Web-based Supplementary Materials for "A Pairwise Likelihood Augmented Cox Estimator for Left-Truncated Data" by

Fan Wu, Sehee Kim, Jing Qin, Rajiv Saran and Yi Li*

*email: yili@umich.edu

A Additional Theoretical Results

A.1 Detailed Proofs

This subsection contains detailed proofs of the asymptotic properties of the proposed pairwise likelihood augmented Cox (PLAC) estimator. The proofs are provided under the following regularity conditions which hold in most practical scenarios, such as in the RRI-CKD study demonstrated in Section 4.

(C1) The true regression coefficient vector β_0 lies in the interior of a compact set $B \subset \mathbb{R}^p$. The true cumulative baseline hazard function $\Lambda_0(t)$ is continuously differentiable and strictly increasing on $[0, \tau]$, and satisfies $\Lambda_0(0) = 0$.

(C2) The vector Z is bounded almost surely. If there exist a deterministic function $\gamma_0(t)$ and a vector $\gamma \in \mathbb{R}^p$, such that $\gamma_0(t) + \gamma^{\mathrm{T}} Z = 0$ with probability one, then $\gamma_0(t) = 0$ and $\gamma = 0$.

(C3) With probability one, there exists a constant $\delta_1 > 0$ such that $\mathrm{pr}(A^* \leqslant T^* \mid Z^*) > \delta_1$, $\mathrm{pr}(A + C \geqslant \tau \mid Z) > \delta_1$, and $\mathrm{pr}(\bar{Y}(\tau) = 1 \mid Z) > \delta_1$, where $\bar{Y}(\tau) = 1$ implies that $Y(t) = 1$ for all $t \in [A, \tau]$.

(C4) Let $b \in \mathbb{R}^p$, and h be a function with bounded total variation on $[0, \tau]$, then the information operator corresponding to the conditional likelihood evaluated at $(\beta_0,\, \Lambda_0)$,

$$J_0^C(b,\, h) = \left(\lim_{n\to\infty} \frac{\partial U^C(\beta, \Lambda)}{\partial(\beta, \Lambda)}\bigg|_{\beta=\beta_0,\, \Lambda=\Lambda_0} \right)(b,\, h),$$

is invertible.

Conditions (C1)-(C4) are standard assumptions for the Cox model under left-truncation that are necessary to prove the identifiability of the parameters as well as the existence and uniqueness of the PLAC estimator. The continuity of $\Lambda_0(t)$ facilitates proving the uniform convergence of $\hat{\Lambda}(t)$, and the strictly monotonicity suggests that events can happen at any time during the study period. The boundedness assumption in (C2) is important for the uniform convergence of the function classes involved, and the second assumption ensures the covariates are not degenerate and that the parameters are identifiable. Condition (C3) implies that there is a positive probability that events are observable during the study period, and that there is a positive probability that some subjects are still at risk at the end of the follow-up. Condition (C4), which is used to show the root of the composite score equations is unique, is adapted from the classic weak convergence proof for the Cox model (see Andersen et al., 1993, Condition VII2.1(e)). In addition, if the probability of occurring $(T^* < A^* \mid Z^*)$ is positive, then $\mathcal{L}_n^P$ is non-degenerate, so that we can attain efficiency gain beyond the conditional approach. When this condition does not hold, $\mathcal{L}_n^P$ is zero, and thus the PLAC estimator becomes identical to the conditional approach estimator.

We use Ω to denote the set of all possible observations. For convenience, we adopt the de Finetti's linear functional notations (Pollard, 2002), where $\mathbb{P}_n$ denotes the empirical measure of the observations $\mathcal{O}_i$, $i = 1, \ldots, n$, P_0 denotes the true probability measure on Ω, and $\mathbb{U}_{n,2}$ is the empirical measure of pairs $(\mathcal{O}_i, \mathcal{O}_j)$ such that $1 \leqslant i < j \leqslant n$.

A.1.1 Strong Consistency

The PLAC estimator falls in the category of Z-estimators. To follow the consistency proof for general Z-estimators, a complication brought by the pairwise structure is to show the uniform convergence of the involved bivariate function classes. We tackle this difficulty through bounding the bracketing numbers (entropies) of these function classes using the U-processes theory (see De la Peña and Giné, 1999, Chapter 5). For $k = 0, 1, 2$, the function classes $\{(z_1, z_2) \mapsto z_1^{\otimes k} e^{z_1^{\mathrm{T}}\beta} - z_2^{\otimes k} e^{z_2^{\mathrm{T}}\beta} : z_1, z_2 \in \mathbb{R}^p;\ \beta \in B\}$ are Euclidean (Nolan and Pollard, 1987); thus, their bracketing numbers in $L_1(P^2)$ are finite, where $P^2 \equiv P \otimes P$, and P is any probability measure. Bounds for classes only consisting of indicator functions can be shown using the Vapnik-Chervonenkis theory (see De la Peña and Giné, 1999, Section 5.2). Denoting the class of cumulative baseline hazard functions satisfying (C1) as $\mathcal{H}_\Lambda$, then

Lemma 1. *The bivariate function class* $\mathcal{H}_\Lambda^D = \{(s,t) \mapsto \Lambda(s) - \Lambda(t) : s, t \in [0, \tau];\ \Lambda \in \mathcal{H}_\Lambda\}$ *has finite bracketing numbers in* $L_1(P^2)$ *for all* $\varepsilon > 0$.

Proof. To avoid technicality, we assume all bivariate function classes involved in this and the following proofs are measurable (see De la Peña and Giné, 1999, Section 3.5). Theorem 2.7.5 of van der Vaart and Wellner (1996) indicates that for a fixed $\varepsilon > 0$, there exists a constant K_1 such that the bracketing entropy

$$\log N_{[]}(\varepsilon, \mathcal{H}_\Lambda, L_1(P)) < \frac{K_1}{\varepsilon} < \infty$$

for any probability measure P. For a given $\Lambda \in \mathcal{H}_\Lambda$, suppose an ε-bracket containing it in $L_1(P)$ is (Λ_l, Λ_u); thus, we have $\Lambda_l(t) < \Lambda(t) < \Lambda_u(t)$, $\forall t \in [0, \tau]$ and that

$$\int |\Lambda_u(s) - \Lambda_l(s)|\, dP < \varepsilon,$$

where the integral without limits here and below are taken from 0 to τ. Then for the

corresponding bivariate function in $\mathcal{H}_\Lambda^D$, we have

$$\Lambda_l(s) - \Lambda_u(t) < \Lambda(s) - \Lambda(t) < \Lambda_u(s) - \Lambda_l(t), \quad \forall s, t \in [0, \tau]\,.$$

By triangle inequality,

$$\begin{aligned} & \iint |\Lambda_u(s) - \Lambda_l(t) - \Lambda_l(s) + \Lambda_u(t)|\, dP^2 \\ \leqslant & \int\int |\Lambda_u(s) - \Lambda_l(s)|\, dPdP + \int\int |\Lambda_u(t) - \Lambda_l(s)|\, dPdP \\ = & \int |\Lambda_u(s) - \Lambda_l(s)|\, dP + \int |\Lambda_u(t) - \Lambda_l(s)|\, dP < 2\varepsilon. \end{aligned}$$

Therefore, $(\Lambda_l(s) - \Lambda_u(t),\ \Lambda_u(s) - \Lambda_l(t))$ is a (2ε)-bracket for $\Lambda(s) - \Lambda(t)$ in $L_1(P^2)$, thus there is a constant $K_2 > 0$ such that the bracketing entropy

$$\log N_{[]}(\varepsilon, \mathcal{H}_\Lambda^D, L_1(P^2)) < \frac{K_2}{\varepsilon} < \infty.$$

Since ε is arbitrary, the class $\mathcal{H}_\Lambda^D$ has finite bracketing numbers in $L_1(P^2)$. □

Remark 1. By Corollary 5.2.5 of De la Peña and Giné (1999), the finite bracketing numbers imply the corresponding function class satisfies the uniform law of large numbers of U-process. The uniform law of large numbers for $U^P(\beta, \Lambda)$ and its derivatives then follows, because they are Lipschitz functions of the component functions with finite bracketing numbers (van der Vaart and Wellner, 1996).

Proof of Theorem 1. We first re-write the modified composite log-likelihood (2) and the composite score functions using the linear functional notations. Let $N_i(s) = \Delta_i I(X_i \leqslant s)$ be the observed event counting process for subject i, then (2) can be written as

$$\begin{aligned} \ell_n^c(\beta, \Lambda) &= \mathbb{P}_n \int_0^\tau \left\{(\log \Lambda\{s\} + Z^{\mathrm{T}}\beta) dN(s) - Y(s) e^{Z^{\mathrm{T}}\beta} d\Lambda(s)\right\} \\ &\quad - \mathbb{U}_{n,2} \log(1 + R(\beta, \Lambda)). \end{aligned}$$

Differentiating it with respect to β yields the composite score function for β:

$$\begin{aligned} U_\beta(\beta,\Lambda) &= \mathbb{P}_n \int_0^\tau Z\left\{dN(s) - Y(s)e^{Z^{\mathrm{T}}\beta}d\Lambda(s)\right\} \\ &\quad -\mathbb{U}_{n,2}\left\{\frac{R(\beta,\Lambda)}{1+R(\beta,\Lambda)}\int_0^\tau Q^{(1)}(s;\beta)d\Lambda(s)\right\}. \end{aligned}$$

