\beta_j| \le C_1 < \infty, \forall j\}$, a compact subset of R^p, and the space for $h_0(t)$ be $A = \{h_0(t) : h_0 \in C^r[a,b], 0 \le h_0(t) \le C_2 < \infty, \forall t \in [a,b]\}$. Then the parameter space for $\boldsymbol{\tau} = (\boldsymbol{\beta}, h_0(t))$ is $\Gamma = \{\boldsymbol{\tau} : \boldsymbol{\beta} \in B, h_0 \in A\} = B * A$. Recall $\widehat{h}_0(t) = \sum_{u=1}^m \widehat{\theta}_u\psi_u(t)$. In order to avoid confusion we define $\widetilde{h}_0(t) = \sum_{u=1}^m \theta_u\psi_u(t)$, which is the approximation to $h_0(t)$. Assume θ_u are bounded and non-negative and $\psi_u(t)$ are bounded for $t \in [a,b]$; see Assumption A3 below. The space for $\widetilde{h}_0(t)$ is denoted by $A_n = \{\widetilde{h}_0(t) : 0 \le \widetilde{h}_0(t) \le C_3 < \infty, \forall t \in [a,b]\}$. The parameter space for $\boldsymbol{\tau}_n = (\boldsymbol{\beta}, \widetilde{h}_0(t))$ is $\Gamma_n = \{\boldsymbol{\tau}_n : \boldsymbol{\beta} \in B, \widetilde{h}_0 \in A_n\} = B * A_n$. The MPL estimator of $\boldsymbol{\tau}_n$ is denoted by $\widehat{\boldsymbol{\tau}}_n = (\widehat{\boldsymbol{\beta}}, \widehat{h}_0(t))$.

Let $\boldsymbol{W}_i = (\delta_i^L, \delta_i^R, \delta_i^I, \delta_i, T_i^L, T_i^R, \mathbf{x}_i)^T$ for $i = 1, \ldots, n$. They are random vectors and are assumed i.i.d. The density function for $\boldsymbol{W}_i$ is

$$f(\boldsymbol{w}_i;\boldsymbol{\tau}) = (h_i(t_i)S_i(t_i))^{\delta_i}(1-S_i(t_i^R))^{\delta_i^L}S_i(t_i^L)^{\delta_i^R}(S_i(t_i^L)-S_i(t_i^R))^{\delta_i^I}\gamma(\mathbf{x}_i),$$

where γ denotes the density function of $\mathbf{x}_i$ which is assumed independent of $\boldsymbol{\tau}$. Let $F(\boldsymbol{w}_i;\boldsymbol{\tau})$ be the cumulative distribution function of $\boldsymbol{W}_i$. For $\boldsymbol{\tau} \in \Gamma$, define $Pl(\boldsymbol{\tau}) = \int \log f(\boldsymbol{w}_i;\boldsymbol{\tau})dF(\boldsymbol{w}_i;\boldsymbol{\tau}_0) = E_0(\log f(\boldsymbol{W}_i;\boldsymbol{\tau}))$ and $P_nl(\boldsymbol{\tau}) = \frac{1}{n}\sum_{i=1}^n \log f(\boldsymbol{W}_i;\boldsymbol{\tau})$. For $\boldsymbol{\tau}_n \in \Gamma_n$, $Pl(\boldsymbol{\tau}_n)$ and $P_nl(\boldsymbol{\tau}_n)$ are similarly defined. Here the expectation E_0 is taken with the "true" $\boldsymbol{\tau}$: $\boldsymbol{\tau}_0 = (\boldsymbol{\beta}_0, h_{00}(t))$, which in fact maximizes $Pl(\boldsymbol{\tau})$.

Assumption A4 below assumes that for any $\boldsymbol{\tau} \in B * A$, there exist $\boldsymbol{\tau}_n \in B * A_n$ such that $\rho(\boldsymbol{\tau}_n, \boldsymbol{\tau}) \to 0$ when $n \to \infty$. Here, the definition of $\rho(\cdot)$ is, if $\boldsymbol{\tau}_1 = (\boldsymbol{\beta}_1, h_{01}(t))$ and $\boldsymbol{\tau}_2 = (\boldsymbol{\beta}_2, h_{02}(t))$ then

$$\rho(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = \left\{ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \sup_{t \in [a,b]} |h_{01}(t) - h_{02}(t)|^2 \right\}^{1/2}. \tag{S1}$$

This assumption can be guaranteed under certain regularity conditions, such as those in Proposition 2.8 in DeBoor and Daniel (1974). Recall the scaled smoothing parameter $\mu_n = \lambda/n$.

S2.1 Sketch proof of Theorem 1

Result of Theorem 1 require regularity conditions stated below.

A1. Matrix $\mathbf{X}$ is bounded and $E(\mathbf{X}\mathbf{X}^T)$ is non-singular.

A2. The penalty function J is bounded over Γ and Γ_n.

A3. For function $\widetilde{h}_0(t)$, there is a constant C_6, independent of n, upper bounds all $\theta_u \geq 0$. Moreover, assume the basis functions $\psi_u(t)$ are bounded for $t \in [a, b]$ and $u = 1, \ldots, m$.

A4. The knots and basis functions are selected in a way such that for any $h(t) \in A$ there exists a $\widetilde{h}_0(t) \in A_n$ such that $\max_t |\widetilde{h}_0(t) - h(t)| \to 0$ as $n \to \infty$; see Proposition 2.8 in DeBoor and Daniel (1974) for conditions for this assumption.

