Published in final edited form as:
Lifetime Data Anal. 2018 July ; 24(3): 509–531. doi:10.1007/s10985-017-9408-1.

Two-Sample Tests for Survival Data from Observational Studies

Chenxi Li
Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI 48824, U.S.A

Abstract

When observational data are used to compare treatment-specific survivals, regular two-sample tests, such as the log-rank test, need to be adjusted for the imbalance between treatments with respect to baseline covariate distributions. Besides, the standard assumption that survival time and censoring time are conditionally independent given the treatment, required for the regular two-sample tests, may not be realistic in observational studies. Moreover, treatment-specific hazards are often non-proportional, resulting in small power for the log-rank test. In this paper, we propose a set of adjusted weighted log-rank tests and their supremum versions by inverse probability of treatment and censoring weighting to compare treatment-specific survivals based on data from observational studies. These tests are proven to be asymptotically correct. Simulation studies show that with realistic sample sizes and censoring rates, the proposed tests have the desired Type I error probabilities and are more powerful than the adjusted log-rank test when the treatment-specific hazards differ in non-proportional ways. A real data example illustrates the practical utility of the new methods.

Keywords

Inverse Probability of Treatment Weighting; Inverse Probability of Censoring Weighting; Weighted Log-Rank Tests; Renyi-Type Tests

1 Introduction

Weighted log-rank tests (Fleming and Harrington 1991, Chapter 7) are the most popular statistical methods to compare survival/hazard function of a time-to-event outcome subject to right censoring between two treatment groups. However, when survival data are from observation studies, one cannot directly use the weighted log-rank tests to compare treatment-specific survivals because of the imbalance with respect to the distribution of baseline covariates between treatment groups. For example, as discussed in Zhang and Schaubel (2012), simultaneous pancreas-kidney (SPK) transplantation and kidney-alone (KA) transplantation are two treatment options for Type I diabetics with end-stage renal disease (ERSD). Receiving a pancreas in addition to a kidney is thought to have the potential to "cure" both ERSD and the diabetes, but the surgery is more complicated and could result in more post-operative complications, meaning that a patient may actually have shorter survival time if going through the simultaneous transplantation than the kidney-alone. So it is of interest to compare the post-operation survival between SPK and KA transplants. Because the decision to receive SPK or KA is related to patients' health conditions at transplant, which also affect the post-transplant survival, one cannot attribute the survival difference detected by applying weighted log-rank tests to survival data of SPK and KA recipients completely to the different transplant types.

Another common issue that invalidates weighted log-rank tests is that the censoring time is not conditionally independent of the survival time given the treatment. This often occurs in studies with long follow-ups, where there may be a considerable percentage of censoring due to loss to follow-up rather than administrative reasons. The subjects who drop out could differ from those who do not with respect to some baseline covariates such as age, socioeconomic status, marital status, and health condition even if they have the same treatment.

To cope with the non-randomness of treatment assignment and censoring that fails the weighted log-rank tests for observational studies, we propose adjusting the weighted log-rank tests by inverse probability of treatment and censoring weighting. The idea is to weight the at-risk and incremental counting processes in weighted log-rank statistics by inverse probabilities of treatment and censoring. This double inverse weighting idea was used by Schaubel and Wei (2011) to estimate cumulative treatment effect on time-to-event outcomes in the presence of confounders and dependent censoring. Here we focus on testing for the difference in treatment-specific hazards across arbitrary time windows. So the proposed methods are preferred over Schaubel and Wei (2011)'s when one is interested in comparing the possibly time-varying effects of two treatments across a specific time period rather than the cumulative effects from the time origin to certain time point. Admittedly, unlike Schaubel and Wei (2011), the proposed methods assume that the adjustment covariates for dependent censoring are time-independent, though this assumption can be easily relaxed to incorporate time-varying covariates with (almost) known trajectories (e.g., under frequent monitoring), as discussed in Section 5.

Xie and Liu (2005) developed an adjusted log-rank test with inverse probability of treatment weighting (IPTW) for group comparisons. We note that their variance formula for the IPTW log-rank statistic does not take the variability of the treatment probability estimates into account. It seems that this would underestimate the variance of the test statistic and lead to a liberal test. However, ignoring the variability of the treatment probability estimates amounts to calculating the variance of the IPTW log-rank statistic as if the treatment probabilities were known; the IPTW cumulative hazard estimators, which are the building blocks of the IPTW log-rank test, actually have larger variances from using known weights than estimated weights, because the influence function of the latter is the residual of projecting the influence function of the former onto the tangent space for the treatment assignment model (see Tsiatis 2006, Page 206). Thus, Xie and Liu (2005)'s variance formula in fact overestimates the variance of their IPTW log-rank statistic and leads to a conservative test. The conservativeness was substantiated by a numerical experiment in Section 3.2. In contrast, our adjusted weighted log-rank tests account for the variability in both the estimated probability of treatment and the estimated probability of censoring. Additionally, our tests allow a time weight and are applicable when censoring depends on time-independent covariates.

It is known that the log-rank test is the most powerful nonparametric test when the hazard functions to be compared are proportional to each other but has little power when the hazard functions cross. In the presence of nonrandom treatment assignment and conditional dependence of censoring on survival endpoint given treatment, one can follow the proof of that property to show that it is also true for our adjusted log-rank test (with weight function being a constant). In general, one has to know the type of the survival difference that may exist between groups to determine the weight function for the weighted log-rank test to be powerful, which is however impossible in many applied situations. To overcome this shortcoming, Gill (1980) proposed a supremum version of weighted log-rank test statistics, which he calls "Renyi-type" statistics, so that they can have robust power to detect various types of difference in hazard function between treatment groups. In this article, we propose an analogue of the Renyi-type test for survival data from observational studies.

The rest of the article is organized as follows. Section 2 formally develops the adjusted weighted log-rank tests and the adjusted Renyi-type tests, and proves that they are asymptotically correct. The section also presents an estimator of treatment-specific survival function and its asymptotic distribution based on the double inverse weighted estimator of treatment-specific cumulative hazard function in Schaubel and Wei (2011). Section 3 evaluates the finite sample performance of the proposed methods through simulations. The new methods are then applied to a data set from the Scientific Registry of Transplant Recipients (SRTR) in Section 4 to compare the post-operative survival between SPK and KA transplants. We conclude the article with some remarks in Section 5.

2 Methods

2.1 Data and Assumptions

We consider a sample of n independent subjects from an observation study. For each subject, let T and C denote the underlying failure and censoring times respectively. The time on study (observation time) is defined to be $U \equiv \min\{T, C\}$ with $\Delta \equiv \mathrm{I}(T \le C)$ being the failure event indicator. Let Z denote the treatment ($Z = 0$ or 1) the subject receives. In addition, we observe a set of baseline covariates $\mathbf{X}$ for every subject. Therefore, the observed data are n i.i.d. replicates of $(U, \Delta, Z, \mathbf{X})$ and denoted by $(U_i, \Delta_i, Z_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$, of which (U_i, Δ_i) can be alternatively represented by counting processes $N_i(t) = \mathrm{I}(U_i \le t, \Delta_i = 1)$ and at-risk processes $Y_i(t) = \mathrm{I}(U_i \ge t)$ $(i = 1, \ldots, n)$. We also define $Z_{ij} = \mathrm{I}(Z_i = j)$ $(j = 0, 1)$ for later use.

Our analysis aim is to compare the survival functions between the two treatment groups. Specifically, letting $T_{(j)}$ $(j = 0, 1)$ denote the two potential failure times of a subject randomly selected from the population under study if s/he received treatment $Z = j$, we want to test whether the two survival functions $S_{(j)}(t) \equiv P(T_{(j)} > t)$ $(j = 0, 1)$, or equivalently, the two hazard functions $\lambda_{(j)}(t) \equiv \lim_{\Delta t \to 0} P(t \le T_{(j)} < t + \Delta t | T \ge t)/\Delta t$ $(j = 0, 1)$ are identical. Since the data are from an observational study, we need to adjust for the imbalance between the two treatment groups with respect to the distribution of the confounders when carrying out the two sample comparison. For this purpose, we assume that all the confounders are captured in $\mathbf{X}$; i.e., $(T_{(0)}, T_{(1)}) \perp Z | \mathbf{X}$. As for censoring, one usually assumes that $T \perp C | Z$ when comparing treatment-specific survivals. This assumption is unrealistic in most observational studies. Hence, we consider a more realistic assumption that $T \perp C|(Z, \mathbf{X})$ in developing the proposed methods.

The building blocks for comparing treatment-specific survivals are the double inverse weighted at-risk processes and incremental counting processes:

$$Y_{ij}^*(t) = \frac{Y_{ij}(t)}{\hat{p}_{ij}\hat{S}_{ij}^C(t-)} \text{ and } dN_{ij}^*(t) = \frac{dN_{ij}(t)}{\hat{p}_{ij}\hat{S}_{ij}^C(t-)}, \quad i = 1, \dots, n, \quad j = 0, 1, \quad (1)$$

where $Y_{ij}(t) = Z_{ij}Y_i(t)$, $N_{ij}(t) = Z_{ij}N_i(t)$, $\hat{p}_{ij}$ is a consistent estimator of $p_{ij} \equiv P(Z_i = j|\mathbf{X}_i)$, and $\hat{S}_{ij}^C(t)$ is a consistent estimator of $\hat{S}_{ij}^C(t) \equiv P(C_i > t | Z_i = j, \mathbf{X}_i)$. Throughout the paper, we define the convention 0/0 = 0. The idea in (1) is the same as Schaubel and Wei (2011), i.e., using inverse probability of treatment weighting to balance the treatment-specific confounder distributions and inverse probability of censoring weighting to account for the dependent censoring. Therefore the valid inference on $S_{(j)}(t)$ $(j = 0, 1)$ depends on correctly modeling the effect of $\mathbf{X}$ on Z and that of $(Z, \mathbf{X})$ on C. We assume that the treatment assignment follows a logistic regression model,

$$\text{logit}\left\{P(Z_i = 1 | \mathbf{X}_i)\right\} = \boldsymbol{\alpha}^T\widetilde{\mathbf{X}}_i^Z, \quad (2)$$

where $\widetilde{\mathbf{X}}_i^Z$ is a vector consisting of an intercept and $\mathbf{X}_i^Z$, a vector made up of (possibly transformed) elements of $\mathbf{X}_i$ with a superscript Z indicating that the vector is related to treatment assignment. If model (2) is correct, the maximum likelihood estimator for $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}$, solving the estimating equation,

$$\sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^Z\left\{Z_i - \text{expit}(\boldsymbol{\alpha}^T\widetilde{\mathbf{X}}_i^Z)\right\} = 0, \quad (3)$$

consistently estimates the true parameter, and thus $\hat{p}_{ij}$ can be obtained by $p_{ij}(\hat{\boldsymbol{\alpha}}) \equiv \text{expit}\left\{(-1)^{j+1}\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i^Z\right\}$.