For $0 \leqslant t \leqslant \tau$ and $h(\cdot) = I(\cdot \leqslant t)$, define a perturbation of Λ by $d\Lambda_\varepsilon = (1+\varepsilon h)d\Lambda$. The derivative of $\ell_n^c(\beta,\Lambda_\varepsilon)$ with respect to ε evaluated at $\varepsilon = 0$ yields the composite score function for Λ in the direction of h:

$$\begin{aligned} U_\Lambda(\beta,\Lambda)(h) &= \mathbb{P}_n \int_0^\tau h(s)\left\{dN(s) - Y(s)e^{Z^{\mathrm{T}}\beta}d\Lambda(s)\right\} \\ &\quad -\mathbb{U}_{n,2}\left\{\frac{R(\beta,\Lambda)}{1+R(\beta,\Lambda)}\int_0^\tau Q^{(0)}(s;\beta)h(s)d\Lambda(s)\right\}. \end{aligned}$$

As in Section 2.3, we can write the composite score function

$$U(\beta,\Lambda) = \begin{pmatrix} U_\beta(\beta,\Lambda) \\ U_\Lambda(\beta,\Lambda)(h) \end{pmatrix}$$

as the summation of $U^C(\beta,\Lambda)$ and $U^P(\beta,\Lambda)$; the former is the conditional approach score function and has expectation zero. We can also show that $\mathrm{E}_0\{U^P(\beta_0,\Lambda_0)\} = 0$, since the summand of U^P satisfies $\mathrm{E}_0\{U_{ij}^P(\beta_0,\Lambda_0)\} = 0$, $1 \leqslant i < j \leqslant n$. To see this, note that the pair $(A_i,\ A_j)$ has a binary distribution after conditioning on (Z_i, Z_j) and the order statistics of $(A_i,\ A_j)$; thus, by double expectation, we have

$$\begin{aligned} \mathrm{E}_0\{U_{ij}^P(\beta,\Lambda)\} &= \mathrm{E}_0\left\{\frac{1}{1+R_{ij}^{-1}(\beta,\Lambda)}\cdot\frac{1}{1+R_{ij}(\beta_0,\Lambda_0)}\begin{pmatrix}\int Q^{(1)}(s;\beta)d\Lambda(s) \\ \int h(s)Q^{(0)}(s;\beta)d\Lambda(s)\end{pmatrix}\right. \\ &\quad \left. -\frac{1}{1+R_{ij}(\beta,\Lambda)}\cdot\frac{1}{1+R_{ij}^{-1}(\beta_0,\Lambda_0)}\begin{pmatrix}\int Q^{(1)}(s;\beta)d\Lambda(s) \\ \int h(s)Q^{(0)}(s;\beta)d\Lambda(s)\end{pmatrix}\right\}. \quad \text{(A.1)} \end{aligned}$$

The two terms in the bracket cancel if and only if $\beta = \beta_0$ and $\Lambda = \Lambda_0$ by the identifiability of the parameters. Specifically, outside a set with zero probability, by Condition (C1), for almost every pair of (z_1, z_2) in the support of the density of the covariates and every pair of $(A_1,\ A_2)$ such that $\Lambda_0(A_1) - \Lambda_0(A_2) > 0$, we have from $R_{ij}(\beta, \Lambda) = R_{ij}(\beta_0, \Lambda_0)$ that

$$\frac{\Lambda(A_1) - \Lambda(A_2)}{\Lambda_0(A_1) - \Lambda_0(A_2)} = \frac{e^{z_1^{\mathrm{T}}\beta_0} - e^{z_2^{\mathrm{T}}\beta_0}}{e^{z_1^{\mathrm{T}}\beta} - e^{z_2^{\mathrm{T}}\beta}}. \tag{A.2}$$

Note that (A.2) implies the ratios on both sides are the same constant c. By Condition (C1), the left-hand side then gives $\Lambda(t) = c\Lambda_0(t)$ for t in the support of A. On the other hand, the right-hand side is degenerate if it equals c when z_1 and z_2 vary, this again implies $\beta = \beta_0$ thus $c = 1$.

Since $\log \mathcal{L}_n^P$ is always negative, by the similar arguments as in Zeng and Lin (2006), we can show that the PLAC estimator has finite jump sizes, and that $\hat{\Lambda}(\tau)$ is bounded almost surely when $n \to \infty$. Because $\ell_n^c(\beta, \Lambda)$ is maximized at the PLAC estimator $(\hat{\beta}, \hat{\Lambda})$ over the whole model, it is certainly maximized along the parametric sub-model $(\hat{\beta}, \Lambda_\varepsilon)$ at $\varepsilon = 0$. Thus by the regularity conditions, the PLAC estimator is the solution to the composite score equations $U_\beta(\beta, \Lambda) = 0$ and $U_\Lambda(\beta, \Lambda)(h) = 0$. Interchanging the summations and integrals in the second equation and rearranging the resulting terms, we have

$$\mathbb{P}_n \int_0^\tau h(s) dN(s) = \int_0^\tau h(s) \left\{ \mathbb{P}_n Y(s) e^{Z^{\mathrm{T}}\hat{\beta}} + \mathbb{U}_{n,2} \frac{R(\hat{\beta}, \hat{\Lambda})}{1 + R(\hat{\beta}, \hat{\Lambda})} Q^{(0)}(s; \hat{\beta}) \right\} d\hat{\Lambda}(s). \tag{A.3}$$

Let

$$M_n(s; \hat{\beta}, \hat{\Lambda}) = \mathbb{P}_n Y(s) e^{Z^{\mathrm{T}}\hat{\beta}} + \mathbb{U}_{n,2} \frac{R(\hat{\beta}, \hat{\Lambda})}{1 + R(\hat{\beta}, \hat{\Lambda})} Q^{(0)}(s; \hat{\beta})$$

denote the random function in the brackets. Replacing $h(s)$ with $h(s)/M_n(s; \hat{\beta}, \hat{\Lambda})$ on both sides of (A.3) yields the self-consistency solution of Λ:

$$\hat{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \hat{\beta}, \hat{\Lambda})}.$$

The rest of the proof follows closely to Murphy et al. (1997), yet the technical details are different due to the pairwise likelihood. Inspired by the form of $\hat{\Lambda}$, we define another random step function

$$\tilde{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \beta_0, \Lambda_0)}.$$

Let $M_0(s; \beta_0, \Lambda_0) = P_0\{Y(s)e^{Z^{\mathrm{T}}\beta_0}\}$. Since $\mathrm{E}_0\{U(\beta_0, \Lambda_0)\} = 0$ and $\mathrm{E}_0\{U^P(\beta_0, \Lambda_0)\} = 0$, the same algebra as we used to get $\hat{\Lambda}$ yields

$$\Lambda_0(t) = P_0 \int_0^t \frac{dN(s)}{M_0(s; \beta_0, \Lambda_0)}.$$

Under the regularity conditions (C2)-(C3), by Lemma 1, and the double expectation argument as we used in (A.1), $s \mapsto M_n(s; \beta_0, \Lambda_0)$ is uniformly bounded away from zero and infinity, and is of uniformly bounded variation when n is sufficiently large. Therefore, by the Glivenko-Cantelli theorem and Remark 1, we have

$$\|M_n(s; \beta_0, \Lambda_0) - M_0(s; \beta_0, \Lambda_0)\|_{L_\infty[0,\tau]} \xrightarrow{a.s.} 0$$

and

$$\left\| \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \beta_0, \Lambda_0)} - P_0 \int_0^t \frac{dN(s)}{M_n(s; \beta_0, \Lambda_0)} \right\|_{L_\infty[0,\tau]} \xrightarrow{a.s.} 0,$$

where $\|\cdot\|_{L_\infty[0,\tau]}$ is the supreme norm on $[0, \tau]$. These results combined with the dominated convergence theorem yield

$$\left\| \tilde{\Lambda}(t) - \Lambda_0(t) \right\|_{L_\infty[0,\tau]} \xrightarrow{a.s.} 0.$$

By the definition of the PLAC estimator, the composite log-likelihood evaluated at

$(\hat{\beta}, \hat{\Lambda})$ is greater than that evaluated at $(\beta_0, \tilde{\Lambda})$:

$$\mathbb{P}_n \int_0^\tau \left\{ \log \frac{\hat{\Lambda}}{\tilde{\Lambda}}\{s\} + Z^{\mathrm{T}}(\hat{\beta} - \beta_0) \right\} dN(s)$$
$$-\mathbb{P}_n \left\{ e^{Z^{\mathrm{T}}\hat{\beta}} \int_0^\tau Y(s) d\hat{\Lambda}(s) - e^{Z^{\mathrm{T}}\beta_0} \int_0^\tau Y(s) d\tilde{\Lambda}(s) \right\} - \mathbb{U}_{n,2} \log \frac{1 + R(\hat{\beta}, \hat{\Lambda})}{1 + R(\beta_0, \tilde{\Lambda})} \quad \geqslant \quad 0.$$

By assumption, β is in a compact set, and that $\hat{\Lambda}(t) \leqslant \hat{\Lambda}(\tau)$ is bounded for $t \in [0, \tau]$ with probability one. Thus, by the Bolzano–Weierstrass theorem and the Helly's selection lemma, for every subsequence of $(\hat{\beta}, \hat{\Lambda})$, we can find a further subsequence (still denoted as $(\hat{\beta}, \hat{\Lambda})$) along which $\hat{\beta} \to \beta^*$ for some β^* and $\hat{\Lambda}(t) \to \Lambda^*(t)$, $\forall t \in [0, \tau]$ for some monotone function Λ^* almost surely.

