Semi-parametric and Parametric Methods for the Analysis of Multi-center Survival Data

by

Zhi He

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2012

Doctoral Committee:

Associate Professor Douglas E. Schaubel, Chair
Professor John D. Kalbfleisch
Professor Hal Morgenstern
Assistant Professor Min Zhang

To Mia, Will, Andy and my parents

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

Introduction

Survival analysis plays an important role in the evaluation of center effects. In this dissertation, we focus on the development of parametric and semi-parametric methods to evaluate centers in the presence of censored data. The data which motivate the proposed methods arise in the kidney transplant setting. Kidney transplant centers may use a variety of surgical protocols, and such center-specific practices may lead to significantly different post-transplant outcomes across centers. It is important to correctly identify those centers with significantly lower mortality or higher mortality. For instance, specific procedures in centers with significantly lower survival need to be corrected or updated. In contrast, various procedures in centers with significantly greater survival can be recommended to other centers. Given the high stakes of center-specific evaluations, it is important that the statistical methods used for identifying outlying centers be accurate.

It is well known that factors may vary considerably across centers, including both patient characteristics and medical practices. These factors can have substantial impact on the expected outcomes at a given center. A fair comparison of center-specific medical practices needs to adequately account for the imbalance of patient characteristics among centers. Otherwise, the inclusion of some high-risk patients

in a particular center can artificially make that center's survival appear poor. The establishment of appropriate methods for analysis of center-specific outcomes has received a lot of attention in the biostatistical literature. Breslow et al. (1983) studied a fully parametric multiplicative model assuming that the death intensity for subject i, $\lambda_i(t)$, is the product of known standard mortality $\lambda_0(t)$ and a factor consisting of a covariate vector,

$$\lambda_i(t) = \lambda_0(t)\exp(\beta'Z_i),$$

where β is an unknown parameter vector. Andersen et al. (1985) introduced a semi-parametric multiplicative regression model with known standard mortality, $\lambda_0(t)$, and unknown relative mortality, $\theta(t)$,

$$(1.1) \qquad\qquad \lambda_i(t) = \lambda_0(t)\theta(t)\exp(\beta'Z_i),$$

using a model which is essentially a special case of the Cox (1972) model. If the model has no adjustment variables, then the death rate at time t is a product of a known standard mortality $\lambda_0(t)$ and an unknown relative mortality $\theta(t)$, as in the approach of Andersen and Væth (1989). In addition, if $\theta(t) = \theta$, a constant, then $\widehat{\theta}$ is the standardized mortality ratio (SMR), defined as the ratio of the observed number of deaths to the number expected in the selected population (if the death rates in the selected population were equivalent to those in the standard population).

Each of the aforementioned methods require a known covariate-specific standard mortality. To relax this requirement, the unstratified Cox model has been used to generate the semiparametric generalization of the SMR, which we here call the Cox SMR. However, the basic properties and implicit assumptions of this Cox SMR are not well understood. Failure to appreciate the limitations of the Cox SMR may lead to the inappropriate use of this measure. To address these concerns, in

Chapter 2, we first provide a rigorous examination of the assumptions required for valid inference through the Cox SMR. We then develop a modification of the Cox SMR based on a stratified Cox model, which remedies some significant limitations of the typically employed unstratified version. We also propose a semiparametric generalization of direct standardization through a Cox model based standardized rate ratio (SRR). The proposed measures are process-based and therefore allow one to not only identify whether the mortality associated with a center is outlying, but also determine when during follow-up the excess mortality tends to occur. Each of the measures we consider are jump-processes due to the nonparametric nature of the cumulative baseline hazard estimator. As such, we develop a smoothed version of the time-dependent SRR, which should generally be more appealing to investigators than a step function. Finally, we propose hypothesis testing procedures to evaluate whether a given center's SRR is constant across follow-up time.

A variety of methodological approaches has been used to evaluate centers. The most frequently used methods rely on statistical comparison of observed versus expected outcomes. The estimator for the expected number of events, which we discussed in Chapter 2, is one such example. An alternative measure is difference in average restricted mean lifetime; e.g., Karrison (1987), Zucker (1998). It is worth noting that existing methods using restricted mean lifetime focus on short- and medium-term clinical outcomes. This is for good reason, since they are relatively easy to measure, and presumably most directly related to stratum-specific performance. However, these measures may fail to capture the long-term effects that reveal the quality of stratum. This is highly relevant in the case of kidney transplants, because the majority of mortality and graft failures occur in the long-term. Long term center-specific profiling may offer important insight into variation in processes with

has an interpretation at the individual level, which is perhaps more relevant to a patient than estimators based on standardized or averaging. Under an additive hazard model, the ratio of baseline survival functions is invariant to the choice of baseline covariate level. That means, the ratio of survival functions represents the contrast between subject i at center j versus subject i at the hypothetical center with baseline hazard function equal to the national average; where subject i can have any covariate value. To implement the proposed estimator, we propose to use a stratified additive model. An interesting feature of the methods proposed in this chapter is the use of an additive hazard model to generate a center effect measure based on relative survival.

For each of the proposed methods in Chapter 2, 3 and 4, we derive the asymptotic properties of the center effect estimators. Finite sample properties are assessed through simulation. Each method is applied to kidney transplant data obtained from a national registry.

CHAPTER II

Modifications of and Alternatives to the Standardized Mortality Ratio in Evaluating Center-Specific Mortality

The standardized mortality ratio (SMR) based on a Cox regression model is often used to evaluate center-specific mortality. However, the asymptotic properties and finite-sample behavior of the Cox SMR are not well-studied. In the first chapter, we describe some strong limitations of the Cox SMR that relate to its underlying assumptions. To address these limitations, we then develop modifications to the Cox SMR based on a stratified Cox model. In addition, since center effects computed through indirect standardization are not comparable, we propose a semiparametric generalization of direct standardization. The measures we consider are process-based and, therefore, allow us to not only identify if a center's mortality is outlying, but also when during the follow-up the excess mortality tends to occur. Kernel-smoothing is performed for the proposed measures. Hypothesis testing procedures are developed to identify outlying centers and to evaluate whether a particular center has an effect that is constant over time. Asymptotic properties of proposed estimators are derived, with finite-sample properties examined through an extensive simulation study. The methods considered and developed are applied to national kidney transplant data.