Our proof below closely follows the proofs in Xue, Lam, and Li (2004), Zhang, Hua, and Huang (2010) and Huang (1996). Recall $\boldsymbol{\tau} = (\boldsymbol{\beta}, h_0(t)) \in B * A$ and $\boldsymbol{\tau}_n = (\boldsymbol{\beta}, \widetilde{h}_0(t)) \in B * A_n \subset B * A$. For $\boldsymbol{\tau}_1 = (\boldsymbol{\beta}_1, h_{01}(t))$ and $\boldsymbol{\tau}_2 = (\boldsymbol{\beta}_2, h_{02}(t))$ (both in $\in B * A$), define the distance measure

$$\rho(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = \{\|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|^2\}^{1/2} = \left\{ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \sup_{t \in [a,b]} |h_{01}(t) - h_{02}(t)|^2 \right\}^{1/2}.$$

The proofs below require the concept of *covering number* of a space; its detailed definition can be found in, for example, Pollard (1984). Briefly, this is the number of spherical balls of a given size required to cover a given space. For a space A with measure $\kappa(A)$, we denote the covering number associated with spheral radius ε by $N(\varepsilon, A, \kappa(A))$.

Results of Theorem 1 can be demonstrated if we are able to show that $\rho(\boldsymbol{\tau}_0, \widehat{\boldsymbol{\tau}}_n) \to 0$ (a.s.), where $\boldsymbol{\tau}_0 = (\boldsymbol{\beta}_0, h_{00}(t))$. Since the scaled smoothing parameter $\mu_n \to 0$ when $n \to \infty$ and the penalty function is bounded, we can concentrated on the log-likelihood function only. The required result can be obtained through the following results.

(1) Let $q(\boldsymbol{w};\boldsymbol{\tau})$ denote the Fréchet derivative of the density functional $f(\boldsymbol{w};\boldsymbol{\tau})$ with respect to $\boldsymbol{\tau}$. Let $\boldsymbol{\xi}$ be a point in between $\widehat{\boldsymbol{\tau}}_n$ and $\boldsymbol{\tau}_0$. Since $\boldsymbol{\xi}$ is not the maximum, the functional $q(\boldsymbol{w};\boldsymbol{\xi})$ is non-zero. Also, both $q(\boldsymbol{w};\boldsymbol{\xi})$ and $f(\boldsymbol{w};\boldsymbol{\xi})$ are bounded. Recall that $Pl(\boldsymbol{\tau})$ has been defined in Section 4. We have

$$\begin{aligned}
&|Pl(\widehat{\boldsymbol{\tau}}_n) - Pl(\boldsymbol{\tau}_0)| = E_0(\log f(\boldsymbol{W}_i;\boldsymbol{\tau}_0) - \log f(\boldsymbol{W}_i;\widehat{\boldsymbol{\tau}}_n)) \\
&\geq \|f^{\frac{1}{2}}(\boldsymbol{W}_i;\boldsymbol{\tau}_0) - f^{\frac{1}{2}}(\boldsymbol{W}_i;\widehat{\boldsymbol{\tau}}_n)\|_2^2 = \left\| \frac{q(\boldsymbol{W}_i;\boldsymbol{\xi})}{2f^{\frac{1}{2}}(\boldsymbol{W}_i;\boldsymbol{\xi})}(\boldsymbol{\tau}_0 - \widehat{\boldsymbol{\tau}}_n) \right\|_2^2 \\
&\geq C_4\|\boldsymbol{\tau}_0 - \widehat{\boldsymbol{\tau}}_n\|_2^2, \qquad \text{(S2)}
\end{aligned}$$

where the first inequality is established since the Kullback-Leibler distance is not less than the Hellinger distance (Wong and Shen, 1995), the second equality comes from the mean value theorem and C_4 is the lower bound of $|q(\boldsymbol{W}_i;\boldsymbol{\xi})/2f^{\frac{1}{2}}(\boldsymbol{W}_i;\boldsymbol{\xi})|$.

(2) It then suffices to show $Pl(\widehat{\boldsymbol{\tau}}_n) - Pl(\boldsymbol{\tau}_0) \to 0$ (a.s.). However, since

$$|Pl(\widehat{\boldsymbol{\tau}}_n) - Pl(\boldsymbol{\tau}_0)| \leq |Pl(\widehat{\boldsymbol{\tau}}_n) - P_n l(\widehat{\boldsymbol{\tau}}_n)| + |P_n l(\widehat{\boldsymbol{\tau}}_n) - Pl(\boldsymbol{\tau}_0)|,$$

we then wish to show that each term on the right hand side converges to 0 (a.s.). For the first term, we just need to implement the result from part (3) below, but the second term demands further analyses. Let $\boldsymbol{\tau}_{0n} = (\boldsymbol{\beta}_0, \widetilde{h}_0(t)) \in B * A_n$ which satisfies $\rho(\boldsymbol{\tau}_{0n}, \boldsymbol{\tau}_0) \to 0$ (as $n \to \infty$) according to Assumption A4. Since $\boldsymbol{\tau}_0$ maximizes $Pl(\boldsymbol{\tau})$ for $\boldsymbol{\tau} \in B * A$ and $\widehat{\boldsymbol{\tau}}_n$ maximizes $P_n l(\boldsymbol{\tau})$ for $\boldsymbol{\tau} \in B * A_n$, we have

$$\begin{aligned}
&P_n l(\boldsymbol{\tau}_{0n}) - Pl(\boldsymbol{\tau}_{0n}) + Pl(\boldsymbol{\tau}_{0n}) - Pl(\boldsymbol{\tau}_0) \leq P_n l(\widehat{\boldsymbol{\tau}}_n) - Pl(\boldsymbol{\tau}_0) \\
&\leq P_n l(\widehat{\boldsymbol{\tau}}_n) - Pl(\widehat{\boldsymbol{\tau}}_n).
\end{aligned}$$

