Published in final edited form as:

Lifetime Data Anal. 2014 October ; 20(4): 619–644. doi:10.1007/s10985-014-9293-9.

Semiparametric methods for center effect measures based on the ratio of survival functions

Kevin He, Ph.D. and

Department of Biostatistics, University of Michigan, 1420 Washington Hts., Ann Arbor, MI, 48109-2029, phone: (734)709-6355

Douglas E. Schaubel, Ph.D.

Department of Biostatistics, University of Michigan, 1420 Washington Hts., Ann Arbor, MI, 48109-2029, phone: (734)395-5992

Kevin He: kevinhe@umich.edu; Douglas E. Schaubel: deschau@umich.edu

Abstract

The survival function is often of chief interest in epidemiologic studies of time to an event. We develop methods for evaluating center-specific survival outcomes through a ratio of survival functions. The proposed method assumes a center-stratified additive hazards model, which provides a convenient framework for our purposes. Under the proposed methods, the center effects measure is cast as the ratio of subject-specific survival functions under two scenarios: the scenario in which the subject is treated at center j ; and that wherein the subject is treated at a hypothetical center with survival function equal to the population average. The proposed measure reduces to the ratio of baseline survival functions, but is invariant to the choice of baseline covariate level. We derive the asymptotic properties of the proposed estimators, and assess finite-sample characteristics through simulation. The proposed methods are applied to national kidney transplant data.

Keywords

Additive hazards model; Failure time data; Observational studies; Stratification

1 Introduction

In epidemiologic studies where the outcome is a possibly censored time to an event, the comparison of survival functions between groups of patients is often of great interest. For instance, for a clinical trial in which group assignment (e.g., treatment) is randomized, Kaplan-Meier (Kaplan and Meier, 1958) estimates of the treatment-specific survival functions would be of interest, along with the log-rank test for comparisons. In observational studies where group assignment is not randomized, valid comparisons of groups generally requires some form of covariate adjustment. For this purpose, Cox regression (Cox, 1972, 1975) is often carried out, such that covariate-adjusted group effects are estimated through hazard ratios (HRs). The use of the Cox model is often based on its ease of implementation

and popularity. When a Cox model is employed, group-specific HRs are convenient measures to estimate covariate-adjusted contrasts. However, the use of the HR generally assumes that the group-specific hazard functions are proportional, an assumption which is often violated in practice. Moreover, the hazard function itself may not be easily understood by the investigators or other stakeholders. Chief interest may still lie in survival probability, whether or not the hazard function is well understood by the investigators. In summary, although survival functions and contrasts therein may be of primary interest, the presence of adjustment covariates often implicitly leads to a perhaps undesirable shift in focus from the survival function to the hazard function.

In this report, we develop methods for evaluating center effects using a ratio of survival functions. The proposed methods do not assume a specific parametric form for the center effects. For concreteness, we focus the discussion hereafter on survival comparisons across medical centers. However, the concepts discussed and the methods proposed are equally applicable for comparing time-to-event by levels of a categorical variate; be it center or some other descriptor used to define group.

The motivating example for the methods proposed in this report is the evaluation of center-specific patient survival among kidney transplant centers in the United States. The evaluation of center-specific survival is important for measuring, and perhaps subsequently modifying, center-specific performance. Centers whose patient survival is significantly lower than the national average may be investigated and face both public and professional scrutiny regarding the potential reasons for their deviation from average performance.

In the context of our motivating example, there are several important issues pertaining to comparisons across centers. First, a given center effect may change systematically as follow-up time increases. For instance, a transplant center could have a substandard surgical protocol (and hence relatively high death rates early in the post-transplant follow-up period), but also have high quality follow-up strategies (resulting in lower death rates at later follow-up stages). Such trends are important to identify, since they can guide strategies for improving the quality of care delivered to patients. An appropriate analysis would permit the quantification of time-varying center effects, an issue addressed by the proposed methods. Second, studies evaluating center performance (including our motivating example) often involve a large number of centers. In such settings, the simultaneous estimation of center effects can be computationally challenging, particularly when allowance is made for time-dependence.

The analysis of center-specific outcomes has been the subject of much discussion; e.g., Wolfe et al. (1992); Lacson et al. (2001); Wolfe et al. (2008). Although the Cox model has become the most popular regression method for censored data, in certain settings, an additive hazards model may also be of interest. Additive hazard models have been studied by several authors as alternatives to multiplicative hazard models. For instance, Aalen (1980, 1989) proposed nonparametric methods for additive hazards models. Huffer and McKeague (1991) studied weighted least squares estimation for the Aalen (1980) model. Lin and Ying (1994) proposed a semiparametric additive hazards model and provided explicit parameter estimators. The Lin and Ying (1994) additive model has received a lot of attention for the

analysis of survival data, in part due to its computational convenience and because it can be considered to be a natural analog of the Cox model. With respect to computational convenience, Schaubel and Wei (2007) evaluated the connections between the Lin and Ying (1994) model and the Cox model. Extensions of the additive hazards models have been made to multivariate survival data (Yin and Cai, 2004), recurrent events (Schaubel, Zeng and Cai, 2006), interval censored data (Lin et al., 1998), competing risks (Shen and Cheng, 1999), and additive transformation models for alternative non-linear link functions (Zeng and Cai, 2010). In addition, Lin and Ying (1995) introduced a general model that included both the multiplicative and additive models as special cases.

In this report, we develop methods for evaluating center-specific outcomes through a ratio of survival functions. The proposed methods assume a center-stratified additive hazards model, which provides a convenient framework for estimating ratio of survival functions. In the next section, we describe the assumed model and proposed methods. Asymptotic properties are derived in Section 3. Finite-sample properties are examined in Section 4 through simulation studies. Section 5 features an application of the proposed methods to kidney transplant data obtained from a national organ failure registry. We then provide some discussion of the proposed and related methods in Section 6.

2 Methods

First, we will provide the notation to be used in this report. Let T_i and C_i represent the survival and censoring times, respectively, for the i 'th patient, where $i = 1, \dots, n$. Let J be the number of centers. The total number of subjects is denoted by $n = \sum_{j=1}^J n_j$, where n_j is the number of subjects in center j . Observation times are denoted by $X_i = T_i \wedge C_i$, with at-risk indicator $Y_i(t) = I(X_i \geq t)$, where $a \wedge b = \min\{a, b\}$ and $I(A)$ is an indicator function taking the value of 1 when condition A holds and 0 otherwise. The observed death indicators are denoted by $\delta_i = I(T_i < C_i)$, and the death counting process is defined as $N_i(t) = \int_0^t \delta_i(X_i) dX_i$. Let G_i denote the center for subject i and set $G_{ij} = I(G_i = j)$. Correspondingly, we set $Y_{ij}(t) = Y_i(t)G_{ij}$ and $N_{ij}(t) = N_i(t)G_{ij}$. The observed data consist of n independent vectors, $(X_i, \delta_i, G_i, \mathbf{Z}_i)$, where $\mathbf{Z}_i$ is a covariate vector. It is worth noting that the proposed methods and corresponding inference procedures are derived under the assumption that the number of centers, J , is fixed.

Based on the additive hazards model, one approach that can be used to estimate center effects is to code centers as indicator variables and assume that the center effects are constant over time. Despite its ease of calculation, this simple approach does have drawbacks when applied in studies of center-specific mortality. One particular problem is that the assumption of constant center effects is often violated in practice. Another problem is that this simple approach assumes not only the covariate effects but also the center effects are additive. To correctly identify the functional form of center effects, it is desirable to allow greater flexibility.

2.1 Stratified Additive Hazards Model

To estimate the center effects, we consider a stratified additive hazards model. The hazard function for subject i in center j with covariate $\mathbf{Z}_i$, is given by

$$\lambda_{ij}(t) \equiv \lambda(t|\mathbf{Z}_i, G_i=j) = \lambda_{0j}(t) + \beta_0^T \mathbf{Z}_i(t), \quad (1)$$

where β_0 is a regression parameter and $\lambda_{0j}(t)$ is an unspecified center-specific baseline hazard function. Note that, in cases where $\mathbf{Z}_i(t) = \mathbf{Z}_i(0)$, the covariate $\mathbf{Z}_i(t)$ would be formed through $\mathbf{Z}_i(0)$ and various functions of t . Such a covariate is of fixed-path and therefore ‘external’, using the terminology of Kalbfleisch and Prentice (2002). The use of time-dependent covariates in (1) would be primarily to remedy non-additivity.