Note that $\hat{\Lambda}(t)$ is absolutely continuous with respect to $\tilde{\Lambda}(t)$. Let $\eta(t) = \lim_{n\to\infty} d\hat{\Lambda}/d\tilde{\Lambda}$ be a bounded measurable function, then $\Lambda^*(t) = \int_0^t \eta(s) d\Lambda_0(s)$ (Zeng and Lin, 2006). By (C1), $\Lambda^*(t)$ is absolutely continuous with respect to the Lebesgue measure and we denote its derivative as $\lambda^*(t)$. Thus we have the ratio $d\hat{\Lambda}/d\tilde{\Lambda}$ converges to $\eta(t) = \lambda^*(t)/\lambda_0(t)$. Again, by the Glivenko-Cantelli theorem, Lemma 1, Remark 1 and the dominant convergence theorem, the difference of the composite log-likelihoods converges to

$$P_0 \int_0^\tau \left\{ \log \frac{\lambda^*}{\lambda_0}(s) + Z^{\mathrm{T}}(\beta^* - \beta_0) \right\} dN(s) \tag{A.4}$$
$$-P_0 \left\{ e^{Z^{\mathrm{T}}\beta^*} \int_0^\tau Y(s) d\Lambda^*(s) - e^{Z^{\mathrm{T}}\beta_0} \int_0^\tau Y(s) d\Lambda_0(s) \right\} - P_0 \log \frac{1 + R(\beta^*, \Lambda^*)}{1 + R(\beta_0, \Lambda_0)} \quad \geqslant \quad 0.$$

The left-hand side of (A.4) is the composite Kullback–Leibler divergence of the composite likelihood associated with the density indexed by (β^*, Λ^*) from the composite likelihood associated with the true density indexed by (β_0, Λ_0) (Varin and Vidoni, 2005), which is the summation of

$$P_0 \int_0^\tau \left\{ \log \frac{\lambda^*}{\lambda_0}(s) + Z^{\mathrm{T}}(\beta^* - \beta_0) \right\} dN(s) - P_0 \left\{ e^{Z^{\mathrm{T}}\beta^*} \int_0^\tau Y(s) d\Lambda^*(s) - e^{Z^{\mathrm{T}}\beta_0} \int_0^\tau Y(s) d\Lambda_0(s) \right\}$$

and

$$-P_0 \log \frac{1 + R(\beta^*, \Lambda^*)}{1 + R(\beta_0, \Lambda_0)}.$$

Both terms are strictly negative unless the corresponding likelihoods are exactly the same. The former can be shown using Jasen's inequality on the density of (X, Δ) given (A, Z). For the later, we use the similar double expectation approach in (A.1) as follows. Note the inner expectation is taken on the Bernoulli distribution of a pair (A_1, A_2) conditional on their order statistics and the covariates. The inequity is from Jasen's inequality applied to the same binary probability mass functions, with the equality holds only when the probability mass functions under different parameters are the same with probability one.

$$\begin{aligned}
&- P_0 \log \frac{1 + R(\beta^*, \Lambda^*)}{1 + R(\beta_0, \Lambda_0)} \\
&= \mathrm{E}\left[\mathrm{E}\left\{\log \frac{1 + R(\beta_0, \Lambda_0)}{1 + R(\beta^*, \Lambda^*)}\right\}\right] \\
&\leqslant \mathrm{E}\left[\log \mathrm{E}\left\{\frac{1 + R(\beta_0, \Lambda_0)}{1 + R(\beta^*, \Lambda^*)}\right\}\right] \\
&= \mathrm{E}\left[\log\left\{\frac{1}{1 + R(\beta_0, \Lambda_0)} \cdot \frac{1 + R(\beta_0, \Lambda_0)}{1 + R(\beta^*, \Lambda^*)} + \frac{1}{1 + R^{-1}(\beta_0, \Lambda_0)} \cdot \frac{1 + R^{-1}(\beta_0, \Lambda_0)}{1 + R^{-1}(\beta^*, \Lambda^*)}\right\}\right] \\
&= \mathrm{E}\left[\log\left\{\frac{1}{1 + R(\beta^*, \Lambda^*)} + \frac{1}{1 + R^{-1}(\beta^*, \Lambda^*)}\right\}\right] \\
&= 0.
\end{aligned}$$

By the negativity of the Kullback–Leibler divergence and (A.4), we have that for a pair of subjects, the composite log-likelihood function $\ell^c_{1,2}(\beta^*, \Lambda^*) = \ell^c_{1,2}(\beta_0, \Lambda_0)$ with probably one under P_0, where the composite log-likelihood for a pair of subjects is defined by setting

$n = 2$ in the expression of $\ell_n^c(\beta, \Lambda)$, i.e.,

$$\begin{aligned}\ell_{1,2}^c(\beta, \Lambda | \mathcal{O}_1, \mathcal{O}_2) = &\sum_{i=1}^{2} \left[\Delta_i \{ \log \lambda(X_i) + Z_i^{\mathrm{T}} \beta \} - e^{Z_i^{\mathrm{T}} \beta} \int_0^\infty Y_i(t) \lambda(t) dt \right] \\ &- 2 \log \left[1 + \exp\{ (e^{Z_1^{\mathrm{T}} \beta} - e^{Z_2^{\mathrm{T}} \beta})(\Lambda(A_1) - \Lambda(A_2)) \} \right],\end{aligned} \tag{A.5}$$

where $\mathcal{O}_i = \{A_i, X_i, \Delta_i, Z_i\}$. Thus, by the Lemma 2 (identifiability) proved below, we have $\beta^* = \beta_0$ and $\Lambda^* = \Lambda_0$. Since every subsequence of $(\hat{\beta}, \hat{\Lambda})$ has a further subsequence converging to (β_0, Λ_0), we have the convergence of the entire sequence to the same limit. Finally, the uniform convergence of $\hat{\Lambda}(t)$ to $\Lambda_0(t)$ over $[0, \tau]$ follows from the continuity of Λ_0. □

Lemma 2. *Under Conditions (C1)-(C3), both β_0 and Λ_0 are identifiable. Specifically, if there exist parameters (β^*, Λ^*) such that Λ^* is absolutely continuous with respect to Λ_0, and $\ell_{1,2}^c(\beta^*, \Lambda^*) = \ell_{1,2}^c(\beta_0, \Lambda_0)$ with probability one under P_0, then we have $\beta^* = \beta_0$ and $\Lambda^* = \Lambda_0$, where $\ell_{1,2}^c$ is the composite log-likelihood function for a pair of subjects as defined in* (A.5).

Proof. Suppose the composite log-likelihood of (β^*, Λ^*) is the same as that of the true parameter (β_0, Λ_0). That is, from (A.5) we have

$$\begin{aligned}&\sum_{i=1}^{2} \left[\Delta_i \{ \log \lambda^*(X_i) + Z_i^{\mathrm{T}} \beta^* \} - e^{Z_i^{\mathrm{T}} \beta^*} \int_0^\infty Y_i(t) \lambda^*(t) dt \right] \\ &- 2 \log \left[1 + \exp\{ (e^{Z_1^{\mathrm{T}} \beta^*} - e^{Z_2^{\mathrm{T}} \beta^*})(\Lambda^*(A_1) - \Lambda^*(A_2)) \} \right] \\ = &\sum_{i=1}^{2} \left[\Delta_i \{ \log \lambda_0(X_i) + Z_i^{\mathrm{T}} \beta_0 \} - e^{Z_i^{\mathrm{T}} \beta_0} \int_0^\infty Y_i(t) \lambda_0(t) dt \right] \\ &- 2 \log \left[1 + \exp\{ (e^{Z_1^{\mathrm{T}} \beta_0} - e^{Z_2^{\mathrm{T}} \beta_0})(\Lambda_0(A_1) - \Lambda_0(A_2)) \} \right],\end{aligned} \tag{A.6}$$

almost surely. Similar to the argument in Zeng and Lin (2006), Condition (C3) implies that, with positive probability, there exists at least one subject with $(\Delta_1 = 1, X_1 = \tau, A_1 = a_1)$ when $A_1 + C_1 \geqslant \tau$, $\bar{Y}_1(\tau) = 1$, $N_1(\tau-) = 0$, and $N_1(\tau) = 1$. In addition, another subject

exists either with

1) $(\Delta_2 = 1, X_2 = \tau, A_2 = a_2)$, when $A_2 + C_2 \geqslant \tau$, $\bar{Y}_2(\tau) = 1$, $N_2(\tau-) = 0$, and $N_2(\tau) = 1$; or

2) $(\Delta_2 = 0, X_2 = \tau, A_2 = a_2)$, when $A_2 + C_2 \geqslant \tau$, $\bar{Y}_2(\tau) = 1$, $N_2(\tau) = 0$.

Note that the equality (A.6) holds for the both cases. By taking the difference between the equalities (A.6) under Case 1) and Case 2), we obtain

$$\log \lambda^*(\tau) + Z_2^{\mathrm{T}} \beta^* = \log \lambda_0(\tau) + Z_2^{\mathrm{T}} \beta_0,$$

which by Condition (C2) implies $\lambda^* = \lambda_0$ and $\beta^* = \beta_0$.

It is worth noting that $\Lambda_0(t)$ is not identifiable for $0 < t < w_1$ (Wang et al., 1993). However, since we assume that the minimum support of A^* contains or is close to zero, a w_1 close to zero is observable by Condition (C3); thus, the identifiability issue is less likely to occur. □

A.1.2 Asymptotic Normality

We first establish a lemma on the $\sqrt{n}$-uniform convergence rate and asymptotic normality of the log-generalized odds ratio. This is achieved by the projection of the U-process.