From part (3) below we have both $P_n l(\widehat{\boldsymbol{\tau}}_n) - Pl(\widehat{\boldsymbol{\tau}}_n)$ and $P_n l(\boldsymbol{\tau}_{0n}) - Pl(\boldsymbol{\tau}_{0n})$ converge to 0 (a.s.). $Pl(\boldsymbol{\tau}_{0n}) - Pl(\boldsymbol{\tau}_0)$ converges to 0 can be established from $\rho(\boldsymbol{\tau}_{0n}, \boldsymbol{\tau}_0) \to 0$ and the fact that $l(\cdot)$ is continuous and bounded.

(3) It suffices to demonstrate $\sup_{\boldsymbol{\tau}_n \in B*A_n} |P_n l(\boldsymbol{\tau}_n) - Pl(\boldsymbol{\tau}_n)| \to 0$ (a.s.). This can be achieved through the following steps:

(i) Firstly, we show that $N(\varepsilon, A_n, L_\infty) \leq (6C_6/\varepsilon)^m$ where constant C_6 will be specified below. This is because for any $\widetilde{h}_1, \widetilde{h}_2 \in A_n$ (note that $\widetilde{h}_e(t) = \sum_u \theta_u^e \psi_u(t)$), $\max_t |\widetilde{h}_1(t) - \widetilde{h}_2(t)| \leq C_5 \max_u |\theta_u^1 - \theta_u^2| \leq C_5 C_6$, where C_5 and C_6 are respectively the upper bound of $\sum_u \psi_u(t)$ and $\{\theta_u, \forall u\}$. Thus, $N(\varepsilon, A_n, L_\infty) \leq N(\varepsilon, \{\boldsymbol{\theta} : 0 \leq \theta_u \leq C_6 : 1 \leq u \leq m\}, L_2)$. From Lemma 4.1 of Pollard (1984) we have that $N(\varepsilon, \{\boldsymbol{\theta} : 0 \leq \theta_u \leq C_6 : 1 \leq u \leq m\}, L_2) \leq (6C_6/\varepsilon)^m$.

(ii) Secondly, we first define the set $\Lambda_n = \{l(\boldsymbol{\tau}) : \boldsymbol{\tau} \in B * A\}$. We wish to demonstrate that $N(\varepsilon, \Lambda_n, L_2 * L_\infty) \leq K/\varepsilon^{p+m}$, where constant K will be defined below. In fact, from Taylor expansion and the assumption that log-likelihood $l(\boldsymbol{\tau})$ is continuous and bounded, we have for $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in B * A$,

$$|l(\boldsymbol{\tau}_1) - l(\boldsymbol{\tau}_2)| < M(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 + |h_{01}(t) - h_{02}(t)|_\infty),$$

where $M > 0$ is a constant. Then (see the Appendix of Xue et al. (2004))

$$\begin{aligned} &N(\varepsilon, \Lambda_n, L_2 * L_\infty) \leq N(\varepsilon/2, B, L_2) \cdot N(\varepsilon/2, A_n, L_\infty) \\ &\leq (6C_1/\varepsilon)^p (6C_6/\varepsilon)^m = K/\varepsilon^{p+m}, \end{aligned}$$

where C_1 is the upper bound of $\{|\beta_j|, \forall j\}$ and $K = (C_1/C_6)^p$. Note that ε in the last term is in fact $\varepsilon/(6C_6)$.

(iii) Select $\alpha_n = n^{-1/2+\phi_1}\sqrt{\log n}$ where $\phi_1 \in (\phi_0/2, 1/2)$ with $\phi_0 < 1$, and define $\varepsilon_n = \varepsilon\alpha_n$. Following the proof of Theorem 1 of Xue et al. (2004) we can show that

$$\begin{aligned} &P\left\{\sup_{\Lambda_n} |P_n l(\tau_n) - Pl(\tau_n)| > 8\varepsilon_n\right\} \leq 8N(\varepsilon_n, \Lambda_n, L_\infty)e^{-n\varepsilon_n^2/128} \\ &\leq 8Ce^{-C'n^{2\phi_1}\log n}, \end{aligned}$$

where$C$ and C' are constants. Hence, $\sum_{n=1}^{\infty} P[\sup_{\Lambda_n} |P_n l(\tau_n) - Pl(\tau_n)| > 8\varepsilon_n]$ is a convergent series, and therefore, by the Borel-Cantelli lemma, we have

$$\sup_{\Lambda_n} |P_n l(\tau_n) - Pl(\tau_n)| \to 0$$

almost surely. Thus $P_n l(\tau_n) - Pl(\tau_n) \to 0$ almost surely $\forall \boldsymbol{\tau}_n \in B * A_n$.

S2.2 Sketch proof of Theorem 2

We first state the following assumptions needed for the asymptotic results in Theorem 2.

B1. Assume the distribution of $\mathbf{x}_i$ is independent of $\boldsymbol{\eta}$.

B2. Assume the limit $\lim_{n\to\infty}[n^{-1}l(\boldsymbol{\eta})]$ exists and has a unique maximum at $\boldsymbol{\eta}_0 \in \Omega$, where Ω is the parameter set for $\boldsymbol{\eta}$ and is a compact subspace of R^{p+m}.