Regarding censoring, we assume a proportional hazards model for each treatment $Z = 0, 1$,

$$\lambda_{ij}^C(t) \equiv \lambda^C(t | Z_i = j, \mathbf{X}_i) = \lambda_{0j}^C(t)\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C), \quad j = 0, 1, \quad (4)$$

where $\lambda^C(t|Z_i = j, \mathbf{X}_i)$ is the conditional hazard of C_i given $Z_i = j$ and $\mathbf{X}_i$, $\lambda_{0j}^C(t)$ is an unspecified treatment-specific baseline hazard function, and $\mathbf{X}_i^C$ is a vector made up of (possibly transformed) elements of $\mathbf{X}_i$ with a superscript C indicating that the vector is related to censoring. If model (4) is correct, consistent estimators for $\boldsymbol{\theta}j$ and $\Lambda_{0j}^{C}(t) \equiv \int_{0}^{t} \lambda_{0j}^{C}(s)ds$ can be obtained by the maximum partial likelihood estimator and the Breslow estimator, respectively, denoted by $\hat{\boldsymbol{\theta}}_j$ and $\hat{\Lambda}_{0j}^{C}(t)$. A consistent estimator for the cumulative hazard $\Lambda_{ij}^{C}(t) \equiv \int_{0}^{t} \lambda_{ij}^{C}(s)ds$ would then be $\hat{\Lambda}_{ij}^{C}(t) \equiv \hat{\Lambda}_{0j}^{C}(t)\exp(\hat{\boldsymbol{\theta}}_j^T \mathbf{X}_i^C)$, and $\hat{S}_{ij}^{C}(t)$ can be obtained by $\exp\left\{-\hat{\Lambda}_{ij}^{C}(t)\right\}$. For notational convenience in deriving the asymptotic theory in the sequel, we set $\boldsymbol{\Lambda}^C \equiv \left\{\Lambda_{ij}^{C}: i = 1, \ldots, n, j = 0, 1\right\}$ and $\hat{\boldsymbol{\Lambda}}^C \equiv \left\{\hat{\Lambda}_{ij}^{C}: i = 1, \ldots, n, j = 0, 1\right\}$.

2.2 Double inverse weighted estimation of treatment-specific survival function

The first step in comparing treatment-specific survivals is usually to estimate the survival curves for a graphical representation. To consistently estimate the treatment-specific survival function $S_{(j)}(t)$ ($j = 0, 1$), we present a double inverse weighted estimator, which builds on the double inverse weighted estimator of treatment-specific cumulative hazard $\Lambda_{(j)}(t) \equiv \int_{0}^{t} \lambda_{(j)}(s)ds$ in Schaubel and Wei (2011). The latter estimator is

$$\hat{\Lambda}_{(j)}(t) \equiv \int_{0}^{t} \frac{\sum_{i=1}^{n} dN_{ij}^{*}(s)}{\sum_{i=1}^{n} Y_{ij}^{*}(s)}, \quad j = 0, 1\,. \qquad (5)$$

So a natural estimator for $S_{(j)}(t)$ is

$$\hat{S}_{(j)}(t) \equiv \exp(-\hat{\Lambda}_{(j)}(t)), \quad j = 0, 1\,. \qquad (6)$$

The uniform consistency of $\hat{\Lambda}_{(j)}(t)$ over an interval $[0, t_u]$, where t_u is chosen to avoid the instability of the estimator in the tail of the observation time distribution, has been proved by Schaubel and Wei (2011). The key is to show that $E[\left\{p_{ij}(\boldsymbol{\alpha})S_{ij}^{C}(s-)\right\}^{-1} dN_{ij}(s)] = dF_{(j)}(s) \equiv -dS_{(j)}(s)$ and $E[\left\{p_{ij}(\boldsymbol{\alpha})S_{ij}^{C}(S-)\right\}^{-1} Y_{ij}(s)] = S_{(j)}(s-)$ via successive conditioning. Schaubel and Wei (2011) also established that $\sqrt{n}(\hat{\Lambda}_{(j)}(t) - \Lambda_{(j)}(t))$ converges asymptotically to a zero-mean Gaussian process. We summarize the asymptotic properties of $\hat{\Lambda}_{(j)}(t)$ in Theorem 1 below. The proof can follow that of Theorem 1 in Schaubel and Wei (2011) and is thus omitted.

Theorem 1—Set $\pi_j(t) = P(Y_{ij}(t) > 0)$ and $\mathcal{T}_j = \left\{t: \pi_j(t) > 0\right\}$ ($j = 0, 1$). Under conditions (a) to (f) in Appendix A, for any $t_u \in \mathcal{T}_j$, $\hat{\Lambda}_{(j)}(t)$ converges almost surely and uniformly to $\Lambda_{(j)}(t)$ for $t \in [0, t_u]$, and $\sqrt{n}\left(\hat{\Lambda}_{(j)}(t) - \Lambda_{(j)}(t)\right)$ converges weakly to a zero-mean Gaussian process in $D[0, t_u]$, the space of cadlag functions on $[0, t_u]$, with covariance function $\sigma_{\Lambda_{(j)}}(s, t) = E\left\{\Phi_{ij}(s)\Phi_{ij}(t)\right\}$, where

$$\Phi_{ij}(t) = \Phi_{ij1}(t) + \Phi_{ij2}(t) + \Phi_{ij3}(t) + \Phi_{ij4}(t)$$

$$\Phi_{ij1}(t) = (-1)^j\left\{\int_0^t \frac{d\xi_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)} - \int_0^t \frac{G_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D_j^2(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dQ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\right\}^T V_Z^{-1}(\boldsymbol{\alpha})\psi_i^Z(\boldsymbol{\alpha})$$

$$\Phi_{ij2}(t) = \left\{\int_0^t \frac{dJ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)} - \int_0^t \frac{H_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D_j^2(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dQ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\right\}^T \left\{\Omega_j^C(\boldsymbol{\theta}_j)\right\}^{-1} U_{ij}^C(\boldsymbol{\theta}_j)$$

$$\Phi_{ij3}(t) = \int_0^t \left[\int_s^t \frac{d\zeta_j(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D_j(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)} - \frac{\gamma_j(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D_j^2(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}\right] \frac{dM_{ij}^C(s)}{r_{Cj}^{(0)}(s;\boldsymbol{\theta}_j)}$$

$$\Phi_{ij4}(t) = \int_0^t \frac{dM_{ij}^*(s)}{D_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)},$$

where $V_Z^{-1}(\boldsymbol{\alpha})\psi_i^Z(\boldsymbol{\alpha})$ and $\left\{\Omega_j^C(\boldsymbol{\theta}_j)\right\}^{-1} U_{ij}^C(\boldsymbol{\theta}_j)$ are the influence function for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\theta}}_j$, respectively (of which explicit expressions are given in Appendix A), $dM_{ij}^C(s) = dN_{ij}^C(s) - Y_{ij}(s)d\Lambda_{ij}^C(s)$ with $N_{ij}^C(s) = Z_{ij}I(U_i \le s, \Delta_i = 0)$, $dM_{ij}^*(s) = \left\{p_{ij}(\boldsymbol{\alpha})S_{ij}^C(s-)\right\}^{-1}\left\{dN_{ij}(s) - Y_{ij}(s)d\Lambda_{(j)}\right\}$, and $d\xi_j(s;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $D_j(s;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $G_j(s;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $dQ_j(s;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $dJ_j(s;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $H_j(s;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $d\zeta_j(u;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$, $\gamma_j(u;\, \boldsymbol{\alpha},\, \boldsymbol{\Lambda}^C)$ and $r_{Cj}^{(0)}(s;\boldsymbol{\theta}_j)$ are defined in Appendix A.

The uniform consistency of $\hat{S}_{(j)}(t)$ and the weak convergence of $\sqrt{n}\left(\hat{S}_{(j)}(t) - S_{(j)}(t)\right)$ are immediate from Theorem 1, the continuous mapping theorem and the functional delta method.

Corollary 1—Under conditions (a) to (g) in Appendix A, for any $t_u \in \mathscr{T}_j$, $\hat{S}_{(j)}(t)$, converges almost surely and uniformly to $S_{(j)}(t)$ for $t \in [0, t_u]$ and $\sqrt{n}\left(\hat{S}_{(j)}(t) - S_{(j)}(t)\right)$ converges weakly to a zero-mean Gaussian process in $D[0, t_u]$ with covariance function $\sigma_{S_{(j)}}(s, t) = S_{(j)}(s)S_{(j)}(t)E\left\{\Phi_{ij}(s)\Phi_{ij}(t)\right\}$.

The covariance function $\sigma_{S_{(j)}}(s,t)$ can be consistently estimated by $\widehat{S}_{(j)}(s)\widehat{S}_{(j)}(t)n^{-1}\sum_{i=1}^{n}\widehat{\Phi}_{ij}(s)\widehat{\Phi}_{ij}(t)$ where $\widehat{\Phi}_{ij}$ is obtained by replacing limiting terms in Φ_{ij} with their empirical counterparts.