Lemma 3. *Under Conditions (C1)-(C3), the class of the log-generalized odds ratios*

$$\mathcal{R} = \{(\mathcal{O}_i, \mathcal{O}_j) \mapsto r_{ij}(\beta, \Lambda) : \mathcal{O}_i, \mathcal{O}_j \in \Omega, \beta \in B, \Lambda \in \mathcal{H}_\Lambda\},$$

where $r_{ij}(\beta, \Lambda) = (e^{Z_i^{\mathrm{T}}\beta} - e^{Z_j^{\mathrm{T}}\beta})(\Lambda(A_i) - \Lambda(A_j))$, *satisfies the uniform central limit theorem for U-processes:*

$$\sqrt{n}(\mathbb{U}_{n,2} r(\beta, \Lambda) - P_0^2 r(\beta, \Lambda)) \rightsquigarrow \mathbb{G}_r,$$

where $\mathbb{G}_r$ *is a tight mean-zero Gaussian process.*

Proof. To establish the weak convergence, we first show that

$$\left\|\mathbb{U}_{n,2}r(\beta,\Lambda) - P_0^2 r(\beta,\Lambda) - \hat{\mathbb{U}}_{n,2}r(\beta,\Lambda)\right\|_{\beta,\Lambda} = o_p(n^{-1/2}),$$

where

$$\hat{\mathbb{U}}_{n,2}r(\beta,\Lambda) = \sum_{i=1}^{n} \mathrm{E}\left(\mathbb{U}_{n,2}r(\beta,\Lambda) - P_0^2 r(\beta,\Lambda)|\mathcal{O}_i\right)$$

is the Hájek projection of $\mathbb{U}_{n,2}r(\beta,\Lambda) - P_0^2 r(\beta,\Lambda)$ (van der Vaart, 2000), and $\|\cdot\|_{\beta,\Lambda}$ is the supreme norm on the parameter space.

It can be verified that $P_0^2 r(\beta,\Lambda) = 2\mathrm{Cov}(e^{Z^{\mathrm{T}}\beta}, \Lambda(A))$. Moreover, since the pair $\mathcal{O}_i$ and $\mathcal{O}_j$ are i.i.d.,

$$\begin{aligned}
\mathrm{E}\left(\, r_{ij}(\beta,\Lambda) \,|\, \mathcal{O}_i\right) &= \mathrm{E}\{(e^{Z_i^{\mathrm{T}}\beta} - e^{Z_j^{\mathrm{T}}\beta})(\Lambda(A_i) - \Lambda(A_j)) \mid A_i, Z_i\} \\
&= e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i) - \Lambda(A_i)\mathrm{E}e^{Z_i^{\mathrm{T}}\beta} - e^{Z_i^{\mathrm{T}}\beta}\mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i)).
\end{aligned}$$

Thus we have

$$\begin{aligned}
\hat{\mathbb{U}}_{n,2}r(\beta,\Lambda) &= \sum_{i=1}^{n} \mathrm{E}\left\{ \binom{n}{2}^{-1} \sum_{j<k} r_{jk}(\beta,\Lambda) - P_0^2 r(\beta,\Lambda) \,\middle|\, \mathcal{O}_i \right\} \\
&= \frac{2}{n}\sum_{i=1}^{n} \left\{ e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i) - \Lambda(A_i)\mathrm{E}e^{Z_i^{\mathrm{T}}\beta} - e^{Z_i^{\mathrm{T}}\beta}\mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i)) \right\} \\
&\quad - 4\mathrm{Cov}(e^{Z^{\mathrm{T}}\beta}, \Lambda(A)).
\end{aligned}$$

Direct calculation gives

$$\tilde{\mathbb{U}}_{n,2} \equiv \mathbb{U}_{n,2}r(\beta,\Lambda) - P^2 r(\beta,\Lambda) - \hat{\mathbb{U}}_{n,2}r(\beta,\Lambda) = \frac{1}{\binom{n}{2}} \sum_{i<j} \tilde{\mathbb{U}}_{n,2}^{(i,j)}.$$

The summand of $\tilde{\mathbb{U}}_{n,2}$ is given by

$$\begin{aligned}
\tilde{\mathbb{U}}_{n,2}^{(i,j)} &= e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i) - e^{Z_j^{\mathrm{T}}\beta}\Lambda(A_i) - e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_j) + e^{Z_j^{\mathrm{T}}\beta}\Lambda(A_j) - 2\mathrm{Cov}(e^{Z^{\mathrm{T}}\beta}, \Lambda(A)) \\
&\quad - \left\{e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i) - \Lambda(A_i)\mathrm{E}e^{Z_i^{\mathrm{T}}\beta} - e^{Z_i^{\mathrm{T}}\beta}\mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{Z_i^{\mathrm{T}}\beta}\Lambda(A_i))\right\} \\
&\quad - \left\{e^{Z_j^{\mathrm{T}}\beta}\Lambda(A_j) - \Lambda(A_j)\mathrm{E}e^{Z_j^{\mathrm{T}}\beta} - e^{Z_j^{\mathrm{T}}\beta}\mathrm{E}\Lambda(A_j) + \mathrm{E}(e^{Z_j^{\mathrm{T}}\beta}\Lambda(A_j))\right\} \\
&\quad + 4\mathrm{Cov}(e^{Z^{\mathrm{T}}\beta}, \Lambda(A)) \\
&= -(e^{Z_i^{\mathrm{T}}\beta} - \mathrm{E}e^{Z_i^{\mathrm{T}}\beta})\left(\Lambda(A_j) - \mathrm{E}\Lambda(A_j)\right) - (e^{Z_j^{\mathrm{T}}\beta} - \mathrm{E}e^{Z_j^{\mathrm{T}}\beta})\left(\Lambda(A_i) - \mathrm{E}\Lambda(A_i)\right),
\end{aligned}$$

where the second equality holds by the definition of the covariance and the i.i.d. property of the observations. Therefore, we have

$$\begin{aligned}
\tilde{\mathbb{U}}_{n,2} &= -\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}(e^{Z_i^{\mathrm{T}}\beta} - \mathrm{E}e^{Z_i^{\mathrm{T}}\beta})\left(\Lambda(A_j) - \mathrm{E}\Lambda(A_j)\right) \\
&\asymp -2\cdot\frac{1}{n}\sum_{i=1}^{n}(e^{Z_i^{\mathrm{T}}\beta} - \mathrm{E}e^{Z_i^{\mathrm{T}}\beta})\cdot\frac{1}{n}\sum_{j=1}^{n}\left(\Lambda(A_j) - \mathrm{E}\Lambda(A_j)\right),
\end{aligned}$$

where $\asymp$ denotes asymptotically equivalent. Since both summations in the last line are empirical processes of Donsker classes, we have

$$\left\|\tilde{\mathbb{U}}_{n,2}\right\|_{\beta,\Lambda} \lesssim \left\|n^{-1/2}\mathbb{G}_n e^{Z^{\mathrm{T}}\beta}\right\|_{\beta}\cdot\left\|n^{-1/2}\mathbb{G}_n\Lambda\right\|_{\Lambda} = O_p(n^{-1/2})O_p(n^{-1/2}) = o_p(n^{-1/2}),$$

where $\lesssim$ means the inequality holds up to a multiplicative constant, and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$.

Therefore, $\mathbb{U}_{n,2}r(\beta,\Lambda) - P_0^2 r(\beta,\Lambda)$ is equivalent to its projection $\hat{\mathbb{U}}_{n,2}r(\beta,\Lambda)$ up to a term of $o_p(n^{-1/2})$. The weak convergence of $\hat{\mathbb{U}}_{n,2}r(\beta,\Lambda)$ can be established by the empirical process theory. Combining these results leads to the weak convergence of $\mathbb{U}_{n,2}r(\beta,\Lambda)$. □

Proof of Theorem 2. Let θ denote the parameters (β,Λ). We proceed by checking the four conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). Note that $\sqrt{n}U(\theta_0)$ can be decomposed into $\sqrt{n}U^C(\theta_0) + \sqrt{n}U^P(\theta_0)$. Following the martingale theory, the first term

converges weakly to a mean-zero Gaussian process $\mathbb{G}_{U^C}$, and the linear functional

$$\sqrt{n}\left\{b_1^{\mathrm{T}} U_\beta^C(\theta_0) + U_\Lambda^C(\theta_0)(h)\right\}$$

converges weakly to a mean-zero normal random variable with variance that can be consistently estimated by $b^{\mathrm{T}}\hat{V}^C b$, where b is defined as in Section 2.3. By Lemma 3, the preservation theorem of Lipschitz functions and Theorem 5.3.1 of De la Peña and Giné (1999), the second term also converges weakly to a mean-zero Gaussian process $\mathbb{G}_{U^P}$, and the linear functional

$$\sqrt{n}\left\{b_1^{\mathrm{T}} U_\beta^P(\theta_0) + U_\Lambda^P(\theta_0)(h)\right\}$$

converges weakly to a mean-zero normal random variable with variance that can be consistently estimated by $b^{\mathrm{T}}\hat{V}^P b$. Note also that given $\{(A_i, Z_i)\}_{i=1}^n$, $U^C(\theta_0)$ is a martingale, whereas $U^P(\theta_0)$ is a function of A_i and Z_i only, thus by double expectation

$$\begin{aligned}
\mathrm{E}_0\{U^C(\theta_0)\cdot U^P(\theta_0)\} &= \mathrm{E}_0\left\{\mathrm{E}_0\left(U^C(\theta_0)\middle|(A_i,Z_i), i=1,\ldots,n\right)\cdot U^P(\theta_0)\right\} \\
&= \mathrm{E}_0\{0\cdot U^P(\theta_0)\} = 0,
\end{aligned}$$

where $\cdot$ denotes the inner product of the underlying space. This indicates that $U^C(\theta_0)$ and $U^P(\theta_0)$ are asymptotically independent (van der Vaart and Wellner, 1996, Example 1.4.6) at θ_0. Therefore, $\sqrt{n}U(\theta_0)$ converges weakly to a mean-zero Gaussian process $\mathbb{G}_U$. In addition, $\sqrt{n}\left\{b_1^{\mathrm{T}} U_\beta(\theta_0) + U_\Lambda(\theta_0)(h)\right\}$ converges weakly to a mean-zero normal random variable with asymptotic variance that can be consistently estimated by $b^{\mathrm{T}}(\hat{V}^C + \hat{V}^P)\, b$. Thus, the two stochastic conditions are satisfied by the consistency of $\hat{\theta}$, Lemma 3 and Lemma 3.3.5 of van der Vaart and Wellner (1996). The fourth condition holds since $\hat{\theta}$ is a zero of $U(\theta)$ and that $u(\theta_0) \equiv \mathrm{E}_0 U(\theta_0) = 0$ by the arguments in the consistency proof.