B3. Assume $l(\boldsymbol{\eta})$ has a finite upper bound, $l(\boldsymbol{\eta})$ is twice continuously differentiable in a neighbourhood of $\boldsymbol{\eta}_0$ and the matrices

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \frac{\partial l_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} \text{ and } \lim_{n\to\infty} \left[-n^{-1}\frac{\partial^2 l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}\right] \tag{S3}$$

exists.

B4. Assume the penalty function $J(\boldsymbol{\eta})$ is twice continuously differentiable on Ω, and their derivatives are bounded.

B5. Assume the matrix $\mathbf{U}^T\mathbf{F}(\boldsymbol{\eta})\mathbf{U}$ is invertible in a neighbuorhood of $\boldsymbol{\eta}_0$, where matrix $\mathbf{U}$ is defined in the main paper.

Assumption B2 simply states that the true parameters can be recovered exactly from maximizing the likelihood if the sample size is infinity.

Let $\bar{l}(\boldsymbol{\eta}) = \lim_{n\to\infty}[n^{-1}l(\boldsymbol{\eta})]$. It follows from the strong law of large numbers that $n^{-1}l(\boldsymbol{\eta}) \to \bar{l}(\boldsymbol{\eta})$ almost surely and uniformly for $\boldsymbol{\eta} \in \Omega$. This result, together with $\mu_n \to 0$ as $n \to \infty$ and $\boldsymbol{\eta}_0$ being the unique maximum of $\bar{l}(\boldsymbol{\eta})$ due to Assumption B2, implies that $\widehat{\boldsymbol{\eta}} \to \boldsymbol{\eta}_0$ (recall $\widehat{\boldsymbol{\eta}}$ maximizes $n^{-1}l(\boldsymbol{\eta})$ for $\boldsymbol{\eta} \in \Omega$) almost surely by applying, for example, Corollary 1 of Honore and Powell (1994).

Next we prove the asymptotic normality result. From the KKT necessary conditions (6) and (7) in the main paper we have that the constrained MPL estimate $\widehat{\boldsymbol{\eta}}$ satisfies

$$\mathbf{U}^\top \frac{\partial \Phi(\widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} = 0 \tag{S4}$$

According to Taylor expansion

$$\frac{\partial \Phi(\widehat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}} = \frac{\partial \Phi(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} + \frac{\partial^2 \Phi(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0), \tag{S5}$$

where $\widetilde{\boldsymbol{\eta}}$ is a vector between $\widehat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_0$. Therefore

$$0 = \mathbf{U}^\top \frac{\partial \Phi(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} + \mathbf{U}^\top \frac{\partial^2 \Phi(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^{\mathrm{T}}}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0). \tag{S6}$$

Next, let $\widehat{\boldsymbol{\chi}}$ be $\widehat{\boldsymbol{\eta}}$ after deleting the active constraints and $\boldsymbol{\chi}_0$ be similarly defined corresponding to $\boldsymbol{\eta}_0$, then

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \mathbf{U}(\widehat{\boldsymbol{\chi}} - \boldsymbol{\chi}_0). \tag{S7}$$

Substituting (S7) into (S6), solving for $\widehat{\boldsymbol{\chi}} - \boldsymbol{\chi}_0$ and then using (S7) to convert the result back to $\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0$ again, we eventually have

$$\sqrt{n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) = -\mathbf{U}\left(\mathbf{U}^\top \frac{1}{n}\frac{\partial^2 \Phi(\widetilde{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^\top}\mathbf{U}\right)^{-1} \mathbf{U}^\top \left(\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\eta}_0)}{\partial \boldsymbol{\eta}} + o(1)\right). \tag{S8}$$

In (S8), when $n \to \infty$ and $\mu_n \to 0$, $-n^{-1}\partial^2\Phi(\widetilde{\boldsymbol{\eta}})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top$ converges to $F(\boldsymbol{\eta}_0)$ (a.s.) by the law of large numbers. On the other hand, after applying the central limit theorem to $n^{-1/2}\partial l(\boldsymbol{\eta}_0)/\partial\boldsymbol{\eta}$ we have the required asymptotic normality result.

S3 M-spline and Gaussian basis functions

If the baseline hazard is approximated by Gaussian basis functions, $\psi_u(t)$ is a truncated Gaussian distribution with location parameter α_u (which are knots), scale parameter σ_u and range $[t_{(1)}, t_{(n)}]$. This leads to the following expressions of $\psi_u(t)$ and its cumulative function $\Psi_u(t)$:

$$\psi_u(t) = \frac{1}{\sigma_u \delta_u} \phi\left(\frac{t-\alpha_u}{\sigma_u}\right),$$
$$\Psi_u(t) = \int_{t_{(1)}}^{t} \psi_u(v)dv = \frac{1}{\Delta_u}\left[\Phi\left(\frac{t-\alpha_u}{\sigma_u}\right) - \Phi\left(\frac{t_{(1)}-\alpha_u}{\sigma_u}\right)\right],$$

where $t_{(1)} \leq t \leq t_{(n)}$, $\phi(\cdot)$ and $\Phi(\cdot)$ respectively are the density and cumulative density functions of the standard Gaussian distribution, $\Delta_u = \Phi((t_{(n)}-\alpha_u)/\sigma_u) - \Phi((t_{(1)}-\alpha_u)/\sigma_u)$ and $\alpha_u \in \mathbb{R}$ and $\sigma_u > 0$. If the baseline hazard is approximated by means of M−spline basis functions of order o (Ramsay, 1988), we get the following expressions of $\psi_u^o(t)$ and $\Psi_u^o(t)$:

$$\psi_u^o(t) = \begin{cases} \frac{\delta(\alpha_u^\star \leq t < \alpha_{u+1}^\star)}{\alpha_{u+1}^\star - \alpha_u^\star} & \text{if } o = 1, \\ \frac{o}{o-1}\frac{\delta(\alpha_u^\star \leq t < \alpha_{u+o}^\star)}{\alpha_{u+o}^\star - \alpha_u^\star}\left[(t-\alpha_u^\star)\,\psi_u^{o-1}(t) + \left(\alpha_{u+o}^\star - t\right)\psi_{u+1}^{o-1}(t)\right] & \text{otherwise,} \end{cases}$$
$$\Psi_u^o(t) = \delta(\alpha_u^\star > t)\left[\sum_{v=u+1}^{\min(u+o,m+1)} \frac{\alpha_{v+o+1}^{\star\star} - \alpha_v^{\star\star}}{o+1}\psi_v^{o+1}(t)\right]^{\delta(\alpha_u^\star < t < \alpha_{u+o}^\star)},$$

where $t_{(1)} \leq t \leq t_{(n)}$, α_u is the uth element of knots vector $\boldsymbol{\alpha}$ whose length is n_α ($\alpha_u \in \mathbb{R}$ and $\alpha_u < \alpha_{u+1}$), $m = n_\alpha + o - 2$, $\boldsymbol{\alpha}^\star = \left[\min(\boldsymbol{\alpha})\mathbf{1}_{o-1}^\top, \boldsymbol{\alpha}^\top, \max(\boldsymbol{\alpha})\mathbf{1}_{o-1}^\top\right]^\top$, and $\boldsymbol{\alpha}^{\star\star} = \left[\min(\boldsymbol{\alpha})\mathbf{1}_{o}^\top, \boldsymbol{\alpha}^\top, \max(\boldsymbol{\alpha})\mathbf{1}_{o}^\top\right]^\top$, where $\mathbf{1}_o$ denotes a vector of 1 of length o. Note that $\Psi_u^o(t)$, the cumulative function of $\psi_u^o(t)$, is referred to as an I−spline. M−spline basis functions have the following properties: $\int_{-\infty}^{\alpha_u^\star} \psi_u^o(v)dv = 0$, $\int_{\alpha_u^\star}^{\alpha_{u+o}^\star} \psi_u^o(v)dv = 1$ and $\int_{\alpha_{u+o}^\star}^{\alpha_\infty^\star} \psi_u^o(v)dv = 0$.

S4 More simulation results

In this section we present more simulation results for MPL in Table S1 where we increased the number of knots for the purpose of demonstrating MPL is less impacted by the number of knots. In Table S2 we provide the regression coefficient estimates by the competitor methods.

		$\pi^E = 0\%$		$\pi^E = 25\%$		$\pi^E = 50\%$	
		$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
$\boldsymbol{\beta}$ estimates:							
Biases							
β_1	MPL-M(+2)	0.004	-0.022	-0.054	-0.028	-0.054	-0.026
	MPL-M(×2)	0.097	0.001	0.022	-0.024	-0.012	-0.024
β_2	MPL-M(+2)	0.013	-0.009	-0.044	-0.017	-0.064	-0.021
	MPL-M(×2)	0.108	0.013	0.031	-0.013	-0.023	-0.019
β_3	MPL-M(+2)	0.020	-0.005	-0.043	-0.018	-0.056	-0.018
	MPL-M(×2)	0.113	0.018	0.033	-0.014	-0.014	-0.016
Mean asymptotic and (Monte Carlo) standard errors							
β_1	MPL-M(+2)	0.332 (0.352)	0.143 (0.146)	0.283 (0.285)	0.124 (0.126)	0.253 (0.251)	0.111 (0.114)
	MPL-M(×2)	0.345 (0.388)	0.144 (0.151)	0.290 (0.313)	0.124 (0.126)	0.255 (0.263)	0.111 (0.115)
β_2	MPL-M(+2)	0.121 (0.137)	0.053 (0.055)	0.103 (0.110)	0.046 (0.046)	0.092 (0.096)	0.041 (0.041)
	MPL-M(×2)	0.127 (0.153)	0.054 (0.057)	0.106 (0.120)	0.046 (0.047)	0.094 (0.100)	0.042 (0.041)
β_3	MPL-M(+2)	0.084 (0.090)	0.036 (0.034)	0.072 (0.072)	0.031 (0.030)	0.064 (0.064)	0.028 (0.026)
	MPL-M(×2)	0.088 (0.097)	0.037 (0.036)	0.074 (0.081)	0.031 (0.030)	0.065 (0.067)	0.028 (0.027)
95% coverage probabilities							
β_1	MPL-M(+2)	0.941	0.944	0.949	0.952	0.955	0.938
	MPL-M(×2)	0.920	0.936	0.940	0.952	0.942	0.935
β_2	MPL-M(+2)	0.922	0.945	0.932	0.939	0.919	0.939
	MPL-M(×2)	0.900	0.947	0.928	0.939	0.940	0.941
β_3	MPL-M(+2)	0.943	0.957	0.948	0.950	0.942	0.957
	MPL-M(×2)	0.923	0.956	0.937	0.952	0.938	0.956
$h_0(t)$ **estimates**							
Biases							
$h_0[F_0^{-1}(0.25)]$	MPL-M(+2)	0.134	0.052	0.114	0.030	0.092	0.017
	MPL-M(×2)	0.077	0.071	0.079	0.035	0.062	0.019
$h_0[F_0^{-1}(0.50)]$	MPL-M(+2)	0.123	0.042	0.073	0.036	0.052	0.025
	MPL-M(×2)	0.306	0.061	0.163	0.036	0.093	0.026
$h_0[F_0^{-1}(0.75)]$	MPL-M(+2)	0.007	-0.034	-0.024	0.000	-0.028	0.004
	MPL-M(×2)	0.256	0.003	0.118	0.004	0.047	0.006
Mean asymptotic (Monte Carlo) standard errors							
$h_0[F_0^{-1}(0.25)]$	MPL-M(+2)	0.556 (0.627)	0.227 (0.240)	0.477 (0.498)	0.197 (0.207)	0.422 (0.431)	0.176 (0.179)
	MPL-M(×2)	0.553 (0.710)	0.240 (0.268)	0.474 (0.583)	0.199 (0.212)	0.416 (0.451)	0.177 (0.180)
$h_0[F_0^{-1}(0.50)]$	MPL-M(+2)	1.142 (1.397)	0.460 (0.500)	0.940 (1.021)	0.399 (0.414)	0.825 (0.835)	0.355 (0.359)
	MPL-M(×2)	1.390 (1.825)	0.484 (0.543)	1.047 (1.193)	0.400 (0.416)	0.868 (0.928)	0.355 (0.360)
$h_0[F_0^{-1}(0.75)]$	MPL-M(+2)	2.025 (2.664)	0.871 (0.965)	1.678 (1.945)	0.772 (0.805)	1.494 (1.614)	0.692 (0.711)
	MPL-M(×2)	2.689 (3.992)	0.944 (1.046)	1.999 (2.414)	0.778 (0.820)	1.635 (1.816)	0.693 (0.714)
95% coverage probabilities							
$h_0[F_0^{-1}(0.25)]$	MPL-M(+2)	0.914	0.947	0.923	0.936	0.929	0.935
	MPL-M(×2)	0.844	0.938	0.866	0.937	0.913	0.936
$h_0[F_0^{-1}(0.50)]$	MPL-M(+2)	0.903	0.936	0.912	0.935	0.923	0.948
	MPL-M(×2)	0.927	0.939	0.928	0.938	0.926	0.948
$h_0[F_0^{-1}(0.75)]$	MPL-M(+2)	0.807	0.874	0.826	0.923	0.851	0.936
	MPL-M(×2)	0.888	0.908	0.894	0.922	0.888	0.940
Integrated discrepancy between $\widehat{h}_0(t)$ and $h_0(t)$ for $t < F_0^{-1}(0.9)$							
	MPL-M(+2)	1.666	0.732	1.351	0.591	1.160	0.518
	MPL-M(×2)	2.171	0.779	1.560	0.602	1.226	0.520