2.3 Adjusted weighted log-rank tests

The adjusted weighted log-rank tests and their supremum versions build on the following stochastic process of cumulative weighted difference between the estimated treatment-specific hazards,

$$W_a(t;\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\Lambda}}^C) \equiv \int_0^t K^*(s;\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\Lambda}}^C)\left\{\frac{d\bar{N}^*_{\cdot 1}(s)}{\bar{Y}^*_{\cdot 1}(s)} - \frac{d\bar{N}^*_{\cdot 0}(s)}{\bar{Y}^*_{\cdot 0}(s)},\right\} \quad (7)$$

where $d\bar{N}^*_{11}(s) = \sum_{i=1}^{n} dN^*_{ij}(s)$, and $\bar{Y}^*_{11}(s) = \sum_{i=1}^{n} Y^*_{ij}(s)$ (j = 0, 1), and

$$K^*(s;\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\Lambda}}^C) = \frac{1}{\sqrt{n}}W(s)\frac{\bar{Y}^*_{\cdot 1}(s)\bar{Y}^*_{\cdot 0}(s)}{\bar{Y}^*_{\cdot 1}(s) + \bar{Y}^*_{\cdot 0}(s)}, \quad (8)$$

where $W(\cdot)$ is a nonnegative, bounded, and predictable process. In the later simulation and real data analysis, we consider the class of weight functions proposed by Fleming and Harrington (1981):

$$W(s) = \left\{\widehat{S}(s-)\right\}^{\rho}\left\{1-\widehat{S}(s-)\right\}^{\gamma}, \;\; \rho \ge 0, \, \gamma \ge 0,$$

where $\hat{S}(s)$ is an estimator of the overall survival function defined by

$$\widehat{S}(s) = \exp\left[-\int_0^s \frac{\sum_{i=1}^{n}\sum_{j=0}^{1}\left\{\widehat{S}^C_{ij}(u-)\right\}^{-1} dN_{ij}(u)}{\sum_{i=1}^{n}\sum_{j=0}^{1}\left\{\widehat{S}^C_{ij}(u-)\right\}^{-1} Y_{ij}(u)}\right].$$

$W_a(t;\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\Lambda}}^C)$ with (ρ, γ) = (0, 0) and (ρ, γ) = (1, 0) correspond to the adjusted log-rank and Prentice-Wilcoxon (Prentice 1978) statistics respectively.

Theorem 2 below gives the large sample representation of the "adjusted weighted log-rank statistic" defined in (7), which will be used to construct the adjusted weighted log-rank tests and their supremum versions.

Theorem 2—Set $\mathscr{T} = \left\{t: \pi_0(t)\pi_1(t) > 0\right\}$ and $t_{\sup} = \sup\mathscr{T}$. Under conditions (a)–(f), (h) and (i) in Appendix A,

$$W_a(t;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Lambda}}^C) = \int_0^t K^*(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\left\{\lambda_{(1)}(s) - \lambda_{(0)}(s)\right\}ds + n^{-1/2}\sum_{i=1}^n \phi_i(t) + o_p(1), \quad (9)$$

where $o_p(1)$ represents a term that converges in probability to zero in $D[0, t_{\text{sup}}]$ equipped with the uniform norm and

$$\phi_i(t) = A_{i1}(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) - A_{i0}(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) + \left\{B_1(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) - B_0(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\right\}^T V_Z^{-1}(\boldsymbol{\alpha})\psi_i^Z(\boldsymbol{\alpha}) + F_0^T(t;\boldsymbol{\alpha},$$
$$\boldsymbol{\Lambda}^C)\left\{\Omega_0^C(\boldsymbol{\theta}_0)\right\}^{-1}U_{i0}^C(\boldsymbol{\theta}_0) - F_1^T(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\left\{\Omega_1^C(\boldsymbol{\theta}_1)\right\}^{-1}U_{i1}^C(\boldsymbol{\theta}_1) + \int_0^t L_0(s,t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\frac{dM_{i0}^C(s)}{r_{C0}^{(0)}(s;\boldsymbol{\theta}_0)}$$
$$- \int_0^t L_1(s,t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\frac{dM_{i1}^C(s)}{r_{C1}^{(0)}(s;\boldsymbol{\theta}_1)},$$

(10)

where $A_{ij}(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = \int_0^t w(s)D_{1-j}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)D^{-1}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)dM_{ij}^*(s)$ with mean zero, and $w(s)$, $D(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)$, $B_j(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)$, $F_j(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)$ and $L_j(s,t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)$ ($j = 0,1$) are defined in Appendix A.

The proof of Theorem 2 is given in Appendix A. The idea of the proof is to decompose $W_a(t;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Lambda}}^C)$ into

$$W_a(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) + \left\{W_a(t;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Lambda}}^C) - W_a(t;\boldsymbol{\alpha},\hat{\boldsymbol{\Lambda}}^C)\right\} + \left\{W_a(t;\boldsymbol{\alpha},\hat{\boldsymbol{\Lambda}}^C) - W_a(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\right\}, \quad (11)$$

obtain the large sample representations of $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ and $\sqrt{n}\left\{\hat{\Lambda}_{ij}^C(\cdot) - \Lambda_{ij}^C(\cdot)\right\}$

using standard maximum likelihood theory and standard partial likelihood theory (Fleming and Harrington 1991) respectively, and finally apply the (functional) delta methods to the second and third summands in (11).

The adjusted weighted log-rank test for

$$H_0: \lambda_{(0)}(t) = \lambda_{(1)}(t) \text{ for all } t \in \mathscr{I}, \quad (12)$$

is based on the test statistic, $Z_{aw}(U^\dagger) \equiv W_a(U^\dagger;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Lambda}}^C)/\hat{\sigma}_W(U^\dagger)$, where $U^\dagger = \sup\left\{t: \bar{Y}_{.0}(t) \wedge \bar{Y}_{.1}(t) > 0\right\}$ and $\hat{\sigma}_W^2(t) = n^{-1}\sum_{i=1}^n \hat{\phi}_i^2(t)$ where $\hat{\phi}_i$ is obtained by replacing limiting terms in ϕ_i with their empirical counterparts. In the sequel, we just study the properties (significance level and power) of the two-sided test based on $Z_{aw}(U^\dagger)$, i.e., rejecting H_0 if $|Z_{aw}(U^\dagger)| \geq z_{\alpha/2}$ where $\alpha \in (0, 1)$ and $z_{\alpha/2}$ is the $\alpha/2$-th upper quantile of the standard normal distribution. One-sided tests' properties can be studied analogously. Theorem 2 implies that under H_0, $W_a(U^\dagger; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Lambda}}^C)$ can be written asymptotically as a sum of independent and identically distributed zero-mean variates, $\phi_i(t_{\text{sup}})$ ($i = 1, \ldots, \text{n}$). Thus the central limit theorem can be used to show that the adjusted weighted log-rank test based on $Z_{aw}(U^\dagger)$ has asymptotic significance level α. From (9) and (8), it is easy to see that the power of the test based on $Z_{aw}(U^\dagger)$ converges to one as $n \to \infty$, i.e., the test is consistent, under an alternative

$$H_A: \int_0^{t_{\text{sup}}} w(s) \frac{D_0(s; \boldsymbol{\alpha}, \boldsymbol{\Lambda}^C) D_1(s; \boldsymbol{\alpha}, \boldsymbol{\Lambda}^C)}{D_0(s; \boldsymbol{\alpha}, \boldsymbol{\Lambda}^C) + D_1(s; \boldsymbol{\alpha}, \boldsymbol{\Lambda}^C)} \left\{ \lambda_{(1)}(s) - \lambda_{(0)}(s) \right\} ds \neq 0 . \quad (13)$$

For any subset $\mathscr{T} \subset \mathscr{I}$, one can construct an adjusted weighted log-rank test for

$$H_0^{\mathscr{T}} : \lambda_{(0)}(t) = \lambda_{(1)}(t) \text{ for all } t \in \mathscr{T} \quad (14)$$

by choosing a weight function $W(t)$ such that $W(t) = 0$ for any $t \notin \mathscr{T}$. This test is also asymptotically correct according to Theorem 2.

2.4 Adjusted Renyi-type tests

Following the definition of the Renyi-type test (Gill 1980), the adjusted Renyi-type test statistic is defined to be

$$Z_R(U^\dagger) = \frac{\sup_{0 \leq t \leq U^\dagger} |W_a(t; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Lambda}}^C)|}{\hat{\sigma}_W(U^\dagger)} . \quad (15)$$

The asymptotic distribution of $Z_R(U^\dagger)$ under H_0 is not easy to derive. However, by the multiplier central limit theorem (sec. 2.9, van der Vaart and Wellner 1996), when n is large enough, the distribution of $n^{-1/2} \sum_{i=1}^{n} \phi_i(t)$ can be approximated by the conditional distribution of $n^{-1/2} \sum_{i=1}^{n} \omega_i \hat{\phi}_i(t)$ given $(U_i, \Delta_i, Z_i, \mathbf{X}_i)$ ($i = 1, \ldots, n$) where ω_i's are independent standard normal random variables. Therefore, the asymptotic null distribution of of $Z_R(U^\dagger)$ can be obtained by Monte Carlo simulation. The implementation procedure of the adjusted Renyi-type test is as follows.

1. Compute $Z_R(U^\dagger)$ based on the data.

2. Generate m independent sets of independent standard random variables $\left\{\omega_i^{(l)}\right\}_{i=1}^{n}(l = 1, \ldots, m)$ and compute the corresponding $\widetilde{Z}_R^{(l)}(U^{\dagger}) \equiv \sup_{0 \le t \le U^{\dagger}} |n^{-1/2}\sum_{i=1}^{n}\omega_i^{(l)}\hat{\phi}_i(t)|/\hat{\sigma}_W(U^{\dagger})$.
3. Compute the simulation-based p-value: $\widetilde{p}_R \equiv m^{-1}\sum_{l=1}^{m} \mathrm{I}\left\{\widetilde{Z}_R^{(l)}(U^{\dagger}) \ge Z_R(U^{\dagger})\right\}$
4. If $\widetilde{p}_R < \alpha$, reject H_0 at level α.

Theorem 2, the multiplier central limit theorem, and the law of large number together imply that this adjusted Renyi-type test has asymptotic significance level α.

In a similar way to the adjusted weighted log-rank test for $H_0^{\mathcal{T}}$, one can construct an adjusted Renyi-type test for $H_0^{\mathcal{T}}$ by choosing a weight function $W(t)$ such that $W(t) = 0$ for any $t \notin \mathcal{T}$ and then following the above test procedure.