2.1 Introduction

Data from clinical and epidemiologic studies are frequently derived from multiple centers. In such multicenter studies, time to an event is often of interest and may not be observed due to random loss to follow-up or administrative censoring. Indeed, the evaluation of center effects frequently plays an important role in survival analysis. For instance, such evaluations could be used to compare mortality among patients from a specific region or center to that in the general population. A frequently used tool for such purposes is the standardized mortality ratio (SMR), defined as the ratio of the observed number of deaths to the number expected in the selected population if the death rates in the selected population were equivalent to those in the standard population. Existing methods of comparing center-specific mortality typically require a table of mortality rates to serve as a standard population. To relax the assumption that standard population mortality rates are known, a semiparametric generalization of the SMR (which we refer to as the Cox SMR) has been adopted based on the well-known Cox regression model (Cox (1972)). Despite its popularity, the implicit assumptions of the Cox SMR are not well understood. Moreover, large-sample properties of the Cox SMR have not been well-studied, and simulation studies are quite limited. In this report, we first carry out a detailed examination of the Cox SMR. We then propose and evaluate alternative methods which remedy the most important limitations of the measure.

Our interest in quantifying center effects is motivated by the study of post-transplant mortality for end-stage renal disease (ESRD) patients. It is well known that post-transplant mortality may differ significantly across centers. To quantify such differences (with a view to ultimately improving patient care), the compari-

son of mortality rates by center is a commonly used strategy. Centers with higher mortality rates relative to the national average could be targeted for more frequent surveillance and, if improvement is not subsequently observed, could ultimately lose their accreditation. Given the high stakes of such evaluations, it is crucial that the measurements used to classify centers be accurate. In this sense, it is not surprising that the appropriateness of methods used for quantifying center effects has been a topic of ongoing debate. Specifically, discussion has focused on whether the SMR is a defensible measurement for studying center effects (Berry (1983), Hazel (2005), Lacson et al. (2001)). Further study of the SMR is required and, in this report, we pay particular attention to the Cox SMR.

The SMR is an example of indirect standardization, a method that has been widely used in epidemiology and sociology to compare mortality among subpopulation. Another commonly used approach in such comparisons is direct standardization, usually through a measure termed the Standardized Rate Ratio (SRR) or Comparative Mortality Figure (CMF). One can express the SRR as the ratio of expected to observed numbers of deaths in the whole study population; the numerator of the SRR represents the expected number of deaths if all patients were treated at the given center, while the denominator equals the total observed number of deaths in the study population. Both indirect standardization and direct standardization were among the earliest statistical techniques developed, as noted in the review of Keiding (1987). Breslow and Day (1985) compared the strengths and weaknesses of these two methods in the framework of person-year methods. The main disadvantage of the SMR is that ratios of SMRs for two groups do not necessarily summarize the ratios of their component covariate-specific rates, a phenomenon analogous to Simpson (1951) Paradox. Essentially, the SMR's comparison of observed to expected

mortality involves center-specific standardization. Hence, SMRs for different centers are not directly comparable since they are standardized to different populations. In contrast, this drawback is not inherent to the SRR, since the same standard population is applied to all centers. However, a criticism of the SRR is that it generally has a larger standard error than the SMR, especially when the sample size is small.

Originally, the SMR was applied in settings where standard mortality rates were known rather than estimated. For example, Wolfe et al. (1992) calculated SMRs among ESRD patients using mortality tables published by the United States Renal Data Systems. Breslow and Langholz (1987) proposed nonparametric estimation of the SMR as a continuous function of time. Andersen et al. (1999) constructed a test of whether the mortality rate in a population is the same as that of the reference population over a given time interval. It is worth noting that these standardization methods involving expected deaths are closely connected to the regression models for relative mortality; e.g., Breslow et al. (1983), Andersen et al. (1985), Andersen and Væth (1989), Andersen et al. (1993).

Each method referenced in the previous paragraph requires that standard mortality rates are known. It is desirable to relax this requirement since the estimation of standard mortality is often required. Moreover, when standard mortality is estimated, accurate inference generally entails incorporating the variability in the standard rates into the estimated standard errors of the center effects. To relax the assumption of known standard mortality rates, the Cox model has been used to generate the semiparametric generalization of standardized mortality ratio, which we here call the Cox SMR. However, the basic properties and implicit assumptions of this Cox SMR are not well understood. In particular, this indirect standardization has not been validated by comparing it with direct standardization in the framework

of semiparametric models. In consequence, failure to appreciate the limitation of Cox SMR may lead to the inappropriate use of this measure.

Another challenging aspect of the evaluation of center-specific mortality is that the center effects often depend on follow-up time; in other words, the center effects are non-proportional. However, most existing methods of comparing center-specific mortality (including the Cox SMR) rely on the proportional hazards assumption. A more appropriate analysis needs to account for the non-proportionality, an issue we wish to address.

In this report, we first provide a rigorous examination of the assumptions required for valid inference through the Cox SMR. We then develop a modification of the Cox SMR based on the stratified Cox model, which remedies some significant limitations of the typically employed unstratified version. We also propose a semiparametric generalization of direct standardization (through a Cox model based SRR). The measures are process-based and allow us not only to identify whether the mortality associated with a center is outlying, but also when during the follow-up the excess mortality tends to occur. Each of the measures we consider are jump-processes due to the nonparametric nature of the cumulative baseline hazard estimator. As such, we develop a smoothed version of the time-dependent SRR, which should generally be more appealing to investigators than a step function. Finally, we propose hypothesis testing procedures to evaluate whether a given center's SRR is constant across follow-up time.

The remainder of this chapter is organized as follows. The Cox SMR, its proposed modifications, as well as the proposed SRR are described in the next section. Asymptotic properties for each estimator are derived in Section 3. Finite-sample properties are examined in Section 4 through extensive simulation studies. Section

5 applies the proposed methods to Canadian kidney transplant data. We provide some discussion of the proposed and related methods in Section 6.

2.2 Methods

First, we provide the notation to be used in this article. Let T_i and C_i represent the survival and censoring time, respectively, for the i'th patient, where $i = 1, ..., n$. 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 1 when condition A holds and 0 otherwise. The observed death indicators are denoted by $\Delta_i = I(T_i \leq C_i)$, and the death counting process is defined as $N_i(t) = \Delta_i I(X_i \leq t)$. 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, $(X_i, \Delta_i, G_i, Z_i)$, where Z_i is a vector of adjustment covariates.