Table S1: MPL-M results for different number of knots using Simulation 1 data. MPL-M(+2): $n_\alpha = 9, 11$ for $n = 100$ and $n = 500$ respectively; MPL-M(×2): $n_\alpha = 14, 18$ for $n = 100$ and $n = 500$ respectively.

Figure S1 displays the true $h_0(t)$, the average MPL $h_0(t)$ estimates and pointwise 95% confidence intervals for simulations 1 and 2. The standard approach (Breslow) and the CM method provide biased noisy estimates in both simulations. The bump observed around $t = 4.5$ for Breslow is caused by the midpoint strategy to left censored data. More specifically, left censored events are shrunk towards zero with a maximum around 0.45 (which is half the maximum of 0.9 chosen for this type of censoring). The ugly kick observed after 0.45 comes from the low event density right after 0.45 created artificially by using the midpoint strategy. Evidence of this is provided in Figure S2 (top panel). Slightly better results may be obtained using the Nelson-Aaalen estimates (see bottom and middle panels), but the issue remains. Further justification of the cause of the problem can be provided by modifying the strategy (e.g. adding a small random noise or using the true event times) which makes the kick disappear – results not shown). In contrast, both the EM-I and MPL methods provide reliable estimates.

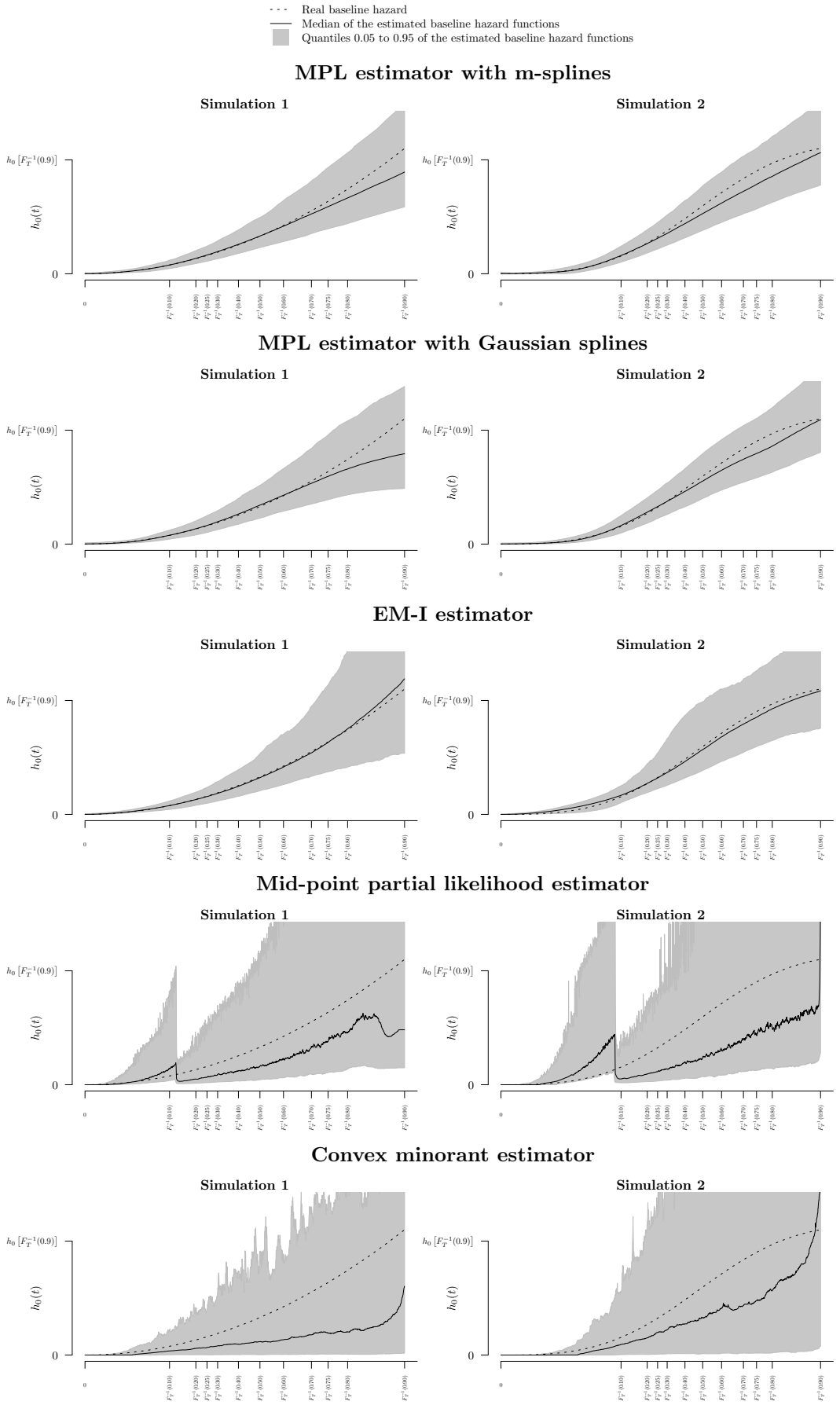

Figure S1: Quantiles 0.025, 0.5 and 0.975 of the baseline hazard estimates for each estimator (rows) and simulation (columns) in the scenario considering 0% event and n=500.