The stratified additive hazards model is analogous to the stratified Cox model, which does not assume any functional form for the relationship among the hazard functions across strata. Such models allow the center effects to differ nonparametrically over time and, hence, allow considerable flexibility. Model (1) is a stratified version of the model developed by Lin and Ying (1994). A similar model was proposed by Yin and Cai (2004) in the context of clustered failure time data. Model (1) can also be considered a special case of the partly parametric additive hazards model (McKeague and Sasieni, 1994),

$\lambda_i(t) = \alpha(t)^T \mathbf{W}_i(t) + \beta_0^T \mathbf{Z}_i(t)$, where $\mathbf{W}(t)$ and $\mathbf{Z}(t)$ are q and p dimensional covariates respectively, and $\alpha(t)$ is an unknown vector of functions.

2.2 Proposed Center Effect Measure

The additive hazards model provides a convenient framework for estimating ratio of survival functions. Specifically, we evaluate center effects using the following ratio of baseline survival functions,

$$\psi_{0j}(t) = \frac{S_{0j}(t)}{S_{00}(t)}, \quad (2)$$

where $S_{0j}(t) = \exp\{-\Lambda_{0j}(t)\}$ is the baseline survival function for center j , $\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(u) du$ is the corresponding baseline cumulative hazard function for center j , and $S_{00}(t)$ is the baseline survival function for a hypothetical center ($j = 0$), intended to reflect the survival probability overall (i.e., across all centers). The rationale for comparing each center to the overall experience is that it is often difficult as well as undesirable to choose an appropriate reference center. Different choices for the reference center may have a strong impact on the results. To address this concern, we define the reference as

$$S_{00}(t) = \sum_{\ell=1}^J p_{\ell} S_{0\ell}(t), \quad (3)$$

where $p_{\ell} = P(G_i = \ell)$ is the marginal probability that subject i belongs to center j . A ratio $\psi_{0j}(t) > 1$ suggests that center j has a greater survival probability than the national average at

time t , while $\psi_{0j}(t) < 1$ indicates this center has a lower survival probability relative to the nation as a whole.

2.3 Ratio of Survival Functions

In this subsection, we explore the interpretation of the measure $\psi_{0j}(t)$ defined in equation (2). Under the assumed additive hazards structure, the ratio of survival functions for a particular subject reduces to the ratio of baseline survival functions. For example, the survival function for subject i (with covariate $\mathbf{Z}_i$) in center j is given by

$$S_{ij}(t) = P(T_i > t | \mathbf{Z}_i, G_i = j) = \exp \left\{ -\Lambda_{0j}(t) - \int_0^t \beta_0^T \mathbf{Z}_i(u) du \right\},$$

where we specify

$$S_{i0}(t) = P(T_i > t | \mathbf{Z}_i, G_i = 0) = \sum_{\ell=1}^J p_{\ell} S_{i\ell}(t)$$

to be the hypothetical survival function if subject i were treated at the hypothetical reference center ($j = 0$) defined by (3). Then, considering the ratio of baseline survival functions,

$$\psi_{ij}(t) \equiv \frac{S_{ij}(t)}{S_{i0}(t)} = \frac{S_{0j}(t) \exp \left\{ -\int_0^t \beta_0^T \mathbf{Z}_i(u) du \right\}}{S_0(t) \exp \left\{ -\int_0^t \beta_0^T \mathbf{Z}_i(u) du \right\}} = \frac{S_{0j}(t)}{S_{00}(t)} \equiv \psi_{0j}(t). \quad (4)$$

That is, terms in (4) involving the subject-specific covariates cancel out, $\psi_{ij}(t) = \psi_{0j}(t)$, such that the ratio of survival functions for subject i reduces to the ratio of baseline survival functions. Therefore, the ratio of baseline survival functions, $\psi_{0j}(t)$, represents the contrast between subject i at center j versus subject i at the hypothetical center ($j = 0$) with baseline survival function equal to the national average, where subject i can have any covariate value. In this light, the proposed measure is invariant to the choice of baseline covariate level. In contrast, the ratio of survival function based on the Cox model does not share this property. In particular, if $\psi_{ij}(t)$ from (2) modeled through Cox regression, $\mathbf{Z}_i$ would not cancel out, and $\psi_{0j}(t)$ would depend on the particular values chosen as the baseline covariate values. This is not an indictment of the Cox model; the additive hazards model appears to be a convenient framework when the focus is on survival as opposed to hazard functions.

2.4 Estimation of Center Effects

To estimate β_0 and $\{\Lambda_{01}(t), \dots, \Lambda_{0J}(t)\}$ from model (1), we combine ideas from Lin and Ying (1994) and Yin and Cai (2004). We repeat various results from these sources for completeness of exposition; noting that the novelty of this report lies not in parameter estimation for model (1), but specifically in the development of the center effect estimators based on (2) and (4).

According to the Doob-Meyer decomposition (Fleming and Harrington, 1991), the counting process $N_{ij}(t)$ for subject i in center j at time t can be uniquely decomposed as

$$N_{ij}(t) = M_{ij}(t) + \int_0^t Y_{ij}(u) \{d\Lambda_{0j}(u) + \beta_0^T \mathbf{Z}_i(u) du\},$$

where $M_{ij}(t)$ is a local square-integrable martingale. In view of this decomposition, the baseline cumulative hazard function for center j , $\Lambda_{0j}(t)$, can be estimated by

$$\hat{\Lambda}_{0j}(t; \hat{\beta}) = \int_0^t \frac{\sum_{i=1}^n \{dN_{ij}(u) - Y_{ij}(u) \hat{\beta}^T \mathbf{Z}_i(u) du\}}{\sum_{i=1}^n Y_{ij}(u)}. \quad (5)$$

Extending the Lin and Ying (1994) estimation approach, the estimating function for β_0 is given by,

$$U(\beta) = \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau \mathbf{Z}_i(t) \{dN_{ij}(t) - Y_{ij}(t) d\hat{\Lambda}_{0j}(t; \beta) - Y_{ij}(t) \beta^T \mathbf{Z}_i(t) dt\}. \quad (6)$$

Substituting (5) into (6) yields

$$U(\beta) = \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_j(t)\} \{dN_{ij}(t) - Y_{ij}(t) \beta^T \mathbf{Z}_i(t) dt\}, \quad (7)$$

where

$$\bar{\mathbf{Z}}_j(t) = \frac{\sum_{i=1}^n Y_{ij}(t) \mathbf{Z}_i(t)}{\sum_{i=1}^n Y_{ij}(t)}.$$

Setting $U(\beta) = 0$, the resulting estimator for β_0 has the following closed form, $\hat{\beta} = \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}}$ where

$$\begin{aligned} \hat{\mathbf{A}} &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau Y_{ij}(t) \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_j(t)\}^{\otimes 2} dt, \\ \hat{\mathbf{B}} &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_j(t)\} dN_{ij}(t), \end{aligned} \quad (8)$$

where, for a vector $\mathbf{a}$, we have $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$ and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$. Similar to previous authors, to ensure that the estimated cumulative hazard $\hat{\Lambda}_{0j}(t)$ is monotone in t , a modification can be defined as

$$\tilde{\Lambda}_{0j}(t) = \max_{s \leq t} \hat{\Lambda}_{0j}(s; \hat{\beta}).$$

Using arguments similar to those in Lin and Ying (1994), $\tilde{\Lambda}_{0j}(t)$ and $\hat{\Lambda}_{0j}(t)$ are asymptotically equivalent.

Combining each of the estimators listed in this section, the center effects are then estimated by

$$\hat{\psi}_{0j}(t) = \frac{\hat{S}_{0j}(t)}{\hat{S}_{00}(t)},$$

where $\hat{S}_{0j}(t) = \exp\{-\hat{\Lambda}_{00}(t)\}$ and $\hat{S}_{00}(t) = \sum_{\ell=1}^J \hat{p}_{\ell} \hat{S}_{0\ell}(t)$, with $\hat{p}_{\ell} = n_{\ell}/n$.

2.5 Model Checking

Since the proposed methods rely on the correctness of the assumed stratified additive model, it is generally important to carry out model checking. A variety of model checking and goodness-of-fit tests have been proposed for the additive hazards models (e.g., Ghosh 2003, Gandy and Jensen 2005, Yin 2007, Zeng and Cai, 2010). In each of these works, model checking involves the use of simulation methods to approximate the distribution of stochastic process, which requires relatively extensive computation (particularly in larger data sets, such as that which motivated our work). Alternatively, a variety of residuals can be employed to check the additive model. For instance, the Cox-Snell residuals are useful for checking the overall fit of the model. The martingale residuals are useful for determining the functional form of covariates. The best functional form may be a transform of the covariates; such as categorization, linear splines and polynomials. In Section 5, we apply some of these techniques to check the model assumption and to verify the assumed functional forms of various covariates.

3 Asymptotic Properties

To derive the large-sample properties for the proposed estimators, we impose the following regularity conditions for $i = 1, \dots, n$ and $j = 1, \dots, J$.