2.2.1 Standardized Mortality Ratio (SMR)

To introduce the Cox SMR, we consider a semiparametric multiplicative model

$$(2.1) \qquad \lambda_{ij}(t) = \lambda_0(t)\theta_j \exp(\beta^T Z_i)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function common for all centers, and θ_j is an unknown constant for relative center effect. In view of the martingale,

$$M_{ij}(\tau) = N_{ij}(\tau) - \int_0^\tau Y_{ij}(u) \exp(\beta^T Z_i)\theta_j d\Lambda_0(u),$$

it is natural to estimate θ_j by the semiparametric generalization of the standardized mortality ratio (Cox SMR),

$$\text{(2.2)} \qquad \widehat{SMR}_j = \frac{O_j(\tau)}{E_j(\tau)}$$

$$O_j(\tau) = \sum_{i=1}^{n} N_{ij}(\tau)$$

$$E_j(\tau) = \sum_{i=1}^{n} \int_0^{\tau} Y_{ij}(u) \exp\{\widehat{\beta}^T Z_i\} d\widehat{\Lambda}_0(u; \widehat{\beta})$$

where the quantity τ satisfies $P(X_i \geq \tau) > 0$ and can be set to the maximum observation time (such that all observed events are included in the analysis). The SMR defined above has an interpretation easily understood by clinical investigators: the ratio of the observed number of events for center j to the expected number of events for center j (if center j subjects belonged to a hypothetical center with hazard function equal to the national average).

Traditionally, the Cox SMR is calculated from an unstratified Cox model,

$$\text{(2.3)} \qquad \lambda_i(t) = \lambda_0(t) \exp\{\beta^T Z_i\}.$$

The partial likelihood (Cox (1975)) estimator of β is denoted by $\widehat{\beta}$ and is given by the solution to $U(\beta) = 0$, where 0 is a vector of zeros,

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i - \overline{Z}(u; \beta)\} dN_i(u),$$

where $\overline{Z}(u; \beta) = S^{(0)}(u; \beta)^{-1} S^{(1)}(u; \beta)$ and $S^{(d)}(u; \beta) = n^{-1} \sum_{i=1}^{n} Y_i(u) Z_i^{\otimes d} \exp\{\beta^T Z_i\}$ for $d = 0, 1, 2$. The cumulative baseline hazard can be estimated by the Breslow (1972) estimator,

$$\text{(2.4)} \qquad \widehat{\Lambda}_0(t; \widehat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{t} \frac{dN_i(u)}{S^{(0)}(u; \widehat{\beta})}.$$

In the denominator of the Cox SMR, the expected number of events is calculated based on (2.3). The validity of this model requires an important assumption:

$$\text{(A.1)} \qquad \lambda_i(t|Z_i, \ G_i) = \lambda_i(t|Z_i).$$

Under assumption (A.1), the hazards are conditionally independent of center given the covariates. Only under assumption (A.1) is $\widehat{\beta}$ from the unstratified Cox model unbiased; otherwise, this is not the case. It is quite common for (A.1) to fail in practice. The motivation for carrying out center-specific evaluations (through the SMR or otherwise) is often that mortality is suspected to differ by center, perhaps greatly. For example, in the data set which motivated our current work, it is suspected that mortality following kidney transplantation may differ significantly across Canadian transplant centers. In such cases, the failure to include center effects in model (2.3) produces biased parameter estimates (i.e., even if Z_i is independent of G_i) because the Cox model employs a non-linear link function. As a consequence, $\widehat{\beta}$ converges to an unknown vector $\beta_* \neq \beta$. Thus, the expected number in the denominator of the Cox SMR would not be accurate. The corresponding standard error estimators and confidence intervals would also be invalid. Moreover, since $\widehat{\beta}$ is biased, so will be $\widehat{\Lambda}_0(t; \widehat{\beta})$ since, as implied by (2.4), the consistency of the baseline hazard estimator depends on that of $\widehat{\beta}$.

2.2.2   SMR based on Stratified Model: SMR*

To address the important limitations identified in the preceding subsection, we now propose a modification of the Cox SMR, which we refer to as SMR*, which is based on the center-stratified Cox model,

$$\text{(2.5)} \qquad \lambda_{ij}(t) \equiv \lambda_i(t|G_i = j) = \lambda_{0j}(t) \exp\{\beta^T Z_i\},$$

where $\lambda_{0j}(t)$ is an unspecified center-specific baseline hazard function and β is a parameter vector. Under this stratified model, proportionality is not assumed to hold across centers, and is assumed only with respect to the adjustment covariates. The partial likelihood estimator of β is denoted by $\widehat{\beta}_\star$ and is given by the solution to $U_\star(\beta) = 0$ where

$$U_\star(\beta) = \sum_{j=1}^{J} \sum_{i=1}^{n} \int_0^\tau \{Z_{ij} - \overline{Z}_j(u; \beta)\} dN_{ij}(u)$$

with $\overline{Z}_j(u; \beta) = S_j^{(0)}(u; \beta)^{-1} S_j^{(1)}(u; \beta)$ and $S_j^{(d)}(u; \beta) = n^{-1} \sum_{i=1}^{n} Y_{ij}(u) Z_i^{\otimes d} \exp\{\beta^T Z_i\}$ for $d = 0, 1, 2$. To estimate $\text{SMR}_j^\star$, we first estimate β, then we use $\widehat{\beta}_\star$ to replace the non-stratified coefficient $\widehat{\beta}$ in formula (2.2). The $\text{SMR}_j^\star$ is then estimated by

$$(2.6) \qquad \widehat{SMR}_j^\star = \frac{O_j(\tau)}{E_j^\star(\tau)}$$

where $O_j(\tau)$ is as previously defined in (2.2) and

$$E_j^\star(\tau) = \sum_{i=1}^{n} \int_0^\tau Y_{ij}(u) \exp\{\widehat{\beta}_\star^T Z_i\} d\widehat{\Lambda}_0(u; \widehat{\beta}_\star)$$

where $\widehat{\Lambda}_0(t; \widehat{\beta}_\star)$ represents an estimator of the average cumulative baseline hazard,

$$(2.7) \qquad \widehat{\Lambda}_0(t; \widehat{\beta}_\star) = \sum_{j=1}^{J} \int_0^t S_j^{(0)}(u; \widehat{\beta}_\star) S^{(0)}(u; \widehat{\beta}_\star)^{-1} d\widehat{\Lambda}_{0j}(u; \widehat{\beta}_\star),$$

with the center-specific cumulative hazards are estimated by

$$(2.8) \qquad \widehat{\Lambda}_{0j}(t; \widehat{\beta}_\star) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{dN_{ij}(u)}{S_j^{(0)}(u; \widehat{\beta}_\star)}.$$

Two important aspects of $\widehat{SMR}_j^\star$ are worth noting. First, $\widehat{\beta}_\star$ is unbiased even if assumption (A.1) fails and, as a consequence, hazards are allowed to be center-dependent. Second, $\widehat{\Lambda}_0(t; \widehat{\beta}_\star)$ is employed in (2.6) since using $\widehat{\Lambda}_{0j}(t; \widehat{\beta}_\star)$ would result in $\widehat{SMR}_j^\star \equiv 1$ for all j.