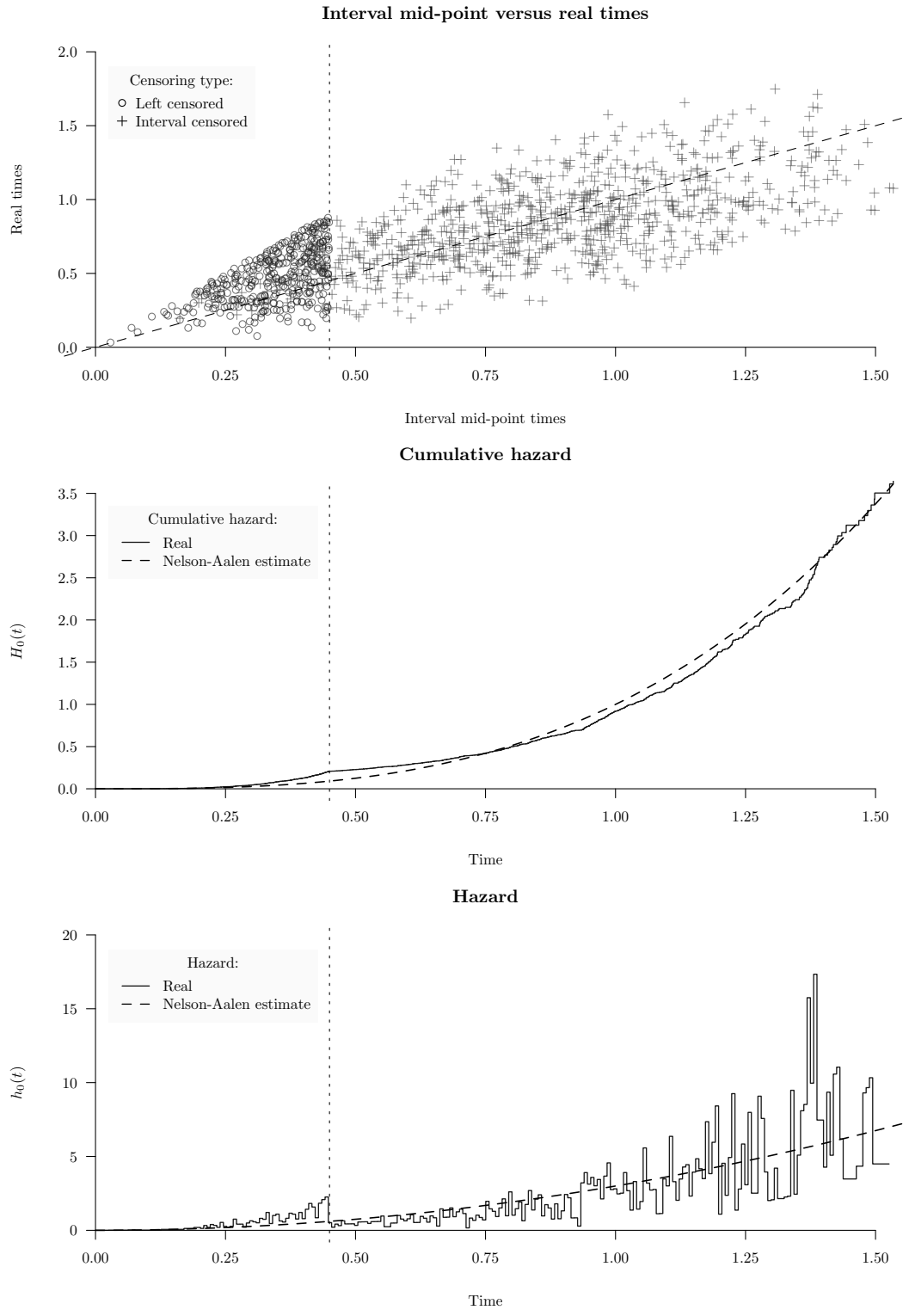

Figure S2: Real versus corresponding interval censored data for a Monte Carlo sample of simulation 1 (Weibull hazard) with 0% event and n=2000 (upper plot) and corresponding Nelson-Aalen cumulative hazard (mid plot) and hazard rate (lower plot) estimates after fitting the Cox regression parameters by means of the partial likelihood estimator with mid interval points.

Predictors	Levels	**Estimators**	**HR estimates**	**HR 95% CI**	p**-value**
Location	Arm	MPL-M	0.570	[0.429;0.757]	0.0001
		MPL-G	0.569	[0.428;0.756]	0.0001
		EM-I	0.573	[0.430;0.762]	0.0001
		PL	0.571	[0.431;0.758]	0.0001
		CM	0.572		
	Leg	MPL-M	1.008	[0.816;1.244]	0.9430
		MPL-G	1.008	[0.816;1.244]	0.9429
		EM-I	1.006	[0.812;1.247]	0.9550
		PL	1.001	[0.812;1.234]	0.9949
		CM	0.988		
	Trunk	MPL-M	0.802	[0.658;0.977]	0.0283
		MPL-G	0.801	[0.658;0.976]	0.0278
		EM-I	0.803	[0.656;0.983]	0.0331
		PL	0.798	[0.656;0.971]	0.0243
		CM	0.794		
Thickness	1 to 2 mm.	MPL-M	1.245	[0.937;1.653]	0.1303
		MPL-G	1.246	[0.937;1.656]	0.1297
		EM-I	1.242	[0.914;1.687]	0.1664
		PL	1.240	[0.936;1.644]	0.1340
		CM	1.255		
	2 to 4 mm.	MPL-M	2.390	[1.804;3.166]	<0.0001
		MPL-G	2.399	[1.810;3.181]	<0.0001
		EM-I	2.372	[1.750;3.214]	<0.0001
		PL	2.399	[1.814;3.171]	<0.0001
		CM	2.438		
	4 mm. and more	MPL-M	3.108	[2.295;4.208]	<0.0001
		MPL-G	3.121	[2.303;4.231]	<0.0001
		EM-I	3.069	[2.213;4.256]	<0.0001
		PL	3.152	[2.332;4.260]	<0.0001
		CM	3.250		
Gender	Female	MPL-M	0.843	[0.717;0.990]	0.0375
		MPL-G	0.842	[0.717;0.990]	0.0377
		EM-I	0.845	[0.719;0.993]	0.0412
		PL	0.838	[0.713;0.983]	0.0304
		CM	0.841		
Centered Age (10 years)	-	MPL-M	1.148	[1.093;1.205]	<0.0001
		MPL-G	1.149	[1.094;1.207]	<0.0001
		EM-I	1.145	[1.091;1.203]	<0.0001
		PL	1.156	[1.100;1.214]	<0.0001
		CM	1.179		