- a. $(X_i, G_i, \mathbf{Z}_i)$ are independent and identically distributed random vectors.
- b. $P(X_i \leq \tau) > 0$ where τ is a pre-specified time point.
- c. Z_{ik} have bounded total variation, i.e., $|Z_{ik}(0)| + \int_0^{\tau} |dZ_{ik}(t)| < \kappa$ for $k = 1, \dots, p$, where κ is a constant and Z_{ik} is the k th component of $\mathbf{Z}_i$.
- d. $\int_0^{\tau} \lambda_j(t) dt < \infty$.
- e. Positive-definiteness of the matrix,

$$\mathbf{A} = E \left[\sum_{j=1}^J \int_0^\tau Y_{1j}(t) \{Z_1(t) - \bar{Z}_j(t)\}^{\otimes 2} dt \right], \quad (9)$$

the limiting value of $\hat{\mathbf{A}}$, defined in (8).

- f. $P(G_{ij} = 1 | \mathbf{Z}_i) > 0$.
- g. The number of centers, J , is fixed.

Condition (a) is employed in the derivation of asymptotic normality. Condition (b) is a standard identifiability requirement. Condition (c) leads to the boundedness of several quantities arising in the proofs, and is applicable in most practical applications. Condition (d) is not essential but simplifies our proofs. Condition (e) will generally hold under an appropriate specification of the covariate vector. With respect to condition (f), the selection probability given covariates is non-zero for all centers. This condition guarantees that the sample size of each center, n_j , goes to ∞ as the total sample size, n , goes to ∞ .

The asymptotic properties related coefficients and the baseline cumulative hazard function estimates are straightforward extensions of existing works (e.g., Lin and Ying, 1994; Yin and Cai, 2004), and we present these results in Appendices for completeness of exposition. Asymptotic properties of the proposed survival ratios are given in the following theorem.

THEOREM 1

The random process $n^{\frac{1}{2}} \{\hat{\psi}_{0j}(t) - \psi_{0j}(t)\}$ converges weakly to a zero mean Gaussian process with covariance function given by $\sigma_{\psi j}(s, t) = E[Y_{1j}(s)Y_{1j}(t)]$ where

$$\Upsilon_{ij}(t) = \mathbf{H}_j^T(t) \boldsymbol{\xi}_i(t),$$

and $\boldsymbol{\xi}_i(t) = [\xi_{i1}(t), \dots, \xi_{iJ}(t)]^T$, with $\xi_{i1}(t)$ defined in Appendix 1. The covariance function can be consistently estimated by $\hat{\sigma}_{\psi j}(s, t) = n^{-1} \sum_{i=1}^n \hat{\Upsilon}_{ij}(s) \hat{\Upsilon}_{ij}(t)$, with $\hat{\Upsilon}_{ij}(t)$ obtained by replacing limiting values in $\Upsilon_{ij}(t)$ with their empirical counterparts.

A proof of Theorem 1 is provided in Appendix 2. Note that Martingale theory (Fleming and Harrington, 1991) can be applied to our set-up, as the stability and negligibility conditions are easy to verify. However, we instead derive the asymptotic properties for the proposed estimator using empirical process theory. The resulting sandwich-type asymptotic variance estimator should be more robust to the model mis-specification. Moreover, the proposed methods could be easily extended to settings with recurrent events or missing data studies with unpredictable weights (e.g., case-cohort studies), to which Martingale results are unlikely to apply.

In practice, the center effects may change over follow-up time. In some situations, it will be of interest to determine whether a particular center j has survival different from the national average at any time point t over a pre-defined interval $[0, L]$. A natural approach in this

context is the supremum test. Using techniques similar to those used to derive simultaneous confidence bands, (e.g., Lin et al., 1994; Lin et al., 2000), we propose Kolmogorov-Smirnov-type statistics, for which the null hypothesis is $H_0 : \psi_j(t) = 1$ for $t \in [0, L]$, where L is a pre-specified time point. The test statistic is given by

$$\hat{F}_j = \sup_{0 \leq t \leq L} \left| n^{\frac{1}{2}} \{ \hat{\psi}_j(t) - 1 \} \right|. \quad (10)$$

Under the null hypothesis, $H_0 : \psi_j(t) = 1$ for $t \in [0, L]$, we have that $n^{\frac{1}{2}} \{ \hat{\psi}_j(t) - 1 \}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^n \hat{Y}_{ij}(t)$. We then use a simulation technique constructed by Lin et al. (1994) and Lin et al. (2000), which suggests that the quantity, $n^{-\frac{1}{2}} \sum_{i=1}^n R_i \{ \hat{Y}_{ij}(t) - 1 \}$, would asymptotically have the same distribution as $n^{\frac{1}{2}} \{ \hat{\psi}_j(t) - 1 \}$ under H_0 , where R_i is a standard normal random variable. Repeatedly generating independent standard normal random samples R_i ($i = 1, \dots, n$), we can empirically determine the critical values of this test.

4 Simulation Study

The finite-sample properties of the proposed methods were evaluated through a series of simulation studies. We considered $J = 10$ centers, with death times generated as

$$T_{ij} = \frac{-\log(U_i)}{\lambda_j + \beta_0^T \mathbf{Z}_i}$$

for $i = 1, \dots, n_j$ and $j = 1, \dots, 10$, where U_i was a Uniform(0,1) random variable; $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})^T$, Z_{i1} followed a Bernoulli distribution, Z_{i2} followed a logistic model, and Z_{i3} came from a Normal distribution. We set $\beta_0^T = (\beta_1, \beta_2, \beta_3) = (0.01, -0.05, 0.1)$. The number of subjects within each center varied as 20, 50 and 100. Censoring times were generated from a Uniform (1, 10) distribution, with the percentage of censored subjects then being approximately 15%. Each data configuration was replicated 1000 times.

We considered two different simulation settings to evaluate the proposed center effect estimators. In Setting 1, the hazard functions and covariate distributions were equal across centers. Specifically, the values of λ_j were chosen to be equal to 0.5 for all centers. The covariate Z_{i1} followed a Bernoulli (0.5) distribution; Z_{i2} followed a logistic model with probability dependent on Z_{i1} (e.g., $\text{logit}\{P(Z_{i2} = 1)\} = 0.5 + \log(0.75)Z_{i1}$); and Z_{i3} came from a Normal distribution with variance 25 and mean $E[Z_{i3}|Z_{i1}, Z_{i2}] = 50 + 0.1Z_{i1} + 0.05Z_{i2}$. Results for the proposed ratio of survival functions were provided in Table 1 for the case of center-specific sample size of $n_j = 100$. The center effect estimator was sufficiently well-behaved, in the sense that the bias was quite small and the ASEs were generally close to the ESDs, while the empirical CPs were generally consistent with the nominal value, 0.95. Results were generally consistent across the observation time distribution. The results for other choices of n_j were similar (data not shown).

In Setting 2, both the hazard functions and covariate distributions were center-dependent. For instance, different values of λ_j ($\lambda_j = 0.45, 0.48, 0.51, 0.54, 0.57, 0.6, 0.63, 0.66, 0.69$ and 0.72 for $j = 1, \dots, 10$) were used, such that the hazard functions increased with increasing center numbers. The ratio of survival functions was calculated at various percentiles of observation time. We calculated the empirical standard deviations (ESD), the average asymptotic standard error (ASE), and the 95% coverage probabilities (CP). Results for the proposed ratio of survival functions were provided in Tables 2–4 for various sample sizes. The $\Lambda_{0j}(t)$ and $\psi_{0j}(t)$ estimators were calculated at various time points which correspond different percentiles of the observation time distribution. When the center-specific sample size was small (e.g., 20 in Table 2), the empirical CPs were slightly underestimated for centers with either extremely large or smaller hazards. As the sample size for each center increased to 50 or 100, the proposed center effect estimator was sufficiently well-behaved (Table 3 and Table 4). Such results suggest that center-specific sample sizes play an important role in the proposed methods. In particular, the minimum sample size (across centers) needs to be reasonably large. Given this concern, centers of size less than 50 were eliminated from the real data analysis in Section 5.

As a Reviewer suggested, another interesting and important practical question is the reliability of the resulting estimated rankings. We explored this issue (under Setting 2) by calculating the percentage of times that the proposed method correctly identifies the best and worst centers. When center-specific sample size was $n_j = 100$, the percentages were 46.3% and 33.8% for the best and the worst center, respectively. When sample sizes were increased to $n_j = 500$, the corresponding percentages were 72.3% for the best center and 58.6% for the worst center. It is worth noting that empirical Bayes methods have been proposed specifically for rank ordering centers (e.g., Laird and Louis, 1989). In contrast, our proposed method aims to address a simpler question: to focus on centers with extreme outcomes, with centers having mortality close to the population average being of little direct interest. From this perspective, it appears that the proposed methods rank the best and worst centers with adequate accuracy.