2.2.3 Time-dependent Standardized Mortality Ratio

The measures defined in (2.2) and (2.6) provide a one-number summary calculated at time $t = \tau$. An advantage of this approach is using all observed events and, thus in a sense, making maximum use of the available data. However, a potential limitation is that each measure reflects only the average center effect over the follow-up time period. In practice, the center effects may change over follow-up time (i.e., non-proportionality, in the context of Cox regression). Therefore, it is often of interest to view each center effect as a process over time. This is particularly relevant in long-term follow-up studies where there may be considerable changes in the mortality pattern within a given center. For example, in the context of our motivating example, it is possible that a transplant center could have a substandard surgery protocol (and hence have relatively high death rates early in the post-transplant follow-up period), but also have high quality follow-up strategies (which result 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. If the hazard functions cross, however, the Cox SMR may fail to detect any difference. To address this problem, a natural solution is to use a time-dependent version of the Cox SMR, which we define as

$$(2.9) \qquad \widehat{SMR}_j(t) = \frac{O_j(t)}{E_j(t)},$$

where $0 < t \leq \tau$. Similarly, the time-dependent SMR* is obtained by

$$(2.10) \qquad \widehat{SMR}_j^\star(t) = \frac{O_j(t)}{E_j^\star(t)}.$$

Each of the estimators in (2.9) and (2.10) can be used to identify both if a center's mortality is outlying, and determine when during the follow-up period the excess

mortality tends to occur. For the reasons described in Section 2.2, $\widehat{SMR}^{\star}_{j}(t)$ would be a better choice than $\widehat{SMR}_{j}(t)$.

2.2.4 Direct Standardization: Standardized Rate Ratio (SRR)

The SMR-type estimators defined thus far are based on indirect standardization. Each of $\mathrm{SMR}_{j}(t)$ and $\mathrm{SMR}^{\star}_{j}(t)$ can be interpreted as the ratio of observed to expected number of deaths for patients treated at center j; where the expected pertains to the average number of deaths at a hypothetical center with the national average hazard function. Estimators based on indirect standardization are generally easily understood by non-statisticians.

It is worth noting, however, that the estimators based on indirect standardization can be viewed as a weighted ratio of center-specific cumulative hazards, with weight functions based on center-specific $S^{(0)}_{j}(t;\beta)$. These weight functions have an obvious disadvantage: they involve center-specific censoring and covariate distributions, which can differ considerably across centers. It is well known that two centers with equal covariate-specific mortality hazards could have different SMRs, merely due to differences in their respective covariate distributions. What is perhaps less known is that, in the presence of censored data, the same phenomenon can occur due to differences in center-specific censoring distributions.

Referring again to our motivating data, there is a certain appeal of an estimator that pertains to patients actually treated at center j. For example, if the goal is to determine whether a center has higher or lower mortality rates than expected, $\mathrm{SMR}^{\star}_{j}(t)$ is still a useful measure for evaluating center j in isolation; which may indeed be the goal of surgeons and clinicians at center j. However, for a governing body (e.g., oversight committee), $\mathrm{SMR}^{\star}_{j}(t)$ is much less useful. In particular, a set of $\mathrm{SMR}^{\star}_{j}(t)$

estimators cannot be validly ranked to determine an ordering of centers, since each estimator is essentially adjusted using different covariate and censoring distribution. These factors can have a substantial impact on the weight function $S_j^{(0)}(t; \beta)$. Based on results from indirect methods, we cannot conclude that center j actually has higher mortality than center j' if the covariate and/or censoring distributions in the two centers are different. To rule out the possibility that differences among centers are due merely to different censoring and/or covariate distributions, the weight function should be specified such that differences among centers with respect to the resulting measure are a function only of corresponding differences in center-specific hazards.

Motivated by such considerations, we propose an alternative method, referred to as the Standardized Rate Ratio (SRR), which can be interpreted as a semiparametric generalization of direct standardization

$$(2.11) \qquad \widehat{SRR}_j(t) = \frac{\mathcal{E}_j(t)}{O(t)}$$

with $O(t) = \sum_{j=1}^{J} O_j(t)$ being the total observed number of deaths in the population; and where we define

$$(2.12) \qquad \mathcal{E}_j(t) = \sum_{\ell=1}^{J} \sum_{i=1}^{n} \int_0^t Y_{i\ell}(u) \exp\{\widehat{\beta}_\star^T Z_i\} d\widehat{\Lambda}_{0j}(u)$$

to represent the total expected number of deaths up to time t if all patients in the population were treated at center j. Note that, although the estimation of $SRR_j(t)$ also involves censoring and covariate distributions, the same weight function is applied across all centers; thus factoring out the impact of imbalances in center-specific censoring and covariate distributions.

2.2.5 Smoothed SRR Estimator

Each of the estimators described in this section is a step function with jumps at observed death times. This property of the estimators may be difficult for non-

statisticians to understand and thus trust. Following Ramlau-Hansen (1983), we can compute a kernel-smoothed version of the SRR. In particular, we define the density-type estimator $\widetilde{f}(t;b)$ as:

$$\widetilde{f}(t;b) = \frac{1}{b}\int_0^\tau \kappa\left(\frac{t-u}{b}\right) d\,\widehat{SRR}(u),$$

where κ is the optimal kernel function (Epanechikov Kernel),

$$\kappa(y) = \frac{3}{4}(1-y^2)I\{|y| < 1\}.$$

and b is a bandwidth. The kernel estimate $\widetilde{SRR}(t;b)$ is then obtained by integrating the density-type estimator $\widetilde{f}(t;b)$ accordingly. For the motivating data set, we tried several fixed bandwidths. Among them, one tenth of the range of event times was chosen as the bandwidth to give a desired degree of smoothness. More formal methods for choosing b include cross-validation techniques.