Table S2: Hazard ratio estimates ($e^{\widehat{\beta}}$), hazard ratio 95% confidence intervals, and p-values of the significant tests per model parameter and estimator. We used 7 quantile based internal knots for the MPL estimator and 4 quantile based internal knots for the EM-I estimator. The latter was selected by minimising the AIC criteria. The CM estimator as implemented in the intcox package did not iterate so that the final estimate correspond to its starting point corresponding to the partial likelihood estimates in which the upper interval bounds of interval censored data were considered as events.

References

Cai, T. and R. A. Betensky (2003): "Hazard regression for interval-censored data with penalized spline," *Biometrics*, 59, 570–579.

DeBoor, C. and J. W. Daniel (1974): "Splines with nonnegative B-spline coefficients," *Mathe-*

matics of Computation, 126, 565–568.

Honore, B. E. and J. L. Powell (1994): "Pairwise difference estimators of censored and truncated regression models," *Journal of Econometrics*, 64, 241–278.

Huang, J. (1996): "Efficient estimation for the proportional hazard model with interval censoring," *The Annals of Statistics*, 24, 540 – 568.

Pollard, D. (1984): *Convergence of stochastic processes*, New York: Springer Verlag.

Ramsay, J. O. (1988): "Monotone regression splines in action," *Statistical Science*, 3, 425–441.

Wong, W. H. and X. Shen (1995): "Probability inequalities for likelihood ratios and convergene rates of sieve MLEs," *Ann. Statist.*, 23, 339–362.

Xue, H., K. F. Lam, and G. Li (2004): "Sieve maximum likelihood estimator for semiparametric regression models with current status data," *J. Amer. Statist. Assoc.*, 99, 346–356.

Zhang, Y., L. Hua, and J. Huang (2010): "A spline-based semiparametric maximum likelihood esimation method for the Cox model with interval-censored data," *Scand. J. Statist.*, 37, 338–354.