We carried out another set of simulations to explore the performance of the supremum test introduced at the end of Section 3. The set-up for the hazard functions and covariate distributions were exactly the same as those for Setting 1, described in the second paragraph of this Section. To avoid unstable estimation in the tails, the supremum test was calculated over $t \in [0, 3]$ in order to evaluate whether a particular center effect was significantly different from the overall average at any point during the first three time units of follow-up. In order to compute $\hat{F}_j$, we obtained 1000 realizations of $(R_1, \dots, R_n)$, where $R_i \sim N(0, 1)$ for each simulated data set. Results are given in Table 5. The proposed supremum test appeared to yield the desired Type I error rate, with empirical significance levels being close to 0.05.

5 Application to Kidney Transplant Data

We applied the proposed methods to evaluate U.S. kidney transplant centers. Data were obtained from the Scientific Registry of Transplant Recipients (SRTR). Included in the analysis were adult patients (≥ 18 years of age at transplant) who underwent kidney transplantation between January 2000 and December 2008. Adjustment covariates were age,

race, gender, diagnosis, donation after cardiac death (DCD), Expanded Criteria Donor (ECD), body mass index (BMI), dialysis time, previous kidney transplant and cold ischemia time. Patients receiving living-donor transplants were excluded. Subsequently, centers with a sample size of $n_j < 50$ were eliminated from additional analysis. The final sample size was $n = 73,517$ from $J = 207$ centers across the United States. Graft failure was considered to occur when the transplanted kidney ceased to function sufficiently to preserve life. Specifically, the event time (recorded in years) was defined as the time from transplantation to graft loss or death, whichever occurred first.

We fitted the stratified additive hazard model and estimated the ratio of survival functions for each center at year 1, 3, and 5. Centers with a lower confidence limit greater than 1 were identified as significantly better than the national average, while centers with an upper confidence bound less than 1 were identified as significantly worse than the national average. The differences across centers tended to increase with increasing t . At year 1, a total of 10 centers had survival functions significantly worse than the national average, while 13 centers were significantly above the average. At year 3, 18 centers had survival function significantly worse than the national average, while 20 centers were significantly above the average. At year 5, 21 centers had survival function significantly worse than the national average, while 25 centers were significantly above the average. In Figure 1 we plot ratio of survival functions at year 3, and associated confidence intervals, with centers ordered by $\psi_{0j}(3)$. Note that approximately 50% of subjects are still at risk at $t = 3$.

The evaluation of time-varying center effects indicated that there were considerable changes in the survival pattern within certain centers. This trend is clear in Figure 2, in which the proposed measure is considered as a function over time. In Figure 2A, the ratio of survival functions steadily increases, which suggests this center had better performance than the national average. However, this trend was not significant in the first two years. In Figure 2B, the center might have both substandard surgery protocol and substandard follow-up treatment since the ratio of survival functions steadily decreases. In contrast, the transplant center in Figure 2C might have a substandard surgery protocol and hence have relatively poor survival functions early in the post-transplant follow-up period; however, improved performance is observed in the later stages of follow-up. Computation of $\psi_{0j}(t)$ at a fixed t alone would fail to identify such patterns.

Finally, we used Cox-Snell residuals to evaluate the overall fit of the assumed stratified additive hazards model. In particular, the Cox-Snell residual was evaluated for each subject at its observed follow-up time. Plots of the estimated cumulative hazards versus the residuals were then generated (Figure 3), with such plots expected to approximately a 45-degree line if the assumed additive model is correct. Here, the estimates seem to be close to the 45 degree line, except in the tail where the estimates are quite variable. To further assess the fit of the model, we explored the functional form of age, which is among the strongest predictors of post-transplant mortality. The resulting figures (not shown) indicate that a linear form is a better fit than either a quadratic form or piecewise-linear spline.

5.1 Control of False Discovery Rate

The issue of multiple testing may be a concern when identifying under- or over-performance among a large number of centers. In settings where the goal is to provide definite results (i.e., guard against any false discoveries), one commonly used approach is the family-wise error rate (FWER), defined as the probability of making one or more false discovery among all the hypotheses (Hochberg and Tamhane, 1987). Controlling FWER is a form of simultaneous inference, where all inferences are jointly corrected up to a pre-specified (global) error rate.

Simultaneous inference could perhaps be considered appropriate if the results of center monitoring were quite extreme; for example, to shut down under-performing centers and financially reward those that over-performed. However, in our motivating example (as in many cases where registry data are used to monitor facilities on a large scale), the purpose is to identify a small percentage of outlying centers that deserve further investigation, such that the center monitoring serves as more of a screening operation. From this perspective, a limited number of false discoveries would be acceptable. Therefore, in the setting of our interest, simultaneous inference would be too conservative. A more relevant control is the false discovery rate (FDR), defined as the expected proportion of false positives among all significant tests (Benjamini and Hochberg; 1995). FDR controlling procedures are more powerful than the FWER procedures, at the cost of increasing the likelihood of false-positives. In large-scale testing situations, Efron (2004) proposed an empirical null method to empirically estimate an appropriate null distribution, which will in turn affect the FDR. In this direction, we applied Efron's local FDR (hereafter denoted by fdr) on our setting. The essence of the fdr approach is to identify a small percentage of ("interesting") cases that deserve further investigation. The details of implementation can be found in Efron (2004). In our analysis, we used the R package *locfdr* to obtain the estimated fdr . Using this approach, only two centers have $fdr(z_i) < 0.2$. This finding can be explained by Figure 4, in that the central peak in the plot is wide, such that most values appear to have arisen from a common distribution, with few outliers. Note that the threshold $fdr(z_i) < 0.2$ was suggested by Efron (2007); this is conservative, from the perspective of requiring a much stronger level of evidence against the null hypothesis. The fdr method controls the Type I errors at a cost of potentially low power; e.g., many of the non-null cases may have large values of $fdr(z_i)$.

Finally, we also applied an empirical null method proposed by Kalbfleisch and Wolfe (2013), which is essentially based on the methods proposed by Efron (2004) and is designed for correction of over-dispersion. In this context, we refer to over-dispersion as the case where a center effect estimator shows substantially more variability than would be expected by chance alone (Spiegelhalter 2005). Over-dispersion often arises due to incomplete covariate information. Ignoring over-dispersion could result in a large percentage of centers being flagged. In this sense, although the empirical null method by Kalbfleisch and Wolfe (2013) does not control FDR directly, it is helpful to reduce FDR. To implement this approach, the p-value for each center was converted to a Z-score, and the corresponding histograms are plotted (Figure 4). The $N(0, 1)$ density is then superimposed on the histogram along with a normal curve fitted to the center of the histograms using a robust M-estimation method. It is clear that the overdispersion of the Z-scores is substantial in this example.

Therefore, significance testing to identify outlying centers was based on the empirical null distribution, instead of the standard Normal density. Table 6 presents the flagging numbers and rates with and without using empirical null method. The results presented are based on a two-tailed test where the relevant p-value is 5% or less. The method without using empirical null would flag about 18.4% of centers. In contrast, the empirical null method makes an appropriate adjustment and yields flagging rates of 5.3%.

Note that clinical requirements and policy-making considerations play important roles in deciding an appropriate approach to reduce FDR. In this light, the empirical null method by Kalbfleisch and Wolfe (2013) reduces FDR and leads to a small percentage of outlying centers for further investigation. Similar techniques for controlling over-dispersion have been applied in Spiegelhalter (2005) and Spiegelhalter et al. (2012).

6 Discussion

In this report, we developed methods for evaluating centers-specific survival through a ratio of survival functions. The center-stratified additive hazards model provides a convenient framework in this setting. For the proposed measure and under the assumed model, methods, the ratio of survival functions (for a particular subject) reduces to ratio of baseline survival functions. In this light, the ratio of survival functions is invariant to the choice of baseline covariate level. The ratio of survival functions then represents the contrast between subject i at center j versus subject i at the hypothetical center with a baseline survival function equal to the national average; where subject i can have any covariate value. In contrast, ratio of survival functions calculated from the Cox model does not share these properties.

An alternative approach to evaluate center effects would be to code centers using indicator variables. Hence, the $\lambda_{0j}(t)$ in the stratified additive hazard model (1) would be replaced, for instance, by $\lambda_0(t) + \sum_{\ell=1}^{J-1} G_{\ell j} \beta_{0\ell}$, with center $j = J$ serving as the reference. Such a model assumes constant center effects. Hence, it is considerably less flexible. In cases where the number of centers is small, time-varying center effects can be estimated through the partly parametric additive hazards model proposed by McKeague and Sasieni (1994), who applied kernel smoothing techniques to construct efficient estimators. Their approaches are appealing in term of efficiency and versatility. However, we did not apply kernel smoothing in this report since we were primarily targeting multi-center observational databases with a large number of centers, in which case smoothing methods often would not be convenient to apply.