2.2.6 Tests for Constant Center Effect

In preceding subsections, we have advocated the use of time-dependent center effect measures. In the interests of parsimony, it is natural to question whether the center effect is constant over time. This issue could be examined using techniques similar to those used to derive simultaneous confidence bands (Lin et al. (1994, 2000)). We propose Kolmogorov-Smirnov-type statistics based on the SRR to evaluate whether center effects are constant over time, for which the null hypothesis is $H_0 : SRR_j(t) = SRR_j$. The test statistic is given by

$$(2.13) \qquad \widehat{T}_j = \sup_{0 \le t \le \tau} \left| n^{\frac{1}{2}}\left\{(\widehat{SRR}_j(t) - \widehat{SRR}_j(\tau)\right\}\right|.$$

The advantage of this test is that its implementation is straightforward. The reference value, $SRR_j(\tau)$, uses all the information until time τ; and it will be the same

as the average SRR if the center effects are constants over time. An alternative test statistic is,

$$\widehat{F}_j = \sup_{0 \le t \le \tau} \left| n^{\frac{1}{2}} \left\{ (\widehat{SRR}_j(t) - \int_0^\tau \widehat{SRR}_j(u)du/\tau \right\} \right|,$$

for which the reference value is perhaps more general, but the computation is much more complicated. We will explore these tests further in the following sections.

2.3 Asymptotic Properties

We now derive the asymptotic properties for each of the measures investigated thus far. Proofs are provided in the Web Appendices.

To derive the large-sample properties for the Cox SMR, we impose the following regularity conditions under the unstratified Cox model.

(a) $(X_i, \Delta_i, G_i, Z_i)$ are independent and identically distributed random vectors.

(b) $P(X_i \ge \tau) > 0$ where τ is a pre-specified time point.

(c) Z_{ik} have bounded total variation, i.e., $|Z_{ik}| < \kappa$ for all $i = 1, ...n$ and $k = 1, ..., p$, where κ is a constant and Z_{ik} is the kth component of Z_i.

(d) $\int_0^\tau \lambda(t)dt < \infty$.

(e) Continuity of the following functions:

$$s^{(1)}(t; \beta) = \frac{\partial}{\partial \beta} s^{(0)}(t; \beta), \ \ s^{(2)}(t; \beta) = \frac{\partial^2}{\partial \beta \partial \beta^T} s^{(0)}(t; \beta)$$

where $s^{(d)}(t; \beta)$ is the limiting value of $S^{(d)}(t; \beta)$ for $d = 0, 1, 2$, with $s^{(1)}(t; \beta)$ and $s^{(2)}(t; \beta)$ bounded and $s^{(0)}(t; \beta)$ bounded away from 0 for $t \in [0, \tau]$ and β in an open set.

The covariance function can be consistently estimated by $\widehat{\sigma}_j(s,t;\widehat{\beta})$, where $\widehat{\sigma}_j(s,t;\widehat{\beta}) = n^{-1}\sum_{i=1}^n \widehat{\xi}_{ij}(s;\widehat{\beta})\widehat{\xi}_{ij}(t;\widehat{\beta})$, with $\widehat{\xi}_{ij}(t;\widehat{\beta})$ obtained by replacing limiting values in $\xi_{ij}(t;\beta)$ with their empirical counterparts. With respect to the proof of Theorem 1, through various results from empirical processes (Bilias et al. (1997); Lin et al. (2000); Pollard (1990)), the process $n^{\frac{1}{2}}\{\widehat{SMR}_j(t) - 1\}$ can be shown to be asymptotically equivalent to $n^{-\frac{1}{2}}\sum_{i=1}^n \xi_{ij}(t;\beta)$. A demonstration of tightness then completes the proof. The consistency of the covariance estimators is established using the Uniform Strong Law of Large Numbers. Details of the proof are provided as follow.

$$\widehat{SMR}_j(t) \equiv \frac{O_j(t)}{E_j(t)} = \frac{\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_{0j}(u)}{\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_0(u)}$$

Based on the uniform consistency, boundedness and monotonicity of $\widehat{\Lambda}_0(u)$, together with condition (e), we can apply Lemma 1 of Lin et al (2000) to prove that $\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_{0j}(u)$ converges almost surely to $\int_0^t s_j^{(0)}(u;\beta)d\Lambda_{0j}(u)$ uniformly in t for $t \in [0,\tau]$. Similarly, $\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_0(u)$ converges almost surely to $\int_0^t s_j^{(0)}(u;\beta)d\Lambda_0(u)$ uniformly in t for $t \in [0,\tau]$. Under the assumption (A.1) defined in the main paper and using stochastic equicontinuity, $\widehat{SMR}_j(t)$ then converges almost surely to 1 uniformly in t for $t \in [0,\tau]$.

To prove the weak convergence, we decompose the quantity as follows

$$(2.14) \quad n^{\frac{1}{2}}\left\{\frac{O_j(t)}{E_j(t)} - 1\right\} = n^{\frac{1}{2}}\left\{\frac{\sum_{i=1}^n \int_0^t dN_{ij}(u)}{n\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_0(u)} - \frac{\int_0^t S_j^{(0)}(u;\beta)d\Lambda_0(u)}{\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_0(u)}\right\}$$

$$(2.15) \quad + n^{\frac{1}{2}}\left\{\frac{\int_0^t S_j^{(0)}(u;\beta)d\Lambda_0(u)}{\int_0^t S_j^{(0)}(u;\widehat{\beta})d\widehat{\Lambda}_0(u)} - \frac{\int_0^t S_j^{(0)}(u;\beta)d\Lambda_0(u)}{\int_0^t S_j^{(0)}(u;\beta)d\Lambda_0(u)}\right\}$$

By the strong law of large number (SLLN), $S_j^{(0)}(u;\beta) \xrightarrow{a.s.} s_j^{(0)}(u;\beta)$. This, combined with the fact that $\widehat{\Lambda}_0(u) \xrightarrow{a.s.} \Lambda_0(u)$ (Anderson and Gill, 1982) results in that

$$(1) = \frac{n^{-\frac{1}{2}}\sum_{i=1}^n \int_0^t dM_{ij}(u;\beta)}{\int_0^t s_j^{(0)}(u;\beta)d\Lambda_0(u)} + o_p(1)$$