3 Simulation

3.1 Performance evaluation of the proposed tests

We investigated the Type I error rates and powers of the proposed tests in finite samples through Monte Carlo simulations, which were performed in SAS 9.3. The number of simulation runs is 1000. The simulation scenarios are similar to those in Zhang and Schaubel (2012) and are described in detail below.

For each simulated data set, we first generated three baseline covariates X_1, X_2 and X_3 each from a standard normal distribution truncated at −0.5 and 0.5 in order to be consistent with the regularity conditions in Appendix A. The correlation between untruncated X_1 and X_3 is 0.2, and all other pairwise correlations equal 0. The treatment indicator Z was then generated from Bernoulli with probability of being one equal to expit($-0.5X_1$ $-0.5X_2$), which results in about 50% of the subjects in a simulated data set receiving treatment 1. The survival time T and censoring time C were generated under two sets of scenarios. The first set of scenarios are:

$$
\begin{aligned}
\textbf{Null:}\quad & \lambda(t|Z=j, \mathbf{X}) = \exp(-3 - X_1 - 0.9X_2 - X_3)\ (j=0,1),\\
& \lambda^C(t|Z=0, \mathbf{X}) = \exp(-5 + X_1 + 1.2X_2),\\
& \lambda^C(t|Z=1, \mathbf{X}) = \exp(-4.5 - 0.2X_1 - 0.7X_2);\\
\textbf{Nearly Proportional:}\quad & \lambda(t|Z=0, \mathbf{X}) = \exp(-2.8 - 0.95X_1 - 0.85X_2 - 0.95X_3),\\
& \lambda(t|Z=1, \mathbf{X}) = \exp(-3 - X_1 - 0.9X_2 - X_3),\\
& \lambda^C(t|Z=0, \mathbf{X}) = \exp(-5 + X_1 + 1.2X_2),\\
& \lambda^C(t|Z=1, \mathbf{X}) = \exp(-4.5 - 0.2X_1 - 0.7X_2);\\
\textbf{Early Departure:}\quad & \lambda(t|Z=0, \mathbf{X}) = \exp(0.5 - 1.5X_1 - X_2 - 0.7X_3)/(1+t)+0.1,\\
& \lambda(t|Z=1, \mathbf{X}) = \exp(-X_1 - 0.9X_2 - X_3)/(1+t) + 0.1,\\
& \lambda^C(t|Z=0, \mathbf{X}) = \exp(-1.9 + X_1 + 1.2X_2),\\
& \lambda^C(t|Z=1, \mathbf{X}) = \exp(-1.9 - 0.2X_1 - 0.7X_2);\\
\textbf{Crossing:}\quad & \lambda(t|Z=0, \mathbf{X}) = 0.21\exp(-0.1X_1 - 0.1X_2 - 0.1X_3)+0.001t,
\end{aligned}
$$

$$\begin{aligned}
&\lambda(t|Z=1,\mathbf{X}) = 0.1\exp(-0.2X_1 - 0.05X_2 - 0.1X_3) + 0.04t,\\
&\lambda^C(t|Z=0,\mathbf{X}) = \exp(-3.1 + X_1 + 1.2X_2),\\
&\lambda^C(t|Z=1,\mathbf{X}) = \exp(-3.1 - 0.2X_1 - 0.7X_2),
\end{aligned}$$

where $\lambda(t|Z=j,\mathbf{X})$ denotes the conditional hazard of T given $Z=j$ and the three baseline covariates $\mathbf{X}$. In the **Null** scenario, it is obvious that $\lambda_{(0)}() = \lambda_{(1)}(\cdot)$. In the **Nearly Proportional** scenario, $\lambda_{(1)}(t)/\lambda_{(0)}(t) \approx 0.85$ for most t's. In the **Early Departure**, $\lambda_{(0)}(t)$ and $\lambda_{(1)}(t)$ have a substantial difference when t is small, but the difference converges to zero as $t \to \infty$. The curves of $\lambda_{(0)}(t)$ and $\lambda_{(1)}(t)$ cross each other in the **Crossing** scenario. Figure 1 shows the plots of $\lambda_{(1)}(t)$ and $\lambda_{(0)}(t)$ against t in the **Nearly Proportional**, **Early Departure** and **Crossing** scenarios. The dependent censoring rate is around 19% in every scenario. The second set of scenarios under which T and C were generated just have different censoring distributions than the first set so that each dependent censoring rate is increased to 40%. The four scenarios are named **Null HC**, **Nearly Proportion HC**, **Early Departure HC** and **Crossing HC** respectively with **HC** meaning heavier censoring. The censoring distributions in these four scenarios are:

$$\begin{array}{rl}
\textbf{Null HC:} & \lambda^C(t|Z=0,\mathbf{X}) = \exp(-3.8 + X_1 + 1.2X_2),\\
 & \lambda^C(t|Z=1,\mathbf{X}) = \exp(-3.3 - 0.2X_1 - 0.7X_2);\\
\textbf{Nearly Proportional HC} & \lambda^C(t|Z=0,\mathbf{X}) = \exp(-3.7 + X_1 + 1.2X_2),\\
 & \lambda^C(t|Z=1,\mathbf{X}) = \exp(-3.2 - 0.2X_1 - 0.7X_2);\\
\textbf{Early Departure HC:} & \lambda^C(t|Z=0,\mathbf{X}) = \exp(-0.7 + X_1 + 1.2X_2),\\
 & \lambda^C(t|Z=1,\mathbf{X}) = \exp(-0.7 - 0.2X_1 - 0.7X_2);\\
\textbf{Crossing HC:} & \lambda^C(t|Z=0,\mathbf{X}) = \exp(-2.1 + X_1 + 1.2X_2),\\
 & \lambda^C(t|Z=1,\mathbf{X}) = \exp(-2.1 - 0.2X_1 - 0.7X_2).
\end{array}$$

Table 1 shows the simulation results for the adjusted weighted log-rank tests with $(\rho, \gamma) = (0, 0)$ and $(\rho, \gamma) = (1, 0)$ (i.e., the adjusted log-rank and Prentice-Wilcoxon tests) as well as the adjusted Renyi-type test with $(\rho, \gamma) = (0, 0)$ and $m = 2000$ in the scenarios where the dependent censoring rate is around 19%. The significance level α was set at 0.05 for all the tests. From the table, one can see that all the three tests have nominal Type I error rates. Under the alternative hypotheses, the powers of the tests increase with the sample size. The adjusted log-rank test is a little more powerful than the adjusted Prentice-Wilcoxon test in the **Nearly Proportional** scenario, but less powerful under the **Early Departure** and the **Crossing**. The adjusted Renyi-type log-rank test has comparable power to the better one of the adjusted log-rank and Prentice-Wilcoxon tests in the **Nearly Proportional** and **Early Departure** scenarios, and it is much more powerful than the other two when the two treatment-specific hazard functions cross. All these results meet the theoretical properties of the tests.

Table 2 shows the simulation results for the above three tests in the scenarios where the dependent censoring rate is around 40%. One can see that the Type I error probabilities of the adjusted log-rank test and its supremum version are a bit higher than the nominal level. We are puzzled by this little inflation. Further investigation is needed. The adjusted Prentice-Wilcoxon test still has the nominal Type I error probability. In terms of power, all the tests' powers are decreased due to the increase of censoring rate. The adjusted Prentice-Wilcoxon test is still the most powerful in the **Early Departure HC** scenario. Although it also has a greater power than the other two in the **Nearly Proportional HC** except when $n = 300$ and in the **Crossing HC**, these simulation results should be explained with caution, because the rise of censoring rate shortens the hazard comparison time window in finite samples. As the time window gets shorter, our scenario of crossing hazards becomes more and more like a scenario of early departure, so does our scenario of nearly proportional hazards. Table 2 also shows that the adjusted Renyi-type log-rank test has comparable power to the better one of the adjusted log-rank and Prentice-Wilcoxon tests under every alternative.

3.2 The adjusted log-rank test versus the IPTW log-rank test

We compared the Type I error rates and powers of our adjusted log-rank test versus the IPTW log-rank test (Xie and Liu 2005), which is implemented by the procedure "lifetest" of SAS 9.4. The comparison was performed in the scenarios of **Null II** and **Nearly Proportional II** defined below, in which the censoring time and the failure time are independent given the treatment, as well as in the **NULL** scenario where there is dependent censoring. The covariate and treatment variables were generated as in Section 3.1. The number of simulation runs is 1000.

$$\begin{aligned}
\textbf{Null II:}\quad & \lambda(t|Z=j,\mathbf{X}) = \exp(-3 - X_1 - 0.9X_2 - X_3)\ (j=0,1),\\
& \lambda^C(t|Z=0,\mathbf{X}) = \exp(-5),\\
& \lambda^C(t|Z=1,\mathbf{X}) = \exp(-4.5);\\
\textbf{Nearly Proportional II:}\quad & \lambda(t|Z=0,\mathbf{X}) = \exp(-2.8 - 0.95X_1 - 0.85X_2 - 0.95X_3),\\
& \lambda(t/Z=1,\mathbf{X}) = \exp(-3 - X_1 - 0.9X_2 - X_3),\\
& \lambda^C(t|Z=0,\mathbf{X}) = \exp(-5),\\
& \lambda^C(t|Z=1,\mathbf{X}) = \exp(-4.5).
\end{aligned}$$

Table 3 shows the comparison results of the adjusted log-rank test versus the IPTW log-rank test. The Type I error rates of the latter with $\alpha = 0.05$ are remarkably lower than the nominal level under **Null II**. This is no surprise given the theoretical arguments in Section 1. Under the **Null** scenario, the IPTW log-rank test has an inflated Type I error probability as a result of ignoring the dependent censoring. In contrast, the adjusted log-rank test has a correct Type I error rate in both the **Null II** and **Null** scenarios. Additionally, the adjusted log-rank test is more powerful than the IPTW log-rank test in the **Nearly Proportional II** scenario. This is again because the latter overestimates the variance of its test statistic, as argued in Section 1.