As pointed out by an anonymous Reviewer, sample size may play an important role in the profiling of medical providers. For example, all other considerations being equal, smaller centers are less likely to be identified as outlying since power is reduced. In settings where the number of centers being profiled is relatively large (e.g., 1000), one option would be to apply the afore-described empirical null methods within strata defined by center size; e.g., centers could be stratified into quartiles based on number of patients. Kalbfleisch and Wolfe (2013) showed that such an approach resulted in flagging rates that were fairly consistent across strata. Based on the work of Efron (2004), stratum sizes should be 200 in order to accurately specify an empirical null distribution. Note that in using such stratification,

centers are not compared to the national average, but to the average among centers of similar size; which may be undesirable in certain settings.

Acknowledgments

This work was supported in part by National Institutes of Health grant 5R01-DK070869. The authors thank the Associate Editor and two Reviewers for constructive comments which strengthened the article. The data reported 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 authors and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

References

- Aalen, OO. Lecture Notes in Statist. Vol. 2. Springer; New York: 1980. A model for nonparametric regression analysis of counting processes. *Mathematical Statistics and Probability Theory*; p. 1-25.
- Aalen OO. A linear regression model for the analysis of life times. *Stat Med*. 1989; 8:907–925. [PubMed: 2678347]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B*. 1995; 57:289–300.
- Chen P, Tsiatis AA. Causal inference on the difference of the restricted mean life between two groups. *Biometrics*. 2001; 57:1030–1038. [PubMed: 11764241]
- Cox DR. Regression models and life tables (with Discussion). *J Roy Stat Soc B*. 1972; 34:187– 200.
- Cox DR. Partial likelihood. *Biometrika*. 1975; 62:269–276.
- Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Stat Assoc*. 2004; 99:96–104.
- Efron B. Size, power and false discovery rates. *Ann Stat*. 2007; 35(4):13511377.
- Fleming, TR.; Harrington, D. *Counting Processes and Survival Analysis*. New York: Wiley; 1991.
- Gandy A, Jensen U. Checking a semiparametric additive hazards model. *Lifetime Data Anal*. 2005; 11:451472.
- Ghosh D. Goodness-of-fit methods for additive-risk models in tumorigenicity experiments. *Biometrics*. 2003; 59:721726.
- Hochberg, Y.; Tamhane, AC. *Multiple comparison procedures*. Wiley; 1987.
- Huffer FW, McKeague IW. Weighted least squares estimation for Aalen's additive risk model. *J Am Stat Assoc*. 1991; 95:238–248.
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2. New York: Wiley; 2002.
- Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. *Statistics in the Biosciences*. 2013 In Press.
- Kosorok MR. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. 2008
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958; 53:457481.
- Lacson E, Teng M, Lazarus JM, Lew N, Lowrie ED, Owen WF. Limitations of the facility-specific standardized mortality ratio for profiling health care quality in dialysis. *American Journal of Kidney Diseases*. 2001; 37:267–275. [PubMed: 11157366]
- Laird NM, Louis TA. Empirical Bayes ranking methods. *Journal of Educational Statistics*. 1989; 14(1):29–46.
- Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards models. *Biometrika*. 1994; 81:73–81.
- Lin DY, Oakes D, Ying Z. Additive hazards regression for current status data. *Biometrika*. 1998; 85:289–298.
- Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*. 1994; 81:61–71.

- Lin DY, Ying Z. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann Stat*. 1995; 23:1712–1734.
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J Roy Stat Soc B*. 2000; 62:11–730.
- McKeague IW, Sasieni PD. A partly parametric additive risk model. *Biometrika*. 1994; 81:501–514.
- O'Neill TJ. Inconsistency of the misspecified proportional hazards model. *Statistics & Probability Letters*. 1986; 4:219–222.
- Schaubel DE, Wei G. Fitting semiparametric additive hazards models using standard statistical software. *Biometrical Journal*. 2007; 49:719–730. [PubMed: 17638295]
- Schaubel DE, Zeng D, Cai J. A semiparametric additive rates model for recurrent event data. *Lifetime Data Anal*. 2006; 12:389–406. [PubMed: 17031499]
- Shen Y, Cheng SC. Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics*. 1999; 55:1093–1100. [PubMed: 11315053]
- Shorack, GR.; Wellner, JA. *Empirical Processes with Applications to Statistics*. New York: Wiley; 1986.
- Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Health Care*. 2005; 14:347351.
- Spiegelhalter DJ, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance (with discussion). *J Roy Statist Soc Ser A*. 2012; 175:1–25.
- van der Vaart, AW. *Asymptotic Statistics*. Cambridge: 1998.
- van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes*. Springer; New York: 1996.
- Wolfe RA, Gaylin DS, Port FK, Held PJ, Wood CL. Using USRDS generated mortality tables to compare local ESRD mortality rates to national rates. *Kidney International*. 1992; 42:991–996. [PubMed: 1453592]
- Wolfe RA, McCullough KP, Schaubel DE, Kalbfleisch JD, Murray S, Stegall MD, Leichtman AB. Calculating life years from transplant (LYFT): Methods for kidney and kidney-pancreas candidates. *American Journal of Transplantation*. 2008; 8:997–1011. [PubMed: 18336702]
- Yin G, Cai J. Additive hazards model with multivariate failure time data. *Biometrika*. 2004; 91:801–818.
- Yin G. Model checking for additive hazards model with multivariate survival data. *Journal of Multivariate Analysis*. 2007; 98:1018–1032.
- Zeng D, Cai J. Additive transformation models for clustere failure time data. *Lifetime Data Anal*. 2010; 16:333–352. [PubMed: 20012358]

APPENDIX

Appendix 1: Large-sample distribution of β

With simple algebraic manipulation, formula (7) yields

$$U(\beta_0) = \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau \{Z_i(t) - \bar{Z}_j(t)\} dM_{ij}(t).$$

In the argument below, we demonstrate that $\bar{Z}_j(t)$ converges almost surely to $\bar{z}_j(t)$ uniformly in $t \in [0, \tau]$, where

$$\bar{\mathbf{z}}_j(t) = \frac{s_j^{(1)}(t)}{s_j^{(0)}(t)},$$

with $s_j^{(1)}(t)$ being the limiting value of $S_j^{(1)}(t)$, and $s_j^{(0)}(t)$ being the limiting value of $S_j^{(0)}(t)$.

We then show below that $n^{-\frac{1}{2}}U(\beta_0)$ is essentially a sum of independent and identically distributed random variables:

$$n^{-\frac{1}{2}}U(\beta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n u_i(\beta_0) + o_p(1), \quad (11)$$

where

$$u_i(\beta_0) = \sum_{j=1}^J \int_0^\tau \{ \mathbf{Z}_i(t) - \bar{\mathbf{z}}_j(t) \} dM_{ij}(t).$$

The Multivariate Central Limit Theorem can be used to demonstrate that the random vector $n^{-\frac{1}{2}}U(\beta_0)$ asymptotically follows a normal distribution with mean zero. Asymptotic properties of the regression coefficient estimator are then given in the following Lemma.

Lemma 1

The random vector, $n^{\frac{1}{2}}(\hat{\beta} - \beta_0)$, converges weakly to a zero-mean normal random vector with a variance-covariance matrix given by $\sigma_\beta = \mathbf{A}^{-1}E[u_1(\beta_0)u_1(\beta_0)^T]\mathbf{A}^{-1}$.

A consistent estimator of the covariance matrix is given by

$$\hat{\sigma}_\beta = \hat{\mathbf{A}}^{-1} \left\{ n^{-1} \sum_{i=1}^n \hat{u}_i(\hat{\beta}) \hat{u}_i(\hat{\beta})^T \right\} \hat{\mathbf{A}}^{-1},$$

with $\hat{u}_i(\hat{\beta})$ obtained by replacing limiting values in $u_i(\beta_0)$ by their empirical counterparts.

To show that (11) holds, i.e.,

$$n^{-\frac{1}{2}} \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau \{ \bar{\mathbf{Z}}_j(t) - \bar{\mathbf{z}}_j(t) \} dM_{ij}(t) \xrightarrow{P} 0,$$

we first prove that $\bar{\mathbf{Z}}_j(t)$ converges almost surely to $\bar{\mathbf{z}}_j(t)$ uniformly in $t \in [0, \tau]$. Note that the collection of all cells $[t, \infty)$ in the real line is a VC class of index 2 (van der Vaart, 1998) and, hence, satisfies the entropy conditions for the Glivenko-Cantelli theorem (van der Vaart

and Wellner, 1996). Therefore, $\{I(X \leq t)I(G=j), t \in [0, \tau]\}$ belong to some Glivenko-Cantelli class, and $\mathbb{P}_n[I(X \geq t)I(G=j)] \xrightarrow{a.s.} \mathbb{P}[I(X \geq t)I(G=j)]$ uniformly in $t \in [0, \tau]$, where $\mathbb{P}$ denotes expectations and $\mathbb{P}_n$ is the empirical measure (i.e.