Through a Taylor expansion,

$$(2) = -\left\{ \int_0^t s_j^{(0)}(u;\beta) d\Lambda_0(u) \right\}^{-1} n^{\frac{1}{2}} \left\{ \int_0^t S_j^{(0)}(u;\widehat{\beta}) d\widehat{\Lambda}_0(u) - \int_0^t s_j^{(0)}(u;\beta) d\Lambda_0(u) \right\} + o_p(1),$$

where

$$n^{\frac{1}{2}} \left\{ \int_0^t S_j^{(0)}(u;\widehat{\beta}) d\widehat{\Lambda}_0(u) - \int_0^t s_j^{(0)}(u;\beta) d\Lambda_0(u) \right\}$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^t \frac{s_j^{(0)}(u;\beta)}{s^{(0)}(u;\beta)} dM_i(u;\beta)$$

$$+ \int_0^t \left\{ s_j^{(1)}(u;\beta) - s_j^{(0)}(u;\beta)\bar{z}(u;\beta) \right\} d\Lambda_0(u) n^{\frac{1}{2}}(\widehat{\beta} - \beta) + o_p(1)$$

with $M_i(u;\beta) = \sum_{\ell=1}^J M_{i\ell}(u;\beta)$ and

$$n^{\frac{1}{2}}(\widehat{\beta} - \beta) = \Omega(\beta)^{-1} n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau [Z_i - \bar{z}(u;\beta)] dM_i(u;\beta) + o_p(1)$$

Therefore, the process $n^{\frac{1}{2}}\{\widehat{SMR}_j - 1\}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^n \xi_{ij}(t;\beta)$ with $\xi_{ij}(t;\beta)$ defined in the main paper. To complete the proof, we now focus on the tightness. It is obvious that $M_i(t)$ is the difference of two monotone functions in t. The boundedness conditions of Z_i, $S_j^{(0)}$ and $\widehat{\Lambda}_0$ enable us to prove that (1) and (2) are also differences of two monotone functions in t with pseudodimension 1, which suffices to prove that they are manageable (Pollard 1990). It then follows that the tightness holds, and thus, $n^{\frac{1}{2}}\{\widehat{SMR}_j - 1\}$ converges weakly to a zero-mean Gaussian process. The consistency of the covariance estimators is established using the Uniform Strong Law of Large Number.

To derive the large-sample properties for SMR* and SRR, we impose the following regularity conditions under the stratified Cox model.

(a2) $(X_i, \Delta_i, G_i, Z_i)$ are independent and identically distributed random vectors.

(b2) $P(X_i \geq \tau) > 0$ where τ is a pre-specified time point.

(c2) Z_{ik} have bounded total variation, i.e., $|Z_{ik}| < \kappa$ for all $i = 1, ...n$ and $k = 1, ..., p$, where κ is a constant and Z_{ik} is the kth component of Z_i.

(d2) $\int_0^\tau \lambda_{0j}(t)dt < \infty$.

(e2) Continuity of the following functions:

$$s_j^{(1)}(t;\beta) = \frac{\partial}{\partial\beta}s_j^{(0)}(t;\beta), \ \ s_j^{(2)}(t;\beta) = \frac{\partial^2}{\partial\beta\partial\beta^T}s_j^{(0)}(t;\beta)$$

where $s_j^{(d)}(t;\beta)$ is the limiting value of $S_j^{(d)}(t;\beta)$ for $d = 0, 1, 2$, with $s_j^{(1)}(t;\beta)$ and $s_j^{(2)}(t;\beta)$ bounded and $s_j^{(0)}(t;\beta)$ bounded away from 0 for $t \in [0, \tau]$ and β in an open set.

(f2) Positive-definiteness of matrix $\Omega_j(\beta)$:

$$\Omega_j(\beta) = \int_0^\tau v_j(t;\beta)s_j^{(0)}(t;\beta)\lambda_{0j}(t)dt,$$

$$v_j(t;\beta) = \frac{s_j^{(2)}(t;\beta)}{s_j^{(0)}(t;\beta)} - \bar{z}_j(t;\beta)^{\otimes 2}$$

and $\bar{z}_j(t;\beta) = s_j^{(0)}(t;\beta)^{-1}s_j^{(1)}(t;\beta)$ is the limiting value of $\overline{Z}_j(t;\beta)$.

THEOREM 2: Under the regularity conditions for the stratified Cox model, the $\widehat{SMR}_j^\star(t)$ converges almost surely to $\text{SMR}_j^\star(t)$ uniformly in $t \in [0, \tau]$, where

$$SMR_j^\star(t) = \frac{\int_0^t s_j^{(0)}(u;\beta)d\Lambda_{0j}(u)}{\int_0^t s_j^{(0)}(u;\beta)d\Lambda_0(u)}$$

with

$$\Lambda_0(t) = \sum_{\ell=1}^J \int_0^t \frac{s_\ell^{(0)}(u;\beta)}{s^{(0)}(u;\beta)}d\Lambda_{0\ell}(u)$$

and $n^{\frac{1}{2}}\{\widehat{SMR}_j^\star(t) - SMR_j^\star(t)\}$ converges weakly to a zero-mean Gaussian process

with covariance function $\sigma_j^\star(s,t) = E[\xi_{ij}^\star(s;\beta)\xi_{ij}^\star(t;\beta)]$, where

$$\xi_{ij}^\star(t;\beta) = \left\{ \sum_{\ell=1}^{J} \int_0^t s_j^{(0)}(u;\beta) \frac{s_\ell^{(0)}(u;\beta)}{s^{(0)}(u;\beta)} d\Lambda_{0\ell}(u) \right\}^{-1} M_{ij}(t;\beta)$$

$$-h_j(t;\beta) \sum_{\ell=1}^{J} \int_0^t \frac{s_j^{(0)}(u;\beta)}{s^{(0)}(u;\beta)} dM_{i\ell}(u;\beta)$$

$$-h_j(t;\beta) \sum_{\ell=1}^{J} \int_0^t k_j^T(u;\beta) s_\ell^{(0)}(u;\beta) d\Lambda_{0\ell}(u) \Omega(\beta)^{-1} \sum_{\ell=1}^{J} \int_0^\tau \{Z_{i\ell} - \bar{z}_\ell(u;\beta)\} dM_{i\ell}(u;\beta),$$

with

$$h_j(t;\beta) = \left\{ \sum_{\ell=1}^{J} \int_0^t s_j^{(0)}(u;\beta) \frac{s_\ell^{(0)}(u;\beta)}{s^{(0)}(u;\beta)} d\Lambda_{0\ell}(u) \right\}^{-2} \int_0^t s_j^{(0)}(u;\beta) d\Lambda_{0j}(u)$$

$$k_j(u;\beta) = \{s^{(0)}(u;\beta)\}^{-1} \left\{ s_j^{(1)}(u;\beta) - s_j^{(0)}(u;\beta)\, \bar{z}(u;\beta) \right\}.$$