3.3 The adjusted log-rank test versus the log-rank test

We also compared the Type I error rates and powers of our adjusted log-rank test versus the regular log-rank test. The comparison was performed in the scenarios of **Null III** and **Proportional** defined below, in which the treatment assignment is random and the censoring time is independent of the failure time given the treatment. Under **Proportional**, the two treatment-specific hazards are exactly proportional to each other. The covariates were generated as in Section 3.1. The number of simulation runs is 1000.

$$\textbf{Null III:}\quad P(Z=j|\mathbf{X}) = 0.5\ (j=0,1),$$

$$\lambda(t|Z=j,\mathbf{X}) = \exp(-3 - X_1 - 0.9\, X_2 - X_3)\ (j=0,1),$$

$$\lambda^C(t|Z=0,\mathbf{X}) = \exp(-5)$$

$$\lambda^C(t|Z=0,\mathbf{X}) = \exp(-4.5);$$

Proportional: $P(Z=j|\mathbf{X}) = 0.5\ (j=0,1),$

$$\lambda(t|Z=0,\mathbf{X}) = \exp(-2.8)$$

$$\lambda(t|Z=1,\mathbf{X}) = \exp(-3),$$

$$\lambda^C(t|Z=0,\mathbf{X}) = \exp(-5),$$

$$\lambda^C(t|Z=1,\mathbf{X}) = \exp(-4.5).$$

Table 4 shows the comparison results of the adjusted log-rank test versus the log-rank test. One can see that the adjusted log-rank test has a correct Type I error rate when the adjustment is actually not necessary, and it almost does not lose efficiency because of the unnecessary adjustment compared to the log-rank test.

3.4 Performance evaluation of the proposed survival function estimator

As the last numerical experiment, we evaluated the finite sample performances of the double inverse weighted estimator (6) for the treatment-specific survival function and its pointwise confidence interval based on the log-log transformation. Table 5 shows the simulation results for the survival function estimation at three time points under the **Nearly Proportional** scenario. To save space, we just present the results for estimating $S_{(0)}(t)$. From the table, one can see that the double inverse weighted estimator $\hat{S}_{(0)}(t)$ is almost unbiased with moderate sample sizes, the theoretical variances based on Corollary 1 agree well with the empirical ones, and the 95% asymptotic confidence intervals for $S_{(0)}(t)$ based on the log-log transformation achieve the nominal coverage rate.

4 Application

We applied the proposed method to a real data set to compare the hazard of graft failure, defined as death or observed graft failure, between SPK and KA transplant recipients. The data set was obtained from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors.

Our study cohort consists of Type I diabetics with ERSD who received a SPK or KA transplant at age ≥18 during January 1, 2000–May 31, 2016. Only the patients who received the transplant for the first time were included, with repeat transplants excluded. The outcome variable is the time from the date of transplant to the date of graft failure. Subjects were censored at loss to follow-up or at the end of the observation period (May 31, 2016). We considered a common set of covariates in the logistic model for transplant type and the Cox models for censoring time, which include age at transplant, gender, race, blood type, pretransplant time on dialysis, and donor age, same as in Zhang and Schaubel (2012). The sample for the analysis has 1636 SPK and 1930 KA transplant recipients without missing data in the outcome or covariates. The longest common follow-up time in the two groups is

13.58 years. The censoring rate of the sample is about 78%. In the fitted logistic model for transplant type, age at transplant, race and donor age are significant at 0.05 level. The c-statistic for the logistic model is 0.841, suggesting that the model fits the data decently. In the fitted Cox models for censoring time, age at transplant is significant for KA recipients, and age at transplant, race and donor age are significant for SPK recipients. The Cox-Snell residual plots in Appendix B show that the Cox models fit the data well.

Figure 2 shows the double inverse weighted estimates for the KA- and SPK-specific survival functions over 10 years as well as the conditional survival functions from Year 5 to 10. One can see that KA recipients have lower graft failure rates during the first 5 years but higher rates from Year 6.3 to 10 than SPK recipients. We applied our two-sided adjusted log-rank, Prentice-Wilcoxon, Renyi-type log-rank, and Renyi-type Prentice-Wilcoxon tests to the data. The regular log-rank test and the IPTW log-rank test (Xie and Liu 2005) were also performed to compare with ours. Every test was applied to compare the treatment-specific hazards over Year 0 to 5 and Year 5 to 10 respectively. The results are tabulated in Table 6. At 0.05 level, all the tests showed that the KA versus SPK difference in the graft failure hazard over the first 5 years after transplant is statistically significant. Our tests and the regular log-rank test also showed that there is a (nearly) significant difference between the two treatment-specific hazards over the second 5 years after transplant. In contrast, the IPTW log-rank test did not detect any significant difference in the graft failure hazard between KA and SPK recipients over that period. Since our class of adjusted weighted log-rank tests are asymptotically correct in the scenario for which the IPTW log-rank test is valid (in terms of the IPTW log-rank statistic value), the above contrast implies that the IPTW log-rank test led to a false insignificant result for the time window of Year 5 to 10, assuming that our tests correctly modeled the transplant type assignment and the censoring mechanism.

5 Discussion

In this paper, we developed a class of adjusted weighted log-rank tests as well as their supremum versions, which are suitable for two-sample hazard comparisons over arbitrary time window using time-to-event data from observational studies. The adjustment is done through inverse probability of treatment and censoring weighting to deal with the non-randomness of treatment assignment and censoring often associated with observational data. The developed tests are shown to be asymptotically correct and have satisfactory finite sample performance in numerical experiments except that the Type I errors of the adjusted log-rank test and its supremum version are inflated a bit when the dependent censoring is about 40%. We have not figured out the reason for this little inflation. Further investigations are warranted. An application to a national kidney transplant data demonstrated our tests' utility.

The new tests can be directly applied to competing risks data to compare average cause-specific hazards between two treatment groups. They can also be directly applied to left-truncated survival data. Moreover, they can be directly applied to survival data with a mixture of dependent and independent censoring, in which case one should only model the distribution of the dependent censoring time given the adjustment covariates.

Three extensions of the developed tests are worthwhile to pursue in future. Firstly, our tests can be extended to the settings with more than two treatment groups by considering a new model for treatment assignment, e.g., a multinomial logistic model. Secondly, using a Cox model with time-dependent covariates for censoring, one can extend our tests to the situation where censoring times and event times are conditionally independent given some possibly timevaring prognostic factors. Thirdly, one can utilize the double-robust estimator of treatment-specific cumulative hazard proposed by Zhang and Schaubel (2012) to construct a class of adjusted weighted log-rank tests that are asymptotically correct provided that either event hazard or coarsening mechanism (treatment assignment and censoring) is modeled correctly.

Acknowledgments

This work was partially supported by National Institute of Dental and Craniofacial Research grant 1R03DE023889. The kidney transplant data used here have been supplied by the Minneapolis Medical Research Foundation (MMRF) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

Appendix A

A.1. Notations and Regularity Conditions

We introduce the following notations that will be used in the regularity conditions below and the proof of Theorem 2:

$$V_Z(\boldsymbol{\alpha}) = E\left[\frac{\exp(\boldsymbol{\alpha}^T\tilde{\mathbf{X}}^Z)\tilde{\mathbf{X}}^{Z\otimes 2}}{\left\{1+\exp(\boldsymbol{\alpha}^T\tilde{\mathbf{X}}^Z)\right\}}\right],$$

$$R_{C_j}^{(d)}(t;\boldsymbol{\theta}_j) = n^{-1}\sum_{i=1}^{n} Y_{ij}(t)\mathbf{X}_i^{C\otimes d}\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C),$$

$$r_{Cj}^{(d)}(t;\,\boldsymbol{\theta}_j) = E\left\{Y_{ij}(t)\mathbf{X}_i^{C\otimes d}\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C)\right\},$$

$$\overline{\mathbf{X}}_j^C(t;\,\boldsymbol{\theta}_j) = \frac{R_{Cj}^{(1)}(t;\,\boldsymbol{\theta}_j)}{R_{Cj}^{(0)}(t;\,\boldsymbol{\theta}_j)},$$

$$\overline{\mathbf{x}}_j^C(t;\,\boldsymbol{\theta}_j) = \frac{r_{Cj}^{(1)}(t;\,\boldsymbol{\theta}_j)}{r_{Cj}^{(0)}(t;\,\boldsymbol{\theta}_j)},$$

$$\Omega_j^C(\boldsymbol{\theta}_j) = \int_0^\infty \left\{ \frac{r_{Cj}^{(2)}(t;\, \boldsymbol{\theta}_j)}{r_{Cj}^{(0)}(t;\, \boldsymbol{\theta}_j)} - \bar{\mathbf{x}}_j^C(t;\, \boldsymbol{\theta}_j)^{\otimes 2} \right\} E\left\{ Y_{ij}(t)\lambda_{ij}^C(t) \right\} dt,$$

and $S_{0j}^C(t)$ denotes the conditional survival function of C given $Z = j$ and

$$\mathbf{X}^C = 0$$

for $j = 0$, 1 and $d = 0$, 1, 2, where for a column vector $\mathbf{b}$, $\mathbf{b}^{\otimes 2} = \mathbf{b}\mathbf{b}^T$, $\mathbf{b}^{\otimes 1} = \mathbf{b}$, and $\mathbf{b}^{\otimes 0} = 1$. The rest notations appearing in Theorems 1 and 2 will be introduced as we prove Theorem 2.

We assume the following regularity conditions for $i = 1, \ldots, n$ and $j = 0$, 1:

a. Model (2) for treatment assignment is correctly specified.

b. Model (4) for censoring is correctly specified.

c. $\mathbf{X}_i^Z$ is bounded almost surely.

d. $V_Z(\boldsymbol{\alpha})$ is positive definite at the true value of $\boldsymbol{\alpha}$.

e. $\mathbf{X}_i^C$ is bounded almost surely.

f. $\Omega_j^C(\boldsymbol{\theta}_j)$ is positive definite at the true value of $\boldsymbol{\theta}_j$.

g. $S_{(j)}(t)$ is absolutely continuous in t.

h. $S_{0j}^C(t)$ is absolutely continuous in t.

i. As $n \longrightarrow \infty$, $W(s) \xrightarrow{p} w(s)$ uniformly on $\mathscr{I}$ where $w(s)$ is a nonnegative, left-function with right-hand limits on $\mathscr{I}$ such that $w(s) < \infty$ and its right-continuous adaptation w^+ has bounded variation on each closed subinterval of $\mathscr{I}$.