$\mathbb{P}_n[I(X \geq t)I(G=j)] = n^{-1} \sum_{i=1}^n Y_{ij}(t)$). Next, note that $\mathbf{Z}(t)$ is a stochastic process with bounded variation, and every function with bounded variation is the difference of two bounded monotonic functions. We then have that $\mathbf{Z}(t)$ belongs to some Glivenko-Cantelli class indexed by t . Moreover, both $I(X \leq t)$ and $\mathbf{Z}(t)$ are uniformly bounded. Hence, the product of these two classes form a Glivenko-Cantelli class. Therefore,

$\mathbb{P}_n[I(X \geq t)I(G=j)\mathbf{Z}(t)] \xrightarrow{a.s.} \mathbb{P}[I(X \geq t)I(G=j)\mathbf{Z}(t)]$ uniformly in $t \in [0, \tau]$. Finally, arguments analogous to the proof of Glivenko-Cantelli Theorem for the empirical distribution function (e.g., Theorem 19.1 of van der Vaart, 1998), the bounded and monotonic conditions lead to $[\mathbb{P}_n\{Y_j(t)\}]^{-1} \xrightarrow{a.s.} [\mathbb{P}\{Y_j(t)\}]^{-1}$ uniformly in $t \in [0, \tau]$. Hence, we have $\bar{Z}_j(t) \xrightarrow{a.s.} \bar{z}_j(t)$ uniformly in $t \in [0, \tau]$.

Next, we prove that $n^{-\frac{1}{2}} \sum_{j=1}^J \sum_{i=1}^n M_{ij}(t)$ converges weakly to a zero-mean Gaussian process.

$$n^{-\frac{1}{2}} \sum_{j=1}^J \sum_{i=1}^n M_{ij}(t) = \mathbb{G}_n \left[\sum_{j=1}^J I(G=j) \left\{ I(X \leq t) \Delta - \int_0^t I(X \geq u) \{d\Lambda_{0j}(u) + \beta_0^T \mathbf{Z}(u) du\} \right\} \right],$$

where $\mathbb{G}_n$ is the empirical process defined by $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - \mathbb{P})f$ for a measurable function f . Since the VC class with finite index also satisfies the entropy conditions for the Donsker Theorem (van der Vaart and Wellner, 1996), $\{I(X \leq t), t \in [0, \tau]\}$ and $\{I(X \geq t), t \in [0, \tau]\}$ belong to some Donsker classes. The same holds for the stochastic process with bounded variation, $\{\mathbf{Z}(t), t \in [0, \tau]\}$, and the bounded monotonic stochastic process $\{\Lambda_{0j}(t), t \in [0, \tau]\}$. The set of functions $\{\beta^T \mathbf{Z}(t), t \in [0, \tau], \beta \in \mathcal{B}\}$ is also well-behaved and belongs to some Donsker class when the parameter space, $\mathcal{B}$, is compact. Finally, the class of functions of Lipschitz transformations of Donsker classes is a Donsker class. Therefore, with the various bounded conditions, we have that the class of functions

$$\sum_{j=1}^J I(G=j) \left[I(X \leq t) \Delta - \int_0^t I(X \geq u) \{d\Lambda_{0j}(u) + \beta^T \mathbf{Z}(u) du\} \right] \text{ indexed by } t \text{ and } \beta \text{ are}$$

Donsker classes and, hence, $n^{-\frac{1}{2}} \sum_{j=1}^J \sum_{i=1}^n M_{ij}(t)$ converges weakly to a zero-mean Gaussian process. It then follows from the Strong Embedding Theorem (Shorack and Wellner, 1986) and Lemma 1 of Lin et al. (2000) that (11) holds.

Note that (11) can be also derived through Lemma 19.24 of van der Vaart (1998). For instance, define the following random map:

$$\Psi(\beta, \eta) = \sum_{j=1}^J \int_0^\tau I(G=j) \eta(t) \left[dN_{ij}(t) - I(X \geq t) \{d\Lambda_{0j}(t) + \beta^T \mathbf{Z}(t) dt\} \right],$$

with $\eta(t) = \mathbf{Z}_j(t)$ and $\eta_0(t) = \mathbf{z}_j(t)$. Let $\mathcal{F}$ be the functional space $\{\Psi(\boldsymbol{\beta}, \eta) : \boldsymbol{\beta} \in \mathbf{B}, \eta \in \mathcal{H}\}$. The regularity conditions, (b), from Section 3 ensures that $\mathbb{P}_n[I(t)I(G=j)]$ is bounded away from zero, and hence, $\mathbf{Z}_j(t)$ has bounded variation. Therefore, the class of functions $\mathcal{F}$ is a Donsker class when $\mathbf{B}$ is compact. With the bounded condition and the conclusion that $\bar{\mathbf{Z}}_j(t) \xrightarrow{a.s.} \mathbf{z}_j(t)$ uniformly in $t \in [0, \tau]$, it follows from the Dominant Convergence Theorem that $\Psi(\boldsymbol{\beta}, \hat{\eta})$ converges to $\Psi(\boldsymbol{\beta}, \eta_0)$ in probability with a norm defined by

$d(\Psi(\cdot, \eta), \Psi(\cdot, \eta')) = \sup_{\boldsymbol{\beta} \in \mathbf{B}} \{\mathbb{P}(\Psi(\boldsymbol{\beta}, \eta) - \Psi(\boldsymbol{\beta}, \eta'))^2\}^{\frac{1}{2}}$. We then have all the conditions to apply Lemma 19.24 of van der Vaart (1998). Thus, $\mathbb{G}_n\{\Psi(\boldsymbol{\beta}, \hat{\eta}) - \Psi(\boldsymbol{\beta}, \eta_0)\} \xrightarrow{P} 0$ and (11) holds.

Formula (11) entails that $n^{-\frac{1}{2}}U(\boldsymbol{\beta}_0)$ is essentially a sum of independent and identically distributed random variables:

$$n^{-\frac{1}{2}}U(\boldsymbol{\beta}_0) = \mathbb{G}_n \left[\sum_{j=1}^J \{ \{ \mathbf{Z}(X) - \bar{\mathbf{z}}_j(X) \} I(G=j) \Delta - \int_0^\tau \{ \mathbf{Z}(t) - \bar{\mathbf{z}}_j(t) \} I(G=j) I(X \geq t) \{ d\boldsymbol{\Lambda}_0(t) + \boldsymbol{\beta}_0^T \mathbf{Z}(t) dt \} \} \right] + o_p(1).$$

Using the Multivariate Central Limit Theorem, the random vector $n^{-\frac{1}{2}}U(\boldsymbol{\beta}_0)$ asymptotically follows a p -variate normal distribution with mean zero. Similarly, we have that

$$\begin{aligned} n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= n^{-\frac{1}{2}} \hat{\mathbf{A}}^{-1} \sum_{j=1}^J \sum_{i=1}^n \int_0^\tau \{ \mathbf{Z}_i(t) - \bar{\mathbf{z}}_j(t) \} dM_{ij}(t) + o_p(1) \\ &= \mathbf{A}^{-1} \mathbb{G}_n \left[\sum_{j=1}^J \{ \{ \mathbf{Z}(X) - \bar{\mathbf{z}}_j(X) \} I(G=j) \Delta - \int_0^\tau \{ \mathbf{Z}(t) - \bar{\mathbf{z}}_j(t) \} I(G=j) I(X \geq t) \{ d\boldsymbol{\Lambda}_0(t) + \boldsymbol{\beta}_0^T \mathbf{Z}(t) dt \} \} \right] + o_p(1). \end{aligned}$$

By the Multivariate Central Limit Theorem and Slutsky's Theorem, Theorem 1 follows.

Appendix 2: Weak convergence of $\Lambda_0(t)$ and $\psi(t)$

Applying the Functional Delta Method (van der Vaart, 1998), we now consider the asymptotic properties of the random processes $n^{\frac{1}{2}}(\hat{\boldsymbol{\Lambda}}_0 - \boldsymbol{\Lambda}_0)$, where $\boldsymbol{\Lambda}_0 = [\Lambda_{01}, \dots, \Lambda_{0J}]^T$, with $\Lambda_{0j} = \{\Lambda_{0j}(t); t \in (0, \tau)\}$.