To complete the proof, we only need to verify the Fréchet derivative of u at θ_0 exists and is continuously invertible. The Fréchet differentiability can be check directly. For the

continuous invertibility, note that $J \equiv -\partial u(\theta)/\partial\theta|_{\theta=\theta_0}$ can be decomposed into J^C and J^P corresponding to the conditional and the pairwise likelihoods. By (C4) and the classic Cox model results, J^C is continuously invertible. Thus, it suffices to show J^P is a compact operator and that J is one-to-one by the Fredholm theory.

Following Example 3.3.10 of van der Vaart and Wellner (1996), we find the derivate J^P has the form

$$\begin{pmatrix} \beta-\beta_0 \\ \Lambda-\Lambda_0 \end{pmatrix} \mapsto \begin{pmatrix} J^P_{\beta\beta} & J^P_{\beta\Lambda} \\ J^P_{\Lambda\beta} & J^P_{\Lambda\Lambda} \end{pmatrix} \begin{pmatrix} \beta-\beta_0 \\ \Lambda-\Lambda_0 \end{pmatrix},$$

where

$$\begin{aligned}
J^P_{\beta\beta}(\beta-\beta_0) &= P_0\left\{\frac{R_0(\int Q_0^{(1)}d\Lambda_0)(\int Q_0^{(1)}d\Lambda_0)^{\mathrm{T}}}{(1+R_0)^2}+\frac{R_0(\int Q_0^{(2)}d\Lambda_0)}{1+R_0}\right\}(\beta-\beta_0)\\
J^P_{\beta\Lambda}(\Lambda-\Lambda_0) &= P_0\left\{\frac{R_0(\int Q_0^{(1)}d\Lambda_0)\int Q_0^{(0)}d(\Lambda-\Lambda_0)}{(1+R_0)^2}\right\}\\
J^P_{\Lambda\beta}(\beta-\beta_0)h &= P_0\left\{\frac{R_0(\int Q_0^{(1)}d\Lambda_0)^{\mathrm{T}}\int Q_0^{(0)}hd\Lambda_0}{(1+R_0)^2}\right\}(\beta-\beta_0)\\
J^P_{\Lambda\Lambda}(\Lambda-\Lambda_0)h &= P_0\left\{\frac{R_0\int Q_0^{(0)}hd\Lambda_0\cdot\int Q_0^{(0)}hd(\Lambda-\Lambda_0)}{(1+R_0)^2}+\frac{R_0\int Q_0^{(0)}hd(\Lambda-\Lambda_0)}{1+R_0}\right\},
\end{aligned}$$

where the functions with subscript zero are evaluated at the true parameter θ_0. Note that for $J^P_{\beta\beta}$ and $J^P_{\Lambda\Lambda}$, the second terms in the brackets have expectation zero, by the similar double expectation argument as in (A.1). Since bounded linear operators with finite dimensional ranges are compact, we only need to show the compactness of $J^P_{\Lambda\beta}$ and $J^P_{\Lambda\Lambda}$. That is to say, for a sequence of functions h_n in the unit ball, $J^P_{\Lambda\beta}(\beta-\beta_0)h_n$ and $J^P_{\Lambda\Lambda}(\Lambda-\Lambda_0)h_n$ have convergent subsequences. In fact, by (C1)-(C2) and the bounded variation property of the functions involved, the convergent subsequences can be selected using the Helly's lemma. Therefore, the operator J^P is compact.

We now show J is one-to-one. For $(\gamma,h)\in\mathbb{R}^p\times BV[0,\tau]$, we need to show $J(\gamma,h)=0$ implies $\gamma=0$ and $h(t)=0$. Similar to the arguments in Zeng and Lin (2006), some algebra

gives

$$J(\gamma, h) = P_0 \left\{ \left(\gamma^{\mathrm{T}} \int_0^\tau Z(dN - Ye^{Z^{\mathrm{T}}\beta_0} d\Lambda_0) + \int_0^\tau h dN - \int_0^\tau Ye^{Z^{\mathrm{T}}\beta_0} h d\Lambda_0 \right)^2 + \frac{1}{R_0} \left\{ \frac{R_0}{1+R_0} \gamma^{\mathrm{T}} \int_0^\tau Q_0^{(1)} d\Lambda_0 + \frac{R_0}{1+R_0} \int_0^\tau Q_0^{(0)} h d\Lambda_0 \right)^2 \right\}.$$

Comparing the expressions of J^C and J^P with V^C and V^P, we note that although the second Bartlett equality for the pairwise likelihood does not hold (Varin et al., 2011), the non-negativity of quadratic functions and R_0 indicate that, with probability one, the conditional score along the path $(\beta_0 + \gamma, \Lambda_0 + \varepsilon \int h d\Lambda_0)$

$$\gamma^{\mathrm{T}} \int_0^\tau Z\{dN(s) - Y(s)e^{Z^{\mathrm{T}}\beta_0} d\Lambda_0(s)\} + \int_0^\tau h(s) dN(s) - \int_0^\tau Y(s)e^{Z^{\mathrm{T}}\beta_0} h(s) d\Lambda_0(s) = 0.$$

By (C1) and (C3), considering the case where $N(\tau) = 0$ and $A + C \geqslant \tau$ and the case where $N(t) = I(t \geqslant t_0)$ for some $t_0 \in [0, \tau]$ and $A + C \geqslant \tau$, we obtain two equalities. Taking difference, we have

$$\int_0^\tau \left(\gamma^{\mathrm{T}} Z + h(s)\right) e^{Z^{\mathrm{T}}\beta_0} d\Lambda_0(s) + \gamma^{\mathrm{T}} Z + h(t_0) = 0.$$

The only solution to the above equations is trivial, thus

$$\gamma^{\mathrm{T}} Z + h(t) = 0, \quad \forall t \in [0, \tau].$$

It follows from the identifiability condition (C2) that $\gamma = 0$ and $h(t) = 0$.

With all four conditions verified, by Theorem 3.3.1 of van der Vaart and Wellner (1996), we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow J^{-1} \mathbb{G}_U,$$

where $\mathbb{G}_U$ is a mean-zero Gaussian process. Since linear maps preserve the Gaussian prop-

erty, $\sqrt{n}(\hat{\theta}-\theta_0)$ also converges weakly to a mean-zero Gaussian process $J^{-1}\mathbb{G}_U$. In addition, the linear functional (7) converges weakly to a mean-zero Gaussian random variable with a variance estimator given by (8). The matrices $\hat{J}^C$ and $\hat{J}^P$ are given by

$$\begin{aligned}\hat{J}^C &= -\frac{1}{n}\sum_{i=1}^{n} \partial U_i^C(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})\big|_{\beta=\hat{\beta},\,\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}\,,\\ \hat{J}^P &= -\frac{1}{n(n-1)}\sum_{i\neq j} \partial U_{ij}^P(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})\big|_{\beta=\hat{\beta},\,\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}\,.\end{aligned}$$

The summands of the above matrices $\partial U_i^C(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})$ and $\partial U_{ij}^P(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})$ take the forms

$$-\begin{pmatrix} Z_i^{\otimes 2} e^{Z_i^{\mathrm{T}}\beta}\sum_{k=1}^m \lambda_k Y_i(w_k) & Z_i e^{Z_i^{\mathrm{T}}\beta} Y_i(w_1) & \cdots & Z_i e^{Z_i^{\mathrm{T}}\beta} Y_i(w_m)\\ Z_i^{\mathrm{T}} e^{Z_i^{\mathrm{T}}\beta} Y_i(w_1) & I(X_i=w_1)\Delta_i/\lambda_1^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ Z_i^{\mathrm{T}} e^{Z_i^{\mathrm{T}}\beta} Y_i(w_m) & 0 & \cdots & I(X_i=w_m)\Delta_i/\lambda_m^2 \end{pmatrix}$$

and

$$-R_{ij}\begin{pmatrix} \frac{\left(\Lambda(Q_{ij}^{(1)})\right)^{\otimes 2}}{(1+R_{ij})^2}+\frac{\Lambda(Q_{ij}^{(2)})}{(1+R_{ij})} & \frac{Q_{ij}^{(0)}(w_1)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2}+\frac{Q_{ij}^{(1)}(w_1)}{(1+R_{ij})} & \cdots & \frac{Q_{ij}^{(0)}(w_m)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2}+\frac{Q_{ij}^{(1)}(w_m)}{(1+R_{ij})}\\ \left\{\frac{Q_{ij}^{(0)}(w_1)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2}+\frac{Q_{ij}^{(1)}(w_1)}{(1+R_{ij})}\right\}^{\mathrm{T}} & \frac{\left(Q_{ij}^{(0)}(w_1)\right)^2}{(1+R_{ij})^2} & \cdots & \frac{Q_{ij}^{(0)}(w_1)Q_{ij}^{(0)}(w_m)}{(1+R_{ij})^2}\\ \vdots & \vdots & \ddots & \vdots\\ \left\{\frac{Q_{ij}^{(0)}(w_m)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2}+\frac{Q_{ij}^{(1)}(w_m)}{(1+R_{ij})}\right\}^{\mathrm{T}} & \frac{Q_{ij}^{(0)}(w_1)Q_{ij}^{(0)}(w_m)}{(1+R_{ij})^2} & \cdots & \frac{\left(Q_{ij}^{(0)}(w_m)\right)^2}{(1+R_{ij})^2} \end{pmatrix},$$

respectively, where

$$\Lambda(Q_{ij}^{(l)}) = \sum_{k=1}^{m}\lambda_k Q_{ij}^{(l)}(w_k), \quad l=1,2.$$