The proof of Theorem 2 is essentially the same as that of Theorem 1. Hence, we focus only on parts that are different.

$$n^{\frac{1}{2}} \left\{ \widehat{SMR}_j^\star(t) - SMR_j^\star(t) \right\} = (3) + (4)$$

$$(3) = n^{\frac{1}{2}} \left\{ \sum_{\ell=1}^{J} \int_0^t S_j^{(0)}(u;\widehat{\beta}_\star) \frac{S_\ell^{(0)}(u;\widehat{\beta}_\star)}{S^{(0)}(u;\widehat{\beta}_\star)} d\widehat{\Lambda}_{0\ell}(u) \right\}^{-1} *$$

$$\left\{ \int_0^t S_j^{(0)}(u;\widehat{\beta}_\star) d\widehat{\Lambda}_{0j}(u) - \int_0^t s_j^{(0)}(u;\beta) d\Lambda_{0j}(u) \right\}$$

$$= n^{-\frac{1}{2}} \left\{ \sum_{\ell=1}^{J} \int_0^t s_j^{(0)}(u;\beta) \frac{s_\ell^{(0)}(u;\beta)}{s^{(0)}(u;\beta)} d\Lambda_{0\ell}(u) \right\}^{-1} \sum_{i=1}^{n} \int_0^t dM_{ij}(u;\beta) + o_p(1)$$

$$(4) = \int_0^t s_j^{(0)}(u;\beta) d\Lambda_{0j}(u)\, n^{\frac{1}{2}} \left\{ \sum_{\ell=1}^{J} \int_0^t S_j^{(0)}(u;\widehat{\beta}_\star) \frac{S_\ell^{(0)}(u;\widehat{\beta}_\star)}{S^{(0)}(u;\widehat{\beta}_\star)} d\widehat{\Lambda}_{0\ell}(u) \right\}^{-1}$$

$$- \int_0^t s_j^{(0)}(u;\beta) d\Lambda_{0j}(u)\, n^{\frac{1}{2}} \left\{ \sum_{\ell=1}^{J} \int_0^t s_j^{(0)}(u;\beta) \frac{s_\ell^{(0)}(u;\beta)}{s^{(0)}(u;\beta)} d\Lambda_{0\ell}(u) \right\}^{-1}$$

$$= - h_j(t, \beta) \, n^{\frac{1}{2}} \sum_{\ell=1}^{J} \left\{ \int_0^t S_j^{(0)}(u; \widehat{\beta}_\star) \frac{S_\ell^{(0)}(u, \widehat{\beta}_\star)}{S^{(0)}(u; \widehat{\beta}_\star)} d\widehat{\Lambda}_{0\ell}(u) - \int_0^t s_j^{(0)}(u; \beta) \frac{s_\ell^{(0)}(u; \beta)}{s^{(0)}(u; \beta)} d\Lambda_{0\ell}(u) \right\}$$

$$+ o_p(1)$$

with

$$h_j(t, \beta) \;\; = \;\; \left\{ \sum_{\ell=1}^{J} \int_0^t s_j^{(0)}(u; \beta) \frac{s_\ell^{(0)}(u; \beta)}{s^{(0)}(u; \beta)} d\Lambda_{0\ell}(u) \right\}^{-2} \int_0^t s_j^{(0)}(u; \beta) d\Lambda_{0j}(u).$$

Note that

$$n^{\frac{1}{2}} \sum_{\ell=1}^{J} \left\{ \int_0^t S_j^{(0)}(u; \widehat{\beta}_\star) \frac{S_\ell^{(0)}(u; \widehat{\beta}_\star)}{S^{(0)}(u; \widehat{\beta}_\star)} d\widehat{\Lambda}_{0\ell}(u) - \int_0^t s_j^{(0)}(u; \beta) \frac{s_\ell^{(0)}(u; \beta)}{s^{(0)}(u; \beta)} d\Lambda_{0\ell}(u) \right\}$$

$$= n^{-\frac{1}{2}} \sum_{\ell=1}^{J} \sum_{i=1}^{n} \int_0^t \frac{s_j^{(0)}(u; \beta)}{s^{(0)}(u; \beta)} dM_{i\ell}(u; \beta) + \sum_{\ell=1}^{J} \int_0^t k_j^T(u; \beta) s_\ell^{(0)}(u; \beta) d\Lambda_{0\ell}(u) \, n^{\frac{1}{2}} (\widehat{\beta}_\star - \beta)$$

$$+ o_p(1),$$

where

$$k_j(u; \beta) \;\; = \;\; \left\{ s^{(0)}(u; \beta) \right\}^{-1} \left\{ s_j^{(1)}(u; \beta) - s_j^{(0)}(u; \beta) \bar{z}(u; \beta) \right\}$$

and

$$n^{\frac{1}{2}} (\widehat{\beta}_\star - \beta) = \left\{ \sum_{\ell=1}^{J} \Omega_\ell(\beta)^{-1} \right\} n^{-\frac{1}{2}} \sum_{\ell=1}^{J} \sum_{i=1}^{n} \int_0^\tau [Z_{i\ell} - \bar{z}_\ell(u; \beta)] dM_{i\ell}(u; \beta) + o_p(1)$$

Therefore, the process $n^{\frac{1}{2}} \{ \widehat{SMR}_j^\star(t) - SMR_j^\star(t) \}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^{n} \xi_{ij}^\star(t; \beta)$ with $\xi_{ij}^\star(t; \beta)$. It then follows from the multivariate central limit theorem, together with a proof of tightness, that $n^{\frac{1}{2}} \{ \widehat{SMR}_j^\star(t) - SMR_j^\star(t) \}$ converges weakly to a zero-mean Gaussian process with covariance function $\sigma_j^\star(s, t) = E[\xi_{ij}^\star(s; \beta) \xi_{ij}^\star(t; \beta)]$.

Note that the limiting value of $\widehat{SMR}_j^\star(t)$ depends not only on the center-specific hazards, but also on the censoring and covariate distributions of center j. The SRR

is desirable in this light since center-specific limiting values would differ only due to corresponding differences in center-specific hazards.