Lemma 2

$n^{\frac{1}{2}}(\hat{\boldsymbol{\Lambda}}_0 - \boldsymbol{\Lambda}_0)$ converges weakly to a zero-mean Gaussian process with a variance-covariance matrix between $n^{\frac{1}{2}}\{\hat{\Lambda}_{0j}(s) - \Lambda_{0j}(s)\}$ and $n^{\frac{1}{2}}\{\hat{\Lambda}_{0k}(t) - \Lambda_{0k}(t)\}$ given by $\sigma_{jk}(s, t) = E[\xi_{1j}(s)\xi_{1k}(t)]$, where

$$\xi_{ij}(s) = \int_0^s \frac{dM_{ij}(u)}{s_j^{(0)}(u)} - \{ \int_0^s \bar{\mathbf{z}}_j(u) du \} \mathbf{A}^{-1} \sum_{\ell=1}^J \int_0^\tau \{ \mathbf{Z}_i(u) - \bar{\mathbf{z}}_\ell(u) \} dM_{i\ell}(u). \quad (12)$$

The covariance function can be consistently estimated by $\hat{\sigma}_{jk}(s, t) = n^{-1} \sum_{i=1}^n \hat{\xi}_{ij}(s) \hat{\xi}_{ik}(t)$, with $\hat{\xi}_{ij}(s)$ obtained by replacing limiting values in $\xi_{ij}(s)$ with their empirical counterparts.

Applying the Functional Delta method again, we have

$$n^{\frac{1}{2}} \{ \hat{\psi}_{0j}(t) - \psi_{0j}(t) \} = \mathbf{H}_j^T(t) n^{\frac{1}{2}} \{ \hat{\Lambda}_0(t) - \Lambda_0(t) \} + o_p(1),$$

where $\mathbf{H}_j^T(t) = [H_{j1}(t), \dots, H_{jJ}(t)]$, with $H_{jk}(t) = \psi_{0j}(t) \psi_{0k}(t) p_k - I(k=j) \psi_{0j}(t)$ for $k = 1, \dots, J$.

We begin the proof by considering the large-sample behavior of the estimator of $\Lambda_{0j}(t)$. In particular,

$$\begin{aligned} & n^{\frac{1}{2}} \{ \hat{\Lambda}_{0j}(t) - \Lambda_{0j}(t) \} \\ &= n^{\frac{1}{2}} \{ \hat{\Lambda}_{0j}(t; \hat{\beta}) - \hat{\Lambda}_{0j}(t; \beta) \} + n^{\frac{1}{2}} \{ \hat{\Lambda}_{0j}(t; \beta) - \Lambda_{0j}(t) \} \\ &= n^{\frac{1}{2}} \int_0^t \frac{\sum_{i=1}^n dM_{ij}(u)}{\sum_{i=1}^n Y_{ij}(u)} - n^{\frac{1}{2}} \int_0^t \frac{\sum_{i=1}^n Y_{ij}(u) (\hat{\beta} - \beta_0)^T \mathbf{Z}_i(u) du}{\sum_{i=1}^n Y_{ij}(u)} \\ &= n^{\frac{1}{2}} \int_0^t \frac{\sum_{i=1}^n dM_{ij}(u)}{\mathbb{P}[I(\geq t) | I(G=j)]} - \{ \int_0^t \bar{\mathbf{z}}_j(u) du \} \mathbf{A}^{-1} \sum_{j=1}^J \int_0^t \{ \mathbf{Z}(u) - \bar{\mathbf{z}}_j(u) \} dM_j(u) + o_p(1) \\ &= \mathbb{G}_n \left[\frac{I(X \leq t) (G=j) \Delta - \int_0^t I(X \geq u) I(G=j) \{ \beta^T \mathbf{Z}_i(u) du + d\Lambda_{0j}(u) \}}{\mathbb{P}[I(X \geq t) | I(G=j)]} - \right. \\ & \quad \left. \{ \int_0^t \bar{\mathbf{z}}_j(u) du \} \mathbf{A}^{-1} \sum_{\ell=1}^J I(G=\ell) [\{ \mathbf{Z}(X) - \bar{\mathbf{z}}_\ell(X) \} \Delta - \right. \\ & \quad \left. \left. \int_0^t \{ \mathbf{Z}(u) - \bar{\mathbf{z}}_\ell(u) \} I(X \geq u) \{ \beta^T \mathbf{Z}_i(u) du + d\Lambda_{0\ell}(u) \} \right] \right] + o_p(1), \end{aligned}$$

where the third equality can be obtained by either Lemma 1 of Lin et al. (2000) or Lemma 19.24 of van der Vaart (1998). Applying the Donsker Theorem once again, Lemma 2 follows.

Note that Lemma 2 can also be established by Functional Delta Method. The proof is essentially the same as those used by Kosorok (2008) for the Nelson-Aalen estimator. Adapting the argument of Kosorok (2008), the center-specific cumulative hazard function can be formulated as:

$$\hat{\Lambda}_{0j}(t) = \int_0^t \frac{dB_{nj}(u)}{Y_{nj}(u)},$$

where

$$B_{nj}(t) = \mathbb{P}_n \left[\Delta I(G=j) I(X \leq t) - \int_0^t I(X \geq u) I(G=j) \hat{\beta}^T \mathbf{Z}(u) du \right]$$

and

$$Y_{nj}(t) = \mathbb{P}_n [I(X \geq t)I(G=j)].$$

After defining

$$\begin{aligned} B_{0j} &= \mathbb{P} \left[\Delta I(G=j)I(X \leq t) - \int_0^t I(X \geq u)I(G=j)\beta^T \mathbf{Z}(u)du \right] \\ &= \mathbb{P} \left[\int_0^t I(X \geq u)I(G=j)d\Lambda_{0j}(u) \right], \end{aligned}$$

we have

$$\begin{aligned} n^{\frac{1}{2}}\{B_{nj} - B_{0j}\} &= \mathbb{G}_n \left[\Delta I(G=j)I(X \leq t) - \int_0^t I(X \geq u)I(G=j)\beta^T \mathbf{Z}(u)du - \right. \\ &\quad \mathbb{P} \left[\int_0^t I(X \geq u)I(G=j)\mathbf{Z}(u)du \right] \mathbf{A}^{-1} \sum_{\ell=1}^J I(G=\ell) [\{\mathbf{Z}(X) - \bar{\mathbf{z}}_\ell(X)\} \Delta - \\ &\quad \left. \int_0^\tau \{\mathbf{Z}(u) - \bar{\mathbf{z}}_\ell(u)\} I(X \geq u) \{\beta^T \mathbf{Z}_i(u)du + d\Lambda_{0\ell}(u)\} \right] \Big] + o_p(1) \end{aligned}$$

and

$$n^{\frac{1}{2}}\{Y_{nj} - s_{0j}\} = \mathbb{G}_n [I(X \geq t)I(G=j)].$$

Using the Donsker Theorem, we have that the pair (B_{nj}, Y_{nj}) is asymptotically normal:

$$n^{\frac{1}{2}} \begin{pmatrix} B_{nj} - B_{0j} \\ Y_{nj} - s_{0j} \end{pmatrix} \Rightarrow \begin{pmatrix} \mathbb{G}_{1j} \\ \mathbb{G}_{2j} \end{pmatrix},$$

where $\mathbb{G}_{1j}$ and $\mathbb{G}_{2j}$ are tight Gaussian processes.

We are ready to show the asymptotic properties of $\Lambda_0(t)$. Note that the center-specific cumulative hazard function is constructed through the map $(A, B) \mapsto (A, \frac{1}{B}) \mapsto \int_0^t \frac{dA}{B}$. This composition map is Hadamard-differentiable (van der Vaart, 1998) on appropriate domains; e.g., the regularity conditions in our setting guarantee that B is bounded away from zero and A has uniformly bounded variation. By the Functional Delta Method (van der Vaart, 1998), we conclude that

$$n^{\frac{1}{2}}(\hat{\Lambda}_{0j} - \Lambda_{0j}) \Rightarrow \int_0^t \frac{d\mathbb{G}_{1j}}{s_{0j}} - \int_0^t \frac{\mathbb{G}_{2j} dB_{0j}}{s_{0j}^2},$$

which leads to Lemma 2. Applying Functional Delta Method once again, Theorem 1 follows.

Figure 1.

Analysis of SRTR Data: Point estimates $\hat{\psi}_{0j}(3)$ ($j = 1, \dots, 207$) and 95% confidence intervals, with $t = 3$ measured in years.

(a) Figure 2A

(b) Figure 2B

(c) Figure 2C

Figure 2.

Evaluation of center: $\hat{\psi}_{0j}(t)$ and 95% point-wise confidence intervals for $t \in [0, 5]$, for $j = 1$ (Fig 2A), $j = 2$ (Fig 2B), $j = 3$ (Fig 2C),

Figure 3.
Model Checking: Cox-Snell residual plots

Figure 4.
Histogram of Z Scores

Table 1

Performance of ratio of survival functions: $\psi_0(t)$; Setting 1: hazards and covariate are center independent; Sample size for each center=100; only odd-numbered centers are shown.