The consistency of variance estimator (8) follows from the Glivenkon-Cantelli theorem and Remark 1. □

A.2 Comparison of Variances: An Example

In this subsection, we calculate the variances of the conditional approach estimator and the PLAC estimator for a specific Cox model under a simple truncation scenario. Suppose there is only one binary covariate Z^* such that $\text{pr}(Z^* = 1) = p$, and the failure time distribution $T^*|Z^*$ follows the Cox model with constant baseline hazard $\lambda(t) = 1$, i.e., $\lambda(t|Z^*) = \exp(\beta Z^*)$. We assume there is no censoring. The truncation time A^* is assumed to be binary with $\text{pr}(A^* = 1) = \text{pr}(A^* = 0) = 1/2$. It means half of the target population were enrolled into the study immediately after their disease onset, whereas the other half would be delayed for one time unit. To make the explicit calculation simpler, we let $\beta = \log 2$.

It is worth noting that all expectations and probabilities should be calculated taking into consideration the *biased sampling* ($T^* \geqslant A^*$). From the independence between T^* and A^* given Z^*

$$\text{pr}(T^* \geqslant A^* \mid Z^* = 1) = \frac{1}{2}(1 + e^{-2});\ \text{pr}(T^* \geqslant A^* \mid Z^* = 0) = \frac{1}{2}(1 + e^{-1}).$$

by Bayes' rule,

$$\tilde{p} = \text{pr}(Z = 1) = \text{pr}(Z^* = 1 \mid T^* \geqslant A^*) = \frac{p(1 + e^{-2})}{1 + pe^{-2} + (1 - p)e^{-1}};$$

and

$$\text{pr}(A = 1 \mid Z = 1) = \text{pr}(A^* = 1 \mid T^* \geqslant A^*, Z^* = 1) = \frac{e^{-2}}{e^{-2} + 1}$$
$$\text{pr}(A = 1 \mid Z = 0) = \text{pr}(A^* = 1 \mid T^* \geqslant A^*, Z^* = 0) = \frac{e^{-1}}{e^{-1} + 1}.$$

Note that $A^* \perp Z^*$, whereas $A \not\perp Z$ (see Section 2.1).

For the conditional likelihood $\mathcal{L}_n^C$, under the simplified model, we have

$$\begin{aligned}
\ell_n^C &= \sum_{i=1}^{n} \{Z_i\beta - (T_i - A_i)e^{Z_i\beta}\}, \\
\frac{\partial \ell_n^C}{\partial \beta} &= \sum_{i=1}^{n} \{Z_i - (T_i - A_i)Z_i e^{Z_i\beta}\}, \\
\frac{\partial^2 \ell_n^C}{\partial \beta^2} &= -\sum_{i=1}^{n} \{(T_i - A_i)Z_i^2 e^{Z_i\beta}\}.
\end{aligned}$$

By the law of large numbers, $V^C = \mathrm{E}\{Z - (T - A)Ze^{Z\beta}\}^2$ and $J^C = \mathrm{E}\{(T - A)Z^2 e^{Z\beta}\}$. Let $V = T - A$, which is $T^* - A^* | T^* \geq A^*$. To calculate these expectations, we first obtain the distribution of $V|(A, Z)$. Using a variable transformation, we can show that the conditional distribution of V given (A, Z) follows an exponential distribution with mean $\exp(\beta Z)$. Next, by double expectations, we have

$$\begin{aligned}
V^C &= \mathrm{E}\{Z - (T - A)Ze^{Z\beta}\}^2 = \tilde{p}\mathrm{E}\{1 - 2(T - A)\}^2 = \tilde{p}\mathrm{E}\{1 - 2T^*\}^2 = \tilde{p}; \\
J^C &= \mathrm{E}\{(T - A)Z^2 e^{Z\beta}\} = \tilde{p}\mathrm{E}\{(T - A)e^{\beta}\} = \tilde{p} \times 2\mathrm{E}T^* = \tilde{p}.
\end{aligned}$$

For the pairwise likelihood $\mathcal{L}_n^P$, under the simplified model, we have

$$\begin{aligned}
\ell_n^P &= -\sum_{i<j} \log(1 + R_{ij}), \\
\frac{\partial \ell_n^P}{\partial \beta} &= -\sum_{i<j} \frac{R_{ij}}{1 + R_{ij}}(A_i - A_j)(Z_i e^{Z_i\beta} - Z_j e^{Z_j\beta}), \\
\frac{\partial^2 \ell_n^P}{\partial \beta^2} &= -\sum_{i<j} \frac{1}{(1 + R_{ij})^2} \Big\{ R_{ij}(A_i - A_j)^2 (Z_i e^{Z_i\beta} - Z_j e^{Z_j\beta})^2 \\
&\qquad + (1 + R_{ij})R_{ij}(A_i - A_j)(Z_i^2 e^{Z_i\beta} - Z_j^2 e^{Z_j\beta}) \Big\}.
\end{aligned}$$

By the U-statistic asymptotic properties and double expectation, we have

$$
\begin{aligned}
V^P &= 4\mathrm{E}\left\{\frac{R_{12}}{1+R_{12}}(A_1-A_2)(Z_1e^{Z_1\beta}-Z_2e^{Z_2\beta})\times\frac{R_{13}}{1+R_{13}}(A_1-A_3)(Z_1e^{Z_1\beta}-Z_3e^{Z_3\beta})\right\}\\
&= \mathrm{pr}(Z_1=1,Z_2=Z_3=0)\times 4\times\mathrm{E}\left\{2\frac{e^{A_1-A_2}}{1+e^{A_1-A_2}}(A_1-A_2)\times 2\frac{e^{A_1-A_3}}{1+e^{A_1-A_3}}(A_1-A_3)\right\}\\
&\quad+\mathrm{pr}(Z_1=0,Z_2=Z_3=1)\times 4\times\mathrm{E}\left\{2\frac{e^{A_2-A_1}}{1+e^{A_2-A_1}}(A_2-A_1)\times 2\frac{e^{A_3-A_1}}{1+e^{A_3-A_1}}(A_3-A_1)\right\}\\
&= \tilde{p}(1-\tilde{p})^2\times 4\times\left\{\frac{e^{-2}}{1+e^{-2}}\times\frac{1}{1+e^{-1}}\times\frac{1}{1+e^{-1}}\times\frac{e}{1+e}\times 2\times\frac{e}{1+e}\times 2\right.\\
&\qquad\left.+\frac{1}{1+e^{-2}}\times\frac{e^{-1}}{1+e^{-1}}\times\frac{e^{-1}}{1+e^{-1}}\times\frac{e^{-1}}{1+e^{-1}}\times(-2)\times\frac{e^{-1}}{1+e^{-1}}\times(-2)\right\}\\
&\quad+\tilde{p}^2(1-\tilde{p})\times 4\times\left\{\frac{e^{-1}}{1+e^{-1}}\times\frac{1}{1+e^{-2}}\times\frac{1}{1+e^{-2}}\times\frac{e^{-1}}{1+e^{-1}}\times 2\times\frac{e^{-1}}{1+e^{-1}}\times 2\right.\\
&\qquad\left.+\frac{1}{1+e^{-1}}\times\frac{e^{-2}}{1+e^{-2}}\times\frac{e^{-2}}{1+e^{-2}}\times\frac{e}{1+e}\times(-2)\times\frac{e}{1+e}\times(-2)\right\}\\
&= 16\tilde{p}(1-\tilde{p})^2\frac{e^2}{(1+e)^4}+16\tilde{p}^2(1-\tilde{p})\frac{e^3}{(1+e)^2(1+e^2)^2};\\
J^P &= \mathrm{E}\left\{\frac{1}{(1+R_{12})^2}\{R_{12}(A_1-A_2)^2(Z_1e^{Z_1\beta}-Z_2e^{Z_2\beta})^2+(1+R_{12})R_{12}(A_1-A_2)(Z_1^2e^{Z_1\beta}-Z_2^2e^{Z_2\beta})\}\right\}\\
&= \mathrm{pr}(Z_1=1,Z_2=0)\times\mathrm{E}\left\{4\frac{e^{A_1-A_2}}{(1+e^{A_1-A_2})^2}(A_1-A_2)^2+2\frac{e^{A_1-A_2}}{1+e^{A_1-A_2}}(A_1-A_2)\right\}\\
&\quad+\mathrm{pr}(Z_1=0,Z_2=Z_3=1)\times\mathrm{E}\left\{4\frac{e^{A_2-A_1}}{(1+e^{A_2-A_1})^2}(A_2-A_1)^2+2\frac{e^{A_2-A_1}}{1+e^{A_2-A_1}}(A_2-A_1)\right\}\\
&= 2\tilde{p}(1-\tilde{p})\left\{\frac{e^{-2}}{1+e^{-2}}\times\frac{1}{1+e^{-1}}\times\left[\frac{4e}{(1+e)^2}+\frac{2e}{1+e}\right]\right.\\
&\qquad\left.+\frac{1}{1+e^{-2}}\times\frac{e^{-1}}{1+e^{-1}}\times\left[\frac{4e^{-1}}{(1+e^{-1})^2}-\frac{2e^{-1}}{1+e^{-1}}\right]\right\}\\
&= \frac{8e^2}{(1+e^2)(1+e)^2}.
\end{aligned}
$$

The asymptotic variance for the conditional approach estimator is the inverse Fisher information of the conditional likelihood $(V^C)^{-1}=(J^C)^{-1}=1/\tilde{p}\approx 2.20$, whereas the variance for the PLAC estimator is the inverse Godambe information of the composite likelihood $(V^C+V^P)/(J^C+J^P)^2\approx 1.71$. Thus, in this simplified example, we will have about 30% improvement of estimation efficiency for the regression coefficient.