THEOREM 3: Under the regularity conditions for the stratified Cox model, $\widehat{SRR}_j(t)$ converges almost surely to $SRR_j(t)$ uniformly in $t \in [0, \tau]$, where

$$SRR_j(t) = \frac{\sum_{\ell=1}^{J} \int_0^t s_\ell^{(0)}(u; \beta) d\Lambda_{0j}(u)}{\sum_{\ell=1}^{J} \int_0^t s_\ell^{(0)}(u; \beta) d\Lambda_{0\ell}(u)}$$

and $n^{\frac{1}{2}}\{\widehat{SRR}_j(t) - SRR_j(t)\}$ converges weakly to a zero-mean Gaussian process with covariance function $\sigma_j^S(s, t) = E[\xi_{ij}^S(s; \beta)\xi_{ij}^S(t; \beta)]$, where

$$\xi_{ij}^S(t; \beta) = w(t; \beta) \int_0^t \frac{s^{(0)}(u; \beta)}{s_j^{(0)}(u; \beta)} dM_{ij}(u; \beta)$$

$$-w(t; \beta)^2 \int_0^t s^{(0)}(u; \beta) d\Lambda_{0j}(u) \sum_{\ell=1}^{J} \int_0^t dM_{i\ell}(u; \beta)$$

$$+w(t; \beta) \int_0^t r_j^T(u; \beta) d\Lambda_{0j}(u) \Omega(\beta)^{-1} \int_0^\tau \{Z_{i\ell} -$$

with

$$w(t; \beta) = \left\{ \sum_{\ell=1}^{J} \int_0^t s_\ell^{(0)}(u; \beta) d\Lambda_{0\ell}(u) \right\}^{-1}$$

$$r_j(u; \beta) = s^{(1)}(u; \beta) - s^{(0)}(u; \beta)\, \overline{z}_j(u; \beta).$$

We can decompose the quantity of interest in Theorem 3 as follows

$$n^{\frac{1}{2}}\{\widehat{SRR}_j(t) - SRR_j(t)\} = (5) + (6),$$

where we have

$$(5) = w(t; \beta)n^{\frac{1}{2}} \left\{ \int_0^t S^{(0)}(u; \widehat{\beta}_\star) d\widehat{\Lambda}_{0j}(u) - \int_0^t s^{(0)}(u; \beta) d\Lambda_{0j}(u) \right\}$$

$$= w(t; \beta) \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^t \frac{s^{(0)}(u; \beta)}{s_j^{(0)}(u; \beta)} dM_{ij}(u; \beta) + \int_0^t r_j^T(u; \beta) d\Lambda_{0j}(u)\, n^{\frac{1}{2}}(\widehat{\beta}_\star - \beta) \right\} + o_p(1),$$

with

$$
w(t; \beta) = \left\{ \sum_{\ell=1}^{J} \int_0^t s_\ell^{(0)}(u; \beta) d\Lambda_{0\ell}(u) \right\}^{-1}
$$

$$
r_j(u; \beta) = s^{(1)}(u; \beta) - s_j^{(1)}(u; \beta)\bar{z}(u; \beta)
$$

$$
n^{\frac{1}{2}}(\widehat{\beta}_\star - \beta) = \left\{ \sum_{\ell=1}^{J} \Omega_\ell(\beta)^{-1} \right\} n^{-\frac{1}{2}} \sum_{\ell=1}^{J} \sum_{i=1}^{n} \int_0^\tau [Z_{i\ell} - \bar{z}_\ell(u; \beta)] dM_{ij}(u; \beta) + o_p(1),
$$

and

$$
(6) = \int_0^t s^{(0)}(u; \beta) d\Lambda_{0j}(u) \ n^{\frac{1}{2}} \left\{ \sum_{\ell=1}^{J} \int_0^t S_\ell^{(0)}(u; \widehat{\beta}_\star) d\widehat{\Lambda}_{0\ell}(u) \right\}^{-1}
$$

$$
- \int_0^t s^{(0)}(u; \beta) d\Lambda_{0j}(u) \ n^{\frac{1}{2}} \left\{ \sum_{\ell=1}^{J} \int_0^t s_\ell^{(0)}(u; \beta) d\Lambda_{0\ell}(u) \right\}^{-1}
$$

$$
= - \frac{\int_0^t s^{(0)}(u; \beta) d\Lambda_{0j}(u)}{\{\sum_{\ell=1}^{J} \int_0^t s_\ell^{(0)}(u; \beta) d\Lambda_{0\ell}(u)\}^2} n^{-\frac{1}{2}} \sum_{\ell=1}^{J} \int_0^t dM_{i\ell}(u; \beta) + o_p(1)
$$

Therefore, the process $n^{\frac{1}{2}}\{\widehat{SRR}_j(t) - SRR_j(t)\}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^{n} \xi_{ij}^s(t; \beta)$. It converges asymptotically to a zero-mean Gaussian process with covariance function $\sigma_j^s(s, t) = E[\xi_{ij}^s(s; \beta)\xi_{ij}^s(t; \beta)]$.

We now focus on the asymptotic properties of the supremum test for constant center effects. Under the null hypothesis, $H_0 : SRR_j(t) = SRR_j$, we have $n^{\frac{1}{2}}\{\widehat{SRR}_j(t) - \widehat{SRR}_j(\tau)\}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^{n}\{\widehat{\xi}_{ij}^S(t; \beta) - \widehat{\xi}_{ij}^S(\tau; \beta)\}$. This suggests that the quantity $n^{-\frac{1}{2}} \sum_{i=1}^{n} R_i\{\widehat{\xi}_{ij}^S(t; \widehat{\beta}) - \widehat{\xi}_{ij}^S(\tau; \widehat{\beta})\}$ would have the same distribution as $n^{\frac{1}{2}}\{\widehat{SRR}_j(t) - \widehat{SRR}_j(\tau)\}$ under H_0, where R_i is a standard normal random variable. Repeatedly generating independent standard normal random samples R_i ($i = 1, ..., n$), we can determine the empirical critical values of this test. The properties of the test based on $\widehat{F}_j$ are obtained similarly.

2.4 Simulation Study

Finite-sample properties of the estimators described in Section 3 were evaluated through a series of simulation studies. We considered $J = 10$ centers. Death times were generated from the Weibull model, $\lambda_{ij}(t) = \alpha_j \gamma_j t^{\gamma_j - 1} \exp(\beta^T Z_i)$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, 10$, where $Z_i = (Z_{i1}, Z_{i2})^T$, Z_{i