J	λ_j	% at risk	t	$\psi_0(t)$	BIAS	ESD	ASE	CP
1	0.50	76%	0.5	1.000	-0.000	0.055	0.053	0.94
		58%	1	1.000	-0.003	0.081	0.081	0.94
		44%	1.5	1.000	0.004	0.109	0.106	0.94
		33%	2	1.000	0.005	0.132	0.132	0.95
		17%	3	1.000	0.001	0.189	0.193	0.95
3	0.50	76%	0.5	1.000	-0.000	0.054	0.053	0.95
		58%	1	1.000	-0.001	0.081	0.081	0.94
		44%	1.5	1.000	-0.006	0.105	0.106	0.95
		33%	2	1.000	0.002	0.132	0.132	0.95
		17%	3	1.000	0.000	0.192	0.193	0.94
5	0.50	76%	0.5	1.000	0.004	0.054	0.053	0.94
		58%	1	1.000	-0.000	0.082	0.081	0.95
		44%	1.5	1.000	-0.008	0.105	0.106	0.94
		33%	2	1.000	-0.008	0.136	0.132	0.93
		17%	3	1.000	-0.008	0.194	0.192	0.95
7	0.50	76%	0.5	1.000	-0.002	0.051	0.053	0.96
		58%	1	1.000	-0.001	0.085	0.081	0.94
		44%	1.5	1.000	0.006	0.103	0.106	0.95
		33%	2	1.000	-0.005	0.131	0.132	0.95
		17%	3	1.000	-0.004	0.191	0.193	0.94
9	0.50	76%	0.5	1.000	-0.002	0.053	0.053	0.95
		58%	1	1.000	0.005	0.078	0.081	0.95
		44%	1.5	1.000	0.003	0.104	0.106	0.94
		33%	2	1.000	0.003	0.131	0.132	0.95
		17%	3	1.000	0.006	0.199	0.193	0.93

Table 2

Performance of ratio of survival functions: $\psi_0(t)$; Setting 2: hazards and covariate are center dependent; Sample size for each center=20; only odd-numbered centers are shown.

J	λ_j	% at risk	t	$\psi_0(t)$	BIAS	ESD	ASE	CP
1	0.45	78%	0.5	1.069	-0.015	0.114	0.115	0.92
		60%	1	1.140	-0.012	0.178	0.190	0.95
		44%	1.5	1.214	-0.033	0.253	0.268	0.94
		33%	2	1.291	-0.045	0.331	0.349	0.94
		18%	3	1.450	-0.065	0.485	0.524	0.91
3	0.51	76%	0.5	1.053	-0.007	0.113	0.116	0.92
		56%	1	1.110	-0.004	0.244	0.266	0.95
		40%	1.5	1.074	-0.001	0.233	0.242	0.96
		28%	2	1.145	-0.033	0.310	0.339	0.95
		14%	3	1.211	-0.063	0.424	0.493	0.93
5	0.57	73%	0.5	1.007	-0.006	0.120	0.121	0.92
		54%	1	1.011	-0.002	0.089	0.089	0.94
		37%	1.5	1.014	0.015	0.248	0.261	0.93
		26%	2	1.015	-0.012	0.302	0.326	0.93
		13%	3	1.012	-0.006	0.416	0.459	0.91
7	0.63	71%	0.5	0.977	0.013	0.121	0.122	0.91
		50%	1	0.952	-0.000	0.179	0.194	0.94
		33%	1.5	0.927	0.007	0.243	0.255	0.93
		21%	2	0.900	0.023	0.292	0.319	0.95
		8%	3	0.845	0.049	0.392	0.433	0.91
9	0.69	68%	0.5	0.948	0.002	0.124	0.127	0.92
		47%	1	0.896	0.009	0.191	0.195	0.93
		30%	1.5	0.847	0.003	0.253	0.251	0.92
		19%	2	0.799	0.030	0.291	0.307	0.93
		7%	3	0.706	0.066	0.392	0.405	0.89

Table 3

Performance of ratio of survival functions: $\psi_0(t)$; Setting 2: hazards and covariate are center dependent; Sample size for each center=50; only odd-numbered centers are shown.

J	λ_j	% at risk	t	$\psi_0(t)$	BIAS	ESD	ASE	CP
1	0.45	78%	0.5	1.069	-0.003	0.073	0.076	0.96
		60%	1	1.140	-0.004	0.120	0.122	0.96
		44%	1.5	1.214	-0.010	0.168	0.171	0.95
		33%	2	1.291	-0.028	0.220	0.224	0.94
3	0.51	18%	3	1.450	-0.050	0.346	0.357	0.94
		76%	0.5	1.053	-0.002	0.076	0.078	0.95
		56%	1	1.110	0.001	0.118	0.123	0.96
		40%	1.5	1.074	-0.011	0.161	0.170	0.96
5	0.57	28%	2	1.145	-0.017	0.213	0.219	0.94
		14%	3	1.211	-0.007	0.316	0.338	0.94
		73%	0.5	1.007	0.003	0.079	0.080	0.94
		54%	1	1.011	-0.004	0.120	0.125	0.95
7	0.63	37%	1.5	1.014	0.006	0.165	0.168	0.94
		26%	2	1.015	-0.001	0.208	0.213	0.94
		12%	3	1.012	0.007	0.319	0.318	0.94
		71%	0.5	0.977	0.005	0.079	0.083	0.95
9	0.69	50%	1	0.952	0.003	0.127	0.126	0.95
		33%	1.5	0.927	0.009	0.161	0.166	0.95
		21%	2	0.900	0.018	0.204	0.208	0.94
		8%	3	0.845	0.014	0.311	0.294	0.92
		68%	0.5	0.948	-0.000	0.085	0.085	0.94
		47%	1	0.896	0.006	0.127	0.126	0.95
		30%	1.5	0.847	0.009	0.162	0.164	0.95
		19%	2	0.799	0.007	0.203	0.200	0.94
		7%	3	0.706	0.015	0.282	0.274	0.92

Table 4

Performance of ratio of survival functions: $\psi_0(t)$; Setting 2: hazards and covariate are center dependent; Sample size for each center=100; only odd-numbered centers are shown.

J	λ_j	% at risk	t	$\psi_0(t)$	BIAS	ESD	ASE	CP
1	0.45	78%	0.5	1.069	-0.001	0.054	0.054	0.94
		60%	1	1.140	-0.005	0.085	0.087	0.96
		44%	1.5	1.214	-0.006	0.119	0.122	0.95
		33%	2	1.291	-0.008	0.155	0.162	0.96
		18%	3	1.450	-0.012	0.265	0.262	0.95
3	0.51	76%	0.5	1.037	-0.001	0.055	0.056	0.95
		56%	1	1.074	-0.001	0.090	0.088	0.95
		40%	1.5	1.110	-0.006	0.119	0.121	0.94
		28%	2	1.145	-0.003	0.157	0.158	0.95
		14%	3	1.211	-0.002	0.252	0.247	0.94
5	0.57	73%	0.5	1.007	0.003	0.056	0.058	0.95
		54%	1	1.011	-0.002	0.089	0.089	0.94
		37%	1.5	1.014	-0.002	0.121	0.120	0.94
		26%	2	1.015	-0.004	0.154	0.154	0.94
		12%	3	1.012	-0.016	0.233	0.231	0.94
7	0.63	71%	0.5	0.975	0.002	0.060	0.059	0.94
		50%	1	0.952	0.004	0.089	0.090	0.95
		33%	1.5	0.927	0.004	0.117	0.119	0.94
		21%	2	0.900	0.006	0.146	0.150	0.96
		8%	3	0.845	0.004	0.225	0.218	0.93
9	0.69	68%	0.5	0.948	-0.0034	0.061	0.061	0.96
		47%	1	0.896	-0.003	0.090	0.090	0.94
		30%	1.5	0.847	0.006	0.120	0.117	0.94
		19%	2	0.799	0.000	0.145	0.144	0.95
		7%	3	0.706	0.010	0.211	0.204	0.93

Table 5

Empirical significance levels (ESL) of supremum-based test for center effect (given in (10)); $n_j=100$; $t \in [0, 3]$.

Center	λ_j	ESL
1	0.5	0.063
2	0.5	0.052
3	0.5	0.068
4	0.5	0.058
5	0.5	0.053
6	0.5	0.079
7	0.5	0.037
8	0.5	0.058
9	0.5	0.051
10	0.5	0.063

Table 6Number and percentage of outlier centers (Significant: p-value $\leq 5\%$)

Method	Non-significant	Significantly-total
Without Empirical Null	169 (81.6%)	38 (18.4%)
With Empirical Null	196 (94.7%)	11 (5.3%)