B Additional Simulation Results

This section contains some additional simulation results comparing the proposed method with the competitors.

Table B.1 displays the results under the same setup as in Section 3 with $n = 200$. Note that the SEs for PLAC in this table are about twice of those in Table 1 with $n = 800$.

Table B.1: Summary of simulation with $n = 200$ and various censoring rates. PC: censoring percentage; True: true values; Bias, SE, SEE and CP: empirical bias ($\times 10^3$), standard error ($\times 10^3$), standard error estimate ($\times 10^3$) and 95% coverage probability; RE: relative efficiency with respect to the conditional approach estimator (ratio of the mean squared errors). The estimate of $\hat{\Lambda}(t)$ is evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

PC			Conditional		LBML			PLAC				
		True	Bias	SE	Bias	SE	RE	Bias	SE	SEE	CP	RE
					Case 1: length-biased sampling							
50	$\hat{\beta}_1$	1	15	233	−58	160	1.87	17	184	177	94	1.60
	$\hat{\beta}_2$	1	3	211	−68	155	1.56	10	173	160	94	1.48
	$\hat{\Lambda}_{\tau_{30}}$	0.212	1	61	22	53	1.12	−1	56	54	93	1.18
	$\hat{\Lambda}_{\tau_{60}}$	0.546	5	121	56	101	1.10	0	112	111	94	1.18
80	$\hat{\beta}_1$	1	43	407	−88	198	3.57	48	257	238	95	2.46
	$\hat{\beta}_2$	1	22	360	−99	188	2.89	47	251	218	92	1.99
	$\hat{\Lambda}_{\tau_{30}}$	0.103	1	56	36	46	0.93	−4	47	42	87	1.41
	$\hat{\Lambda}_{\tau_{60}}$	0.329	4	136	60	94	1.50	−6	114	106	90	1.43
					Case 2: non-length-biased sampling							
50	$\hat{\beta}_1$	1	8	218	−251	141	0.57	17	186	182	96	1.37
	$\hat{\beta}_2$	1	7	224	−246	148	0.61	20	191	183	94	1.36
	$\hat{\Lambda}_{\tau_{30}}$	0.207	−2	59	103	70	0.22	−4	57	53	92	1.06
	$\hat{\Lambda}_{\tau_{60}}$	0.538	−7	91	232	108	0.13	−9	89	90	94	1.02
80	$\hat{\beta}_1$	1	43	403	−388	170	0.92	59	303	264	92	1.73
	$\hat{\beta}_2$	1	61	400	−382	162	0.95	71	296	265	92	1.76
	$\hat{\Lambda}_{\tau_{30}}$	0.099	−5	46	107	65	0.14	−7	44	41	85	1.07
	$\hat{\Lambda}_{\tau_{60}}$	0.270	−10	86	215	105	0.13	−12	83	80	90	1.06

We also investigated the transformation approach for the non-length-biased case (Case 2), and compare its performance with the proposed method (Huang and Qin, 2012). Specifically, we first transformed the survival and truncation times using the distribution function of exponential(1), and then used the EM algorithm proposed by Qin et al. (2011) on the transformed data. The simulation with $n = 400$ and censoring rates 20%, 50% and 80% are reported in Table B.2.

Table B.2: Summary of simulation using transformation approach suggested in Huang and Qin (2012). PC: censoring percentage; True: true values; Bias, SE, SEE and CP: empirical bias ($\times 10^3$), standard error ($\times 10^3$), standard error estimate ($\times 10^3$) and 95% coverage probability; RE: relative efficiency with respect to the conditional approach estimator (ratio of the mean squared errors). The estimate of $\hat{\Lambda}(t)$ is evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

PC			Conditional		LBML-trans			PLAC				
		True	Bias	SE	Bias	SE	RE	Bias	SE	SEE	CP	RE
20	$\hat{\beta}_1$	1	9	117	−10	103	1.29	7	105	104	95	1.25
	$\hat{\beta}_2$	1	10	117	−8	104	1.26	8	107	104	94	1.19
	$\hat{\Lambda}_{\tau_{30}}$	0.298	0	43	4	39	1.19	−1	42	42	94	1.04
	$\hat{\Lambda}_{\tau_{60}}$	0.764	3	70	9	64	1.16	2	68	69	95	1.04
50	$\hat{\beta}_1$	1	9	152	−51	122	1.33	14	129	129	95	1.38
	$\hat{\beta}_2$	1	7	148	−55	124	1.21	11	130	129	94	1.29
	$\hat{\Lambda}_{\tau_{30}}$	0.207	−2	39	13	37	0.99	−3	38	38	94	1.04
	$\hat{\Lambda}_{\tau_{60}}$	0.538	−1	68	25	64	0.97	−3	67	64	93	1.03
80	$\hat{\beta}_1$	1	3	260	−121	158	1.71	26	191	181	93	1.83
	$\hat{\beta}_2$	1	2	262	−127	160	1.65	22	194	182	95	1.82
	$\hat{\Lambda}_{\tau_{30}}$	0.099	−1	34	26	36	0.57	−2	33	31	90	1.05
	$\hat{\Lambda}_{\tau_{60}}$	0.270	−5	60	46	59	0.64	−7	59	58	92	1.03

Lastly, we studied the performance of $\hat{\Lambda}$ at the right tail. Specifically, we checked the performance at the 90%, 95%, 97.5%, and 99% percentiles of the observed survival times $X = \min(T, A+C)$. The censoring rate was set at 50%, and two sample sizes ($n = 200, 800$) were considered. The summary statistics and the median numbers of observed events after the corresponding percentiles are provided in Table B.3

Table B.3: Simulation results for $\Lambda(t)$ estimation at the right tail. The estimates of $\hat{\Lambda}(t)$ at $t \in \{\tau_{90}, \tau_{95}, \tau_{97.5}, \tau_{99}\}$, where τ_p denotes the p% percentile of the observed survival times X, are reported. Simulations under 50% of censoring and various sample sizes (n) were considered. True: true values; n_{event}: number of events (the median of 1000 datasets); Bias, SE, SEE and CP: empirical bias, standard error, standard error estimate and 95% coverage probability; RE: relative efficiency with respect to the conditional approach estimator (ratio of mean squared errors).

n		True	n_{event}	Conditional		LBML			PLAC				
				Bias	SE	Bias	SE	RE	Bias	SE	SEE	CP	RE
				Case 1: length-biased sampling									
200	$\hat{\Lambda}_{\tau_{90}}$	1.722	9	0.013	0.344	0.107	0.253	1.57	0.009	0.321	0.303	94	1.15
	$\hat{\Lambda}_{\tau_{95}}$	2.545	5	0.037	0.610	0.137	0.414	1.97	0.034	0.561	0.471	91	1.19
	$\hat{\Lambda}_{\tau_{97.5}}$	3.498	2	0.035	0.952	0.222	0.667	1.84	0.035	0.876	0.665	87	1.18
	$\hat{\Lambda}_{\tau_{99}}$	4.892	1	−0.240	1.427	0.261	1.029	1.86	−0.230	1.361	0.840	74	1.10
800	$\hat{\Lambda}_{\tau_{90}}$	1.722	17	−0.005	0.160	0.078	0.131	1.11	−0.004	0.148	0.153	95	1.17
	$\hat{\Lambda}_{\tau_{95}}$	2.545	11	−0.003	0.258	0.100	0.204	1.29	0.000	0.243	0.243	94	1.13
	$\hat{\Lambda}_{\tau_{97.5}}$	3.498	5	−0.021	0.435	0.076	0.308	1.89	−0.016	0.406	0.381	94	1.15
	$\hat{\Lambda}_{\tau_{99}}$	4.892	2	−0.045	0.804	0.104	0.573	1.91	−0.037	0.761	0.647	88	1.12
				Case 2: non-length-biased sampling									
200	$\hat{\Lambda}_{\tau_{90}}$	1.453	10	−0.007	0.221	0.546	0.227	0.14	−0.002	0.217	0.206	94	1.04
	$\hat{\Lambda}_{\tau_{95}}$	2.048	5	−0.003	0.350	0.735	0.328	0.19	0.009	0.336	0.315	93	1.08
	$\hat{\Lambda}_{\tau_{97.5}}$	2.704	3	0.008	0.578	1.013	0.514	0.26	0.033	0.555	0.470	91	1.08
	$\hat{\Lambda}_{\tau_{99}}$	3.673	1	0.028	1.032	1.544	0.934	0.33	0.065	0.962	0.749	88	1.15
800	$\hat{\Lambda}_{\tau_{90}}$	1.453	23	0.002	0.108	0.544	0.117	0.04	0.001	0.104	0.103	94	1.08
	$\hat{\Lambda}_{\tau_{95}}$	2.048	12	−0.002	0.168	0.749	0.170	0.05	−0.002	0.159	0.158	94	1.11
	$\hat{\Lambda}_{\tau_{97.5}}$	2.704	5	−0.000	0.261	0.969	0.243	0.07	0.002	0.244	0.241	94	1.14
	$\hat{\Lambda}_{\tau_{99}}$	3.673	2	0.013	0.468	1.314	0.411	0.12	0.018	0.434	0.406	92	1.16

C Additional Data Analysis Results

Graphical check of the uniform truncation assumption The significant test result in Section 4 for the uniform truncation assumption was confirmed by the graphical checking method proposed by Asgharian et al. (2006). When the assumption holds, the estimated survival functions of A and V should coincide. However, as shown in Figure C.1, the estimated survival curve of V is above that of A throughout, and the point-wise confidence intervals for A always stay beyond those of V. This means that in the RRI-CKD dataset, the uniform truncation assumption is violated.