Conditions (a), (c) and (d) ensure the consistency and asymptotic normality of $\hat{\boldsymbol{\alpha}}$. Conditions (b), (e) and (f) ensure the uniform consistency of $\hat{\Lambda}_{ij}^C(\,\cdot\,)$ over $[0, t_{\text{sup}}]$ and the weak convergence of $\sqrt{n}\left\{\hat{\Lambda}_{ij}^C(\,\cdot\,) - \Lambda_{ij}^C(\,\cdot\,)\right\}$ to·a zero-mean Gaussian process on $D[0, t_{\text{sup}}]$. Conditions (g) and (h) ensure that $\exp\left\{-\hat{\Lambda}_{(j)}(\,\cdot\,)\right\}$ and $\exp\left\{-\hat{\Lambda}_{ij}^C(\,\cdot\,)\right\}$ are consistent estimators for $S_{(j)}(\cdot)$ and $S_{ij}^C(\,\cdot\,)$ respectively given the consistency of $\hat{\Lambda}_{(j)}(\,\cdot\,)$ and $\hat{\Lambda}_{ij}^C(\,\cdot\,)$. Condition (i) is needed for weak convergence of $W_a(t;\, \hat{\boldsymbol{\alpha}},\, \hat{\boldsymbol{\Lambda}}^C)$ to a zero-mean Gaussian process on $D[0, t_{\text{sup}}]$ under H_0.

A.2. Proof of Theorem 2

We first decompose $W_a(t;\, \hat{\boldsymbol{\alpha}},\, \hat{\boldsymbol{\Lambda}}^C)$ as follows.

$$W_a(t;\,\hat{\boldsymbol{\alpha}},\,\hat{\boldsymbol{\Lambda}}^C) = W_a(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C) + \left\{W_a(t;\,\hat{\boldsymbol{\alpha}},\,\hat{\boldsymbol{\Lambda}}^C) - W_a(t;\,\boldsymbol{\alpha},\,\hat{\boldsymbol{\Lambda}}^C)\right\} + \left\{W_a(t;\,\boldsymbol{\alpha},\,\hat{\boldsymbol{\Lambda}}^C) - W_a(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)\right\}. \tag{16}$$

By algebra, $W_a(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)$ can be written as

$$\begin{aligned} W_a(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C) = {} & \sum_{i=1}^{n}\int_0^t \frac{K^*(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)}{\sum_{k=1}^{n}[p_{k1}(\boldsymbol{\alpha})\exp\left\{-\Lambda_{k1}^C(s-)\right\}]^{-1}Y_{k1}(s)}dM_{i1}^*(s) \\ & -\sum_{i=1}^{n}\int_0^t \frac{K^*(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)}{\sum_{k=1}^{n}[p_{k0}(\boldsymbol{\alpha})\exp\left\{-\Lambda_{k0}^C(s-)\right\}]^{-1}Y_{k0}(s)}dM_{i0}^*(s) \\ & +\int_0^t K^*(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)\left\{\lambda_{(1)}(s)-\lambda_{(0)}(s)\right\}ds \end{aligned} \tag{17}$$

Define $D_j(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C) \equiv E\left[[p_{ij}(\boldsymbol{\alpha})\exp\left\{-\Lambda_{ij}^C(s-)\right\}]^{-1}Y_{ij}(s)\right](j=0,\,1)$ and $D(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C) \equiv \sum_{j=0}^{1} D_j(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)$. Applying the Law of Large Number, then using conditions (a), (b), (h) and (i), when $n \to \infty$, we can re-express $W_a(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)$ as

$$\begin{aligned} W_a(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C) = {} & n^{-1/2}\sum_{i=1}^{n}\left\{A_{i1}(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C) - A_{i0}(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)\right\} \\ & +\int_0^t K^*(s;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)\left\{\lambda_{(1)}(s)-\lambda_{(0)}(s)\right\}ds + o_p(1) \end{aligned} \tag{18}$$

with $E\{A_{ij}(t;\,\boldsymbol{\alpha},\,\boldsymbol{\Lambda}^C)\} = 0\; j = (0,\,1)$.

To obtain the asymptotic expressions of the second and third summands in (16), we first derive the asymptotic expressions of $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ and $\sqrt{n}\left\{\hat{\Lambda}_{ij}^C(t) - \Lambda_{ij}^C(t)\right\}$.

Under conditions (a), (c) and (d), we have from standard maximum likelihood theory $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}$ and

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = V_Z^{-1}(\boldsymbol{\alpha})n^{-1/2}\sum_{i=1}^{n}\psi_i^Z(\boldsymbol{\alpha}) + o_p(1), \tag{19}$$

where $\psi_i^Z(\alpha) = \tilde{\mathbf{X}}_i^Z\left\{Z_i - \text{expit}(\boldsymbol{\alpha}^T\tilde{\mathbf{X}}_i^Z)\right\}$.

Under conditions (b), (e) and (f), standard partial likelihood theory Fleming and Harrington (1991) leads to $\hat{\boldsymbol{\theta}}_j \xrightarrow{p} \boldsymbol{\theta}_j$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) = \left\{\Omega_j^C(\boldsymbol{\theta}_j)\right\}^{-1} n^{-1/2} \sum_{i=1}^{n} U_{ij}^C(\boldsymbol{\theta}_j) + o_p(1), \quad (20)$$

where $U_{ij}^C(\boldsymbol{\theta}_j) = \int_0^\infty \left\{\mathbf{X}_i^C - \bar{\mathbf{x}}_j^C(t;\, \boldsymbol{\theta}_j)\right\} dM_{ij}^C(t)$ and

$$\sup_{t \in [0,\, t_{\sup}]} |\hat{\Lambda}_{0j}^C(t) - \Lambda_{0_j}^C(t)| \xrightarrow{p} 0 \quad (21)$$

for j=0, 1. We express $\sqrt{n}\left\{\hat{\Lambda}_{ij}^C(t) - \Lambda_{ij}^C(t)\right\}$ as

$$\sqrt{n}\left\{\hat{\Lambda}_{ij}^C(t) - \Lambda_{ij}^C(t)\right\} = \sqrt{n}\left\{\hat{\Lambda}_{0j}^C(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C) - \Lambda_{0j}^C(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C)\right\} + \sqrt{n}\left\{\Lambda_{0j}^C(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C) - \Lambda_{0j}^C(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C)\right\}. \quad (22)$$

Using a Taylor expansion,

$$\sqrt{n}\left\{\Lambda_{0j}^C(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C) - \Lambda_{0j}^C(t)\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C)\right\} = \exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C)\Lambda_{0j}^C(t)\mathbf{X}_i^{CT}\sqrt{n}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) + o_p(1). \quad (23)$$

By algebra, a Taylor expansion around $\boldsymbol{\theta}_j$, the Law of Large Number, the consistency of $\hat{\boldsymbol{\theta}}_j$, and (21),

$$\sqrt{n}\left\{\hat{\Lambda}_{0j}^{C}(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C) - \Lambda_{0j}^{C}(t)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C)\right\} = \sqrt{n}$$

$$\left\{\int_0^t \frac{\sum_{l=1}^n dN_{lj}^C(s)}{\sum_{k=1}^n Y_{kj}(s)\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_k^C)} - \int_0^t \frac{\sum_{l=1}^n dN_{lj}^C(s)}{\sum_{k=1}^n Y_{kj}(s)\exp(\boldsymbol{\theta}_j^T\mathbf{X}_k^C)}\right.$$

$$\left. + \sum_{l=1}^n \int_0^t \frac{dM_{lj}^C(s)}{\sum_{k=1}^n Y_{kj}(s)\exp(\boldsymbol{\theta}_j^T\mathbf{X}_k^C)}\right\}\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C)$$

$$= \sqrt{n}\left[-\left\{\int_0^t \overline{\mathbf{X}}_j^C(s;\,\boldsymbol{\theta}_j)\frac{\sum_{l=1}^n dN_{lj}^C(s)}{\sum_{k=1}^n Y_{kj}(s)\exp(\boldsymbol{\theta}_j^T\mathbf{X}_k^C)}\right\}^T(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)\right.$$

$$\left. + \sum_{l=1}^n \int_0^t \frac{dM_{lj}^C(s)}{\sum_{k=1}^n Y_{kj}(s)\exp(\boldsymbol{\theta}_j^T\mathbf{X}_k^C)}\right]\exp(\hat{\boldsymbol{\theta}}_j^T\mathbf{X}_i^C) + o_p(1) = -\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C$$

$$)\left\{\int_0^t \bar{\mathbf{x}}_j(s;\boldsymbol{\theta}_j)d\Lambda_{0j}^C(s)\right\}^T\sqrt{n}(\hat{\boldsymbol{\theta}}_j - \theta_j) + n^{-1/2}\sum_{l=1}^n\int_0^t \frac{dM_{lj}^C(s)}{r_{Cj}^{(0)}(s;\,\boldsymbol{\theta}_j)}\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C) + o_p(1)\,. = -\exp$$

$$(\boldsymbol{\theta}_j^T\mathbf{X}_i^C)\left\{\int_0^t \bar{\mathbf{x}}_j(s;\boldsymbol{\theta}_j)d\Lambda_{0j}^C(s)\right\}^T\sqrt{n}(\hat{\boldsymbol{\theta}}_j - \theta_j) + n^{-1/2}\sum_{l=1}^n\int_0^t \frac{dM_{lj}^C(s)}{r_{Cj}^{(0)}(s;\,\boldsymbol{\theta}_j)}\exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C) + o_p(1)\,.$$

(24)

Combining (20) (22), (23) and (24), we get

$$\sqrt{n}\left\{\hat{\Lambda}_{ij}^{C}(t) - \Lambda_{ij}^{C}(t)\right\} = K_{ij}^T(t;\,\boldsymbol{\theta}_j)\left\{\Omega_j^C(\boldsymbol{\theta}_j)\right\}^{-1}n^{-1/2}\sum_{k=1}^n U_{kj}^C(\boldsymbol{\theta}_j) + \exp(\boldsymbol{\theta}_j^T\mathbf{X}_i^C \quad (25)$$

$$)n^{-1/2}\sum_{k=1}^n\int_0^t \frac{dM_{kj}^C(s)}{r_{Cj}^{(0)}(s;\,\boldsymbol{\theta}_j)} + o_p(1),$$

where $K_{ij}(t;\,\boldsymbol{\theta}_j) = \int_0^t\left\{\mathbf{X}_i^C - \bar{\mathbf{x}}_j^C(s;\,\boldsymbol{\theta}_j)\right\}d\Lambda_{ij}^C(s)$.

Now consider $W_a(t;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Lambda}}^C) - W_a(t;\boldsymbol{\alpha},\hat{\boldsymbol{\Lambda}}^C)$ in (16). Using Taylor expansion around $\boldsymbol{\alpha}$, root-n consistency of $\hat{\Lambda}_{ij}^C(\cdot)$, the Law of Large Number and condition (i), then substituting (19), after a lot of algebra, we obtain

$$W_a(t;\hat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\Lambda}}^C) - W_a(t;\boldsymbol{\alpha},\widehat{\boldsymbol{\Lambda}}^C) = \left\{B_1(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) - B_0(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\right\}^T V_{-1}^Z(\boldsymbol{\alpha}) n^{-1/2} \sum_{i=1}^{n} \psi_i^Z(\boldsymbol{\alpha}) + o_p(1$$

),

(26)

where

$$B_j(t;\alpha,\boldsymbol{\Lambda}^C) = \int_0^t w(s) \frac{(-1)^{1-j}\sum_{k=0}^{1} D_{1-k}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) G_k(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D^2(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)} dQ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) + \int_0^t w(s$$

$$) \frac{(-1)^{1-j}\sum_{k=0}^{1} D_{1-k}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) G_k(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D^2(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)} dQ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) + \int_0^t w(s) \frac{(-1)^j D_{1-j}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)} d\xi_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C),$$

$$G_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[\widetilde{\mathbf{X}}_k^Z Y_{kj}(s) \exp\left\{\Lambda_{kj}^C(s-)\right\}\left\{p_{kj}^{-1}(\boldsymbol{\alpha}) - 1\right\}\right],$$

$$dQ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[\frac{\exp\left\{\Lambda_{ij}^C(s-)\right\} dN_{ij}(s)}{p_{ij}(\boldsymbol{\alpha})}\right]$$

and

$$d\xi_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[\widetilde{\mathbf{X}}_i^Z \exp\left\{\Lambda_{ij}^C(s-)\right\}\left\{p_{ij}^{-1}(\boldsymbol{\alpha}) - 1\right\} dN_{ij}(s)\right]$$

for $j = 0, 1$.

Next consider $W_a(t;\boldsymbol{\alpha},\widehat{\boldsymbol{\Lambda}}^C) - W_a(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)$ in (16). Using Taylor expansions around $\Lambda_{ij}^C(\cdot)$ $(i = 1, \ldots, n, j = 0, 1)$, the Law of Large Number and condition (i), then substituting (25), after a lot of algebra, we obtain

$$
\begin{aligned}
& W_a(t;\boldsymbol{\alpha},\widehat{\boldsymbol{\Lambda}}^C) - W_a(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) \\
& = F_0^T(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\left\{\boldsymbol{\Omega}_0^C(\boldsymbol{\theta}_0)\right\}^{-1} n^{-1/2}\sum_{i=1}^{n} U_{i0}^C(\boldsymbol{\theta}_0) - F_1^T(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\left\{\boldsymbol{\Omega}_1^C(\boldsymbol{\theta}_1)\right\}^{-1} n^{-1/2}\sum_{i=1}^{n} U_{i1}^C(\boldsymbol{\theta}_1) \\
& + n^{-1/2}\sum_{i=1}^{n}\int_0^t L_0(s,t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\frac{dM_{i0}^C(s)}{r_{C0}^{(0)}(s;\boldsymbol{\theta}_0)} - n^{-1/2}\sum_{i=1}^{n}\int_0^t L_1(s,t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\frac{dM_{i1}^C(s)}{r_{C1}^{(0)}(s;\boldsymbol{\theta}_1)} + o_p(1),
\end{aligned}
$$

(27)

where

$$
\begin{aligned}
F_j(t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = \int_0^t w(s)\Bigg[& \frac{H_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)D_{1-j}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D^2(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dQ_{1-j}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) + \frac{H_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)D_{1-j}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D^2(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dQ_j(s;\boldsymbol{\alpha}, \\
& \boldsymbol{\Lambda}^C) - \frac{D_{1-j}(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dJ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\Bigg],
\end{aligned}
$$

$$
H_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[K_{kj}(s-;\boldsymbol{\theta}_j)\frac{\exp\left\{\Lambda_{kj}^C(s-)\right\}Y_{kj}(s)}{p_{kj}(\boldsymbol{\alpha})}\right],
$$

$$
dJ_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[K_{kj}(s-;\boldsymbol{\theta}_j)\frac{\exp\left\{\Lambda_{kj}^C(s-)\right\}dN_{kj}(s)}{p_{kj}(\boldsymbol{\alpha})}\right],
$$

$$
\begin{aligned}
L_j(s,t;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = \int_s^t w(u)\Bigg[& \frac{\gamma_j(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)D_{1-j}(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D^2(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dQ_{1-j}(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) + \frac{\gamma_j(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)D_{1-j}(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D^2(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}dQ_j(u; \\
& \boldsymbol{\alpha},\boldsymbol{\Lambda}^C) - \frac{D_{1-j}(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}{D(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)}d\zeta_j(u;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C)\Bigg],
\end{aligned}
$$

$$
\gamma_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[\exp(\boldsymbol{\theta}_j^T\mathbf{X}_k^C)\frac{\exp\left\{\Lambda_{kj}^C(s-)\right\}Y_{kj}(s)}{p_{kj}(\boldsymbol{\alpha})}\right]
$$

and

$$d\zeta_j(s;\boldsymbol{\alpha},\boldsymbol{\Lambda}^C) = E\left[\exp(\boldsymbol{\theta}_j^T\mathbf{X}_k^C)\frac{\exp\left\{\Lambda_{kj}^C(s-)\right\}dN_{kj}(s)}{p_{kj}(\boldsymbol{\alpha})}\right]$$

for $j = 0, 1$.

Finally, substituting (18), (26) and (27) into (16) completes the proof of Theorem 2.

Appendix B

Assessment of the Cox Models for Censoring Time of the SRTR Data

References

Fleming TR, Harrington DP. A class of hypothesis tests for one and two sample censored survival data. Communications in Statistics - Theory and Methods. 1981; 10(8):763–794.

Fleming, TR., Harrington, DP. Counting processes and survival analysis. John Wiley & Sons; 1991.

Gill, R. Mathematical Centre tracts. Mathematisch Centrum; 1980. Censoring and stochastic integrals. URL https://books.google.com/books?id=Qh7vAAAAMAAJ

Prentice RL. Linear rank tests with right censored data. Biometrika. 1978; 65(1):167–179. URL http://biomet.oxfordjournals.org/content/65/1/167.abstract, http://biomet.oxfordjournals.org/content/65/1/167.full.pdf+html. DOI: 10.1093/biomet/65.1.167

Schaubel DE, Wei G. Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. Biometrics. 2011; 67(1):29–38. URL http://dx.doi.org/10.1111/j.1541-0420.2010.01449.x. DOI: 10.1111/j.1541-0420.2010.01449.x [PubMed: 20560935]

Tsiatis, A. Semiparametric theory and missing data. Springer; 2006.

van der Vaart, AW., Wellner, JA. Weak Convergence and Empirical Processes. Springer; 1996.

Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in Medicine. 2005; 24(20):3089–3110. URL http://dx.doi.org/10.1002/sim.2174. DOI: 10.1002/sim.2174 [PubMed: 16189810]

Zhang M, Schaubel DE. Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. Biometrics. 2012; 68(4):999–1009. URL http://dx.doi.org/10.1111/j.1541-0420.2012.01759.x. DOI: 10.1111/j.1541-0420.2012.01759.x [PubMed: 22471876]

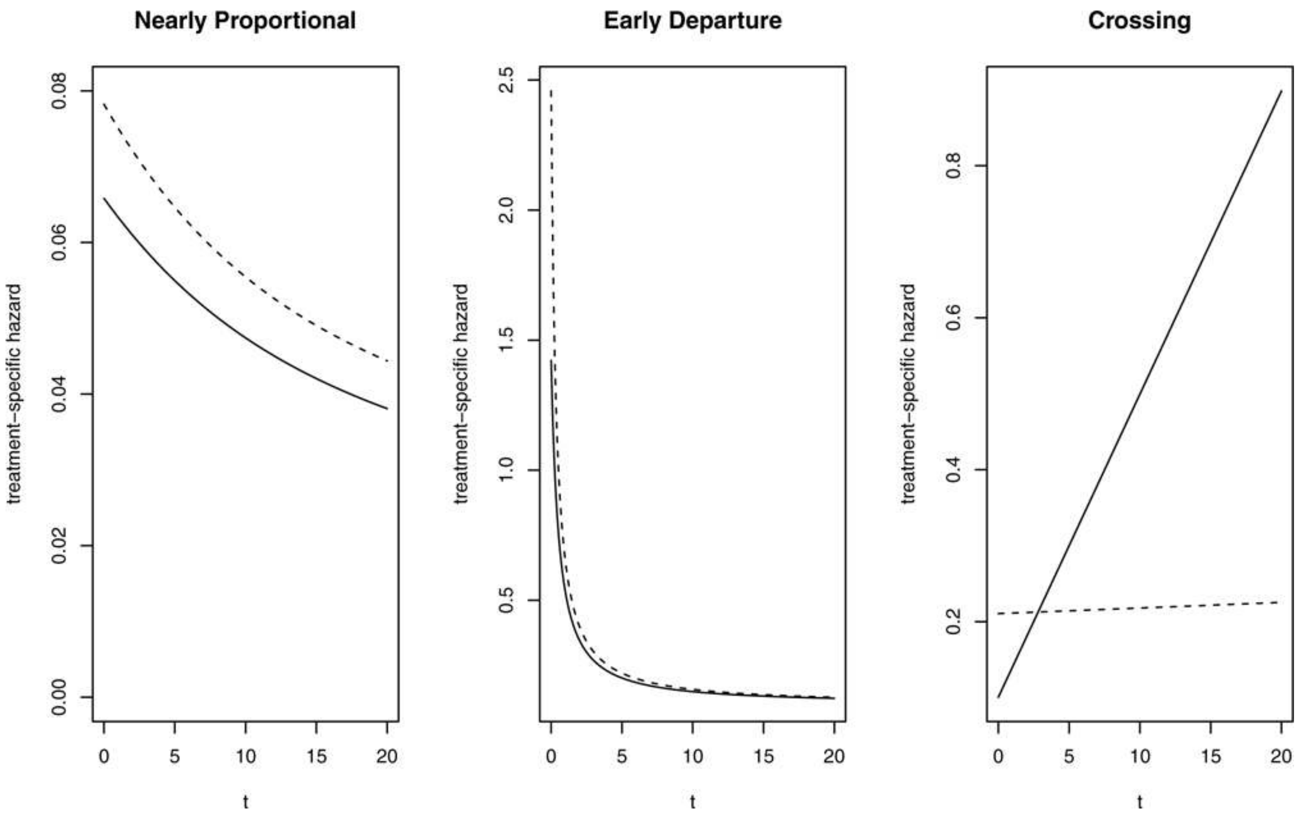

Fig. 1.