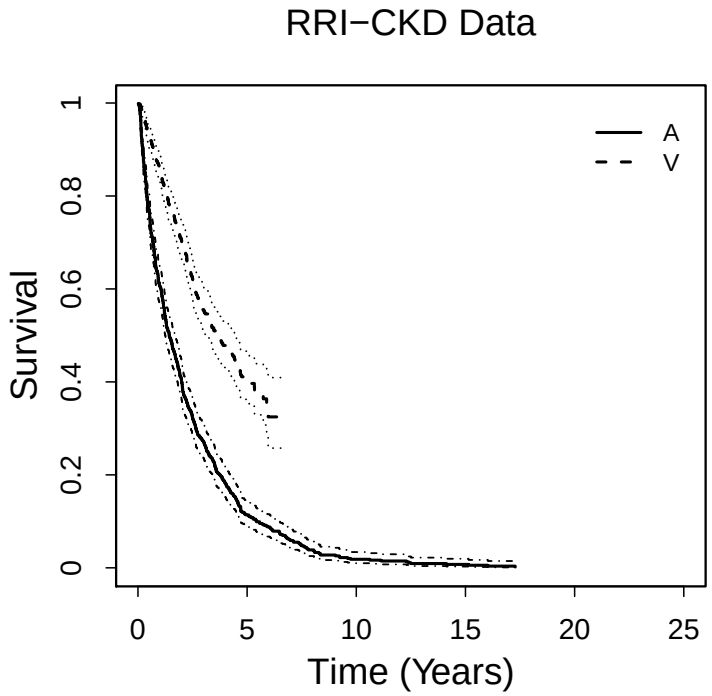

Figure C.1: Estimated survival curves for the truncation time A (solid) and the residual survival time V (dashed) of the RRI-CKD data. The 95% point-wise confidence intervals are shown as dashed or dotted lines around the estimates.

Regression coefficients estimates compared with the competitors A forest plot of the hazards ratios of the risk factors is shown in Figure C.2 to visualize the accuracy, precision and significance of the estimates. All coefficients have similar point estimates,

including the diabetes status and the CKD stage. However, the proposed estimator estimates all coefficients with improved precision indicated by narrower confidence intervals. The point estimates and confidence intervals using the full maximum likelihood estimator by Qin et al. (2011) are also displayed in the plot. We can see the point estimates for Male and CKD stage using the LBML estimator are quite different from those of the other two estimators.

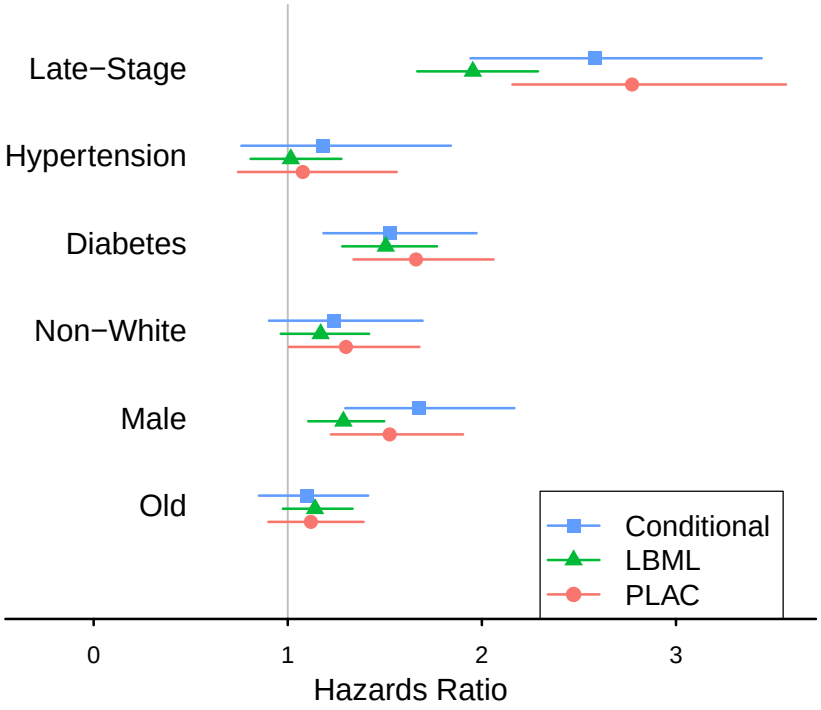

Figure C.2: Estimated hazards ratios of the covariates in the RRI-CKD data. The squares, triangles and dots represent the estimates using the conditional approach, the EM algorithm in Qin et al. (2011), and the proposed method (PLAC), respectively. The horizontal lines around the points represent the corresponding 95% confidence intervals.

Graphical check of the independence between A^* and Z^* We developed a graphical tool to check the independence assumption between the *underlying* truncation time A^* and the covariates Z^*. To be specific, we estimate the unbiased truncation distribution G for each level of the covariate under investigation. The estimation of G follows closely to the inverse-probability weighted estimator proposed in Wang (1991) and Huang and Qin

(2013). First, we calculate the right-truncation probability for observed A_i's using the survival probability the fitted Cox model with the conditional approach, and then use their reciprocals as the weights to estimate G using a weighted empirical cumulative distribution function.

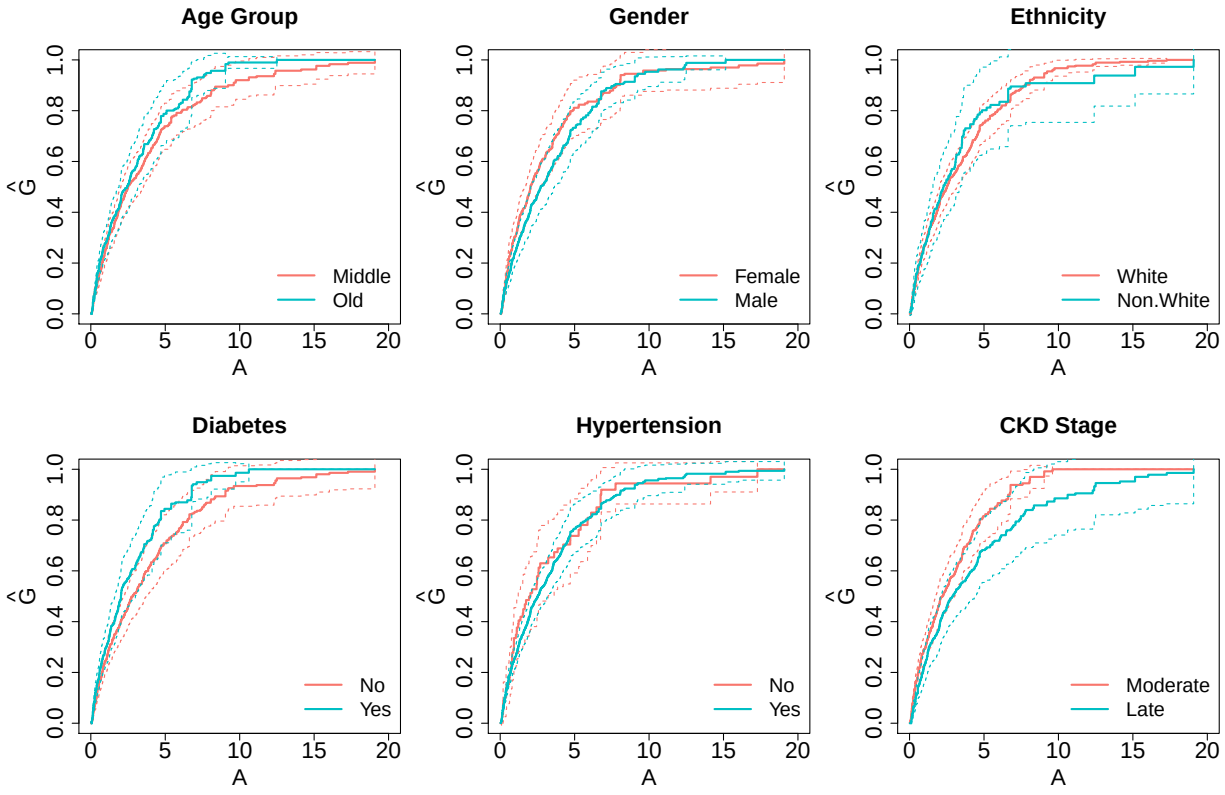

Figure C.3: Estimated $\hat{G}$ for each level of the covariates included in the RRI-CKD data analysis. The solid lines are the estimates, and the dashed lines in the same colors are the corresponding 95% point-wise confidence intervals.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Science & Business Media.

Asgharian, M., Wolfson, D. B., and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in Medicine*, 25(10):1751–1767.

De la Peña, V. and Giné, E. (1999). *Decoupling: from dependence to independence*. Springer, New York, NY.

Huang, C.-Y. and Qin, J. (2012). Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *Journal of the American Statistical Association*, 107(499):946–957.

Huang, C.-Y. and Qin, J. (2013). Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100(4):877–888.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, 15(2):780–799.

Pollard, D. (2002). *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press.

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and EM algorithms with length-biased data. *Journal of the American Statistical Association*, 106(496):1434–1449.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York, NY.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143.

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, 49(1):1–11.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.