The treatment-specific hazard plots for the **Nearly Proportional**, **Early Departure** and **Crossing** scenarios. The solid curves are $\lambda_{(1)}(t)$'s, and the dashed curves are $\lambda_{(0)}(t)$'s.

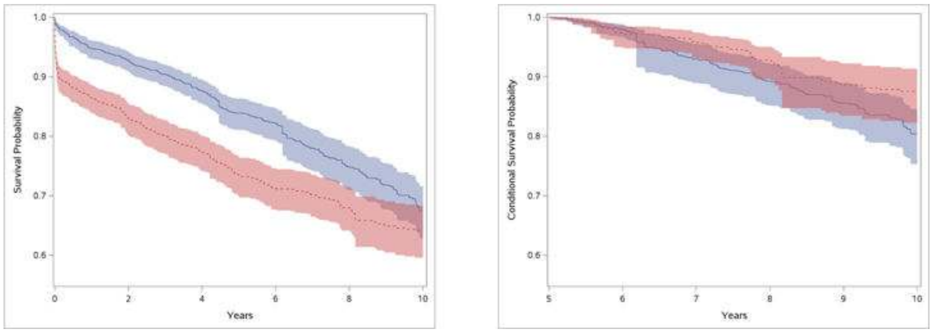

Fig. 2.

The double inverse weighted estimates (with 95% pointwise asymptotic confidence intervals based on the log-log transformation) of average survival function (left panel) and conditional survival function (right panel) of graft failure time for KA recipients (solid line) and SPK recipients (dashed line).

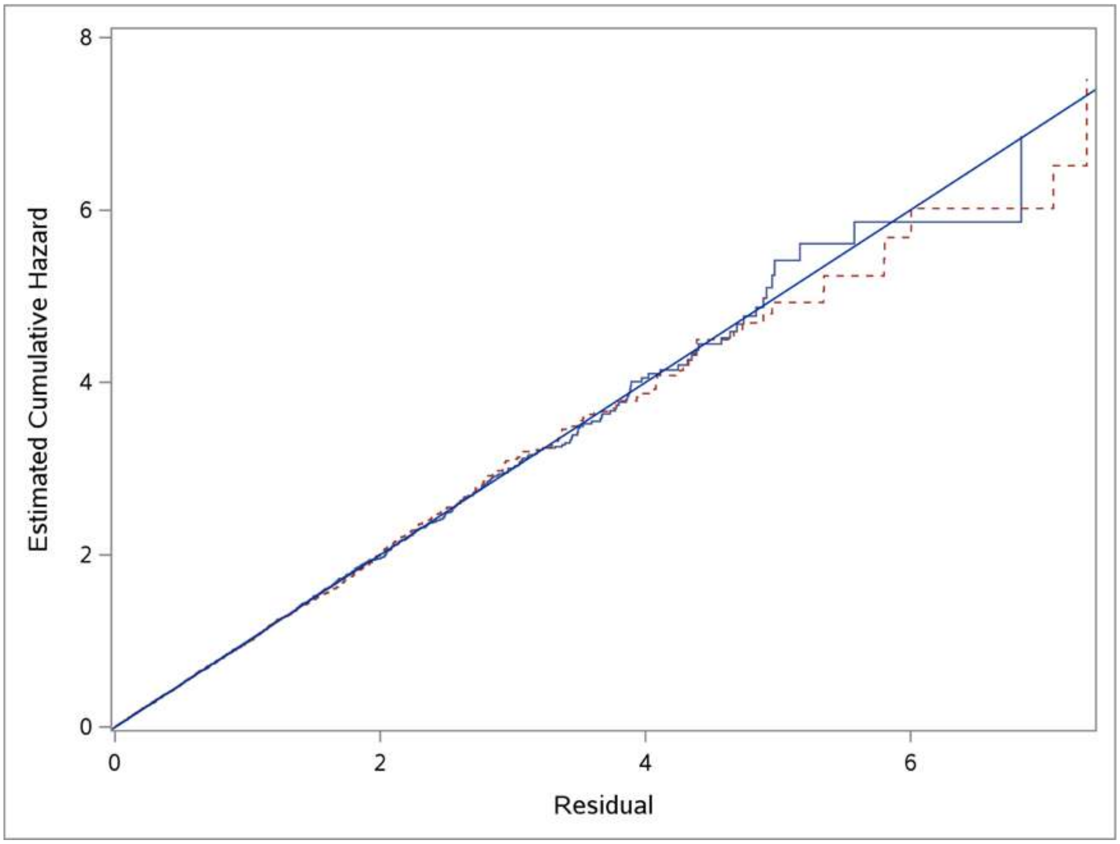

Fig. 3.

Cox-Snell residual plots for the Cox models for censoring time of KA recipients (solid line) and of SPK recipients (dashed line).

Table 1

Type I error rates and powers of the adjusted log-rank (ALR), Prentice-Wilcoxon (APW) and Renyi-type log-rank (ARLR) tests with $\alpha = 0.05$ in the scenarios where the dependent censoring rate is around 19%.

n	Scenario	ALR	APW	ARLR
300	Null	0.059	0.044	0.052
	Nearly Proportional	0.276	0.251	0.300
	Early Departure	0.637	0.856	0.779
	Crossing	0.103	0.161	0.190
600	Null	0.059	0.043	0.052
	Nearly Proportional	0.474	0.438	0.490
	Early Departure	0.870	0.991	0.971
	Crossing	0.178	0.253	0.376
900	Null	0.054	0.049	0.053
	Nearly Proportional	0.614	0.593	0.644
	Early Departure	0.972	1.000	0.994
	Crossing	0.249	0.353	0.596

TABLE 2

Type I error rates and powers of the adjusted log-rank (ALR), Prentice-Wilcoxon (APW) and Renyi-type log-rank (ARLR) tests with $\alpha = 0.05$ in the scenarios where the dependent censoring rate is around 40%.

n	Scenario	ALR	APW	ARLR
300	Null HC	0.064	0.050	0.060
	Nearly Proportional HC	0.246	0.237	0.259
	Early Departure HC	0.550	0.785	0.702
	Crossing HC	0.083	0.157	0.150
600	Null HC	0.065	0.055	0.064
	Nearly Proportional HC	0.344	0.390	0.369
	Early Departure HC	0.757	0.977	0.930
	Crossing HC	0.101	0.241	0.214
900	Null HC	0.067	0.047	0.069
	Nearly Proportional HC	0.450	0.531	0.508
	Early Departure HC	0.823	0.998	0.980
	Crossing HC	0.123	0.324	0.321

Table 3

Comparison of Type I error rate and power between the adjusted log-rank test (ALR) and the IPTW log-rank test (IPTW LR)

n	Scenario	ALR	IPTW LR
600	Null	0.059	0.062
	Null II	0.057	0.030
	Nearly Proportional II	0.501	0.424
900	Null	0.054	0.084
	Null II	0.059	0.028
	Nearly Proportional II	0.650	0.588
1200	Null	0.052	0.100
	Null II	0.051	0.031
	Nearly Proportional II	0.777	0.726

Table 4

Comparison of Type I error rate and power between the adjusted log-rank test (ALR) and the log-rank test (LR)

n	Scenario	ALR	LR
600	Null III	0.048	0.043
	Proportional	0.630	0.628
900	Null III	0.057	0.047
	Proportional	0.786	0.802
1200	Null III	0.046	0.042
	Proportional	0.899	0.901

Table 5

Simulation results for the double inverse weighted estimation of treatment-specific survival function. $E(\cdot)$ and $Var(\cdot)$ denote the empirical mean and variance respectively, and $V\hat{a}r(\cdot)$ denotes the estimator of the asymptotic variance.

n	t	$S_{(0)}(t)$	$E\{\hat{S}_{(0)}(t)\}$	$V\hat{a}r\{\hat{S}_{(0)}(t)\}$	$E[V\hat{a}r\{\hat{S}_{(0)}(t)\}]$	Coverage of 95% C.I.
600	4	0.749	0.749	6.85×10^{-4}	6.48×10^{-4}	0.945
	11	0.492	0.492	8.43×10^{-4}	8.18×10^{-4}	0.950
	26	0.247	0.247	6.60×10^{-4}	6.24×10^{-4}	0.942
900	4	0.749	0.748	4.49×10^{-4}	4.34×10^{-4}	0.945
	11	0.492	0.491	5.37×10^{-4}	5.47×10^{-4}	0.955
	26	0.247	0.247	4.30×10^{-4}	4.19×10^{-4}	0.945

Table 6

The *p*-values for testing the KA versus SPK difference in the graft failure hazard over Year 0 to 5 and Year 5 to 10 after transplant. ALR: adjusted log-rank; APW: adjusted Prentice-Wilcoxon; ARLR: adjusted Renyi-type log-rank; ARPW: adjusted Renyi-type Prentice-Wilcoxon; LR: log-rank; IPTW LR: IPTW log-rank.

Time window	Test					
	ALR	APW	ARLR	ARPW	LR	IPTW LR
[0, 5]	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
[5, 10]	0.033	0.036	0.051	0.052	0.024	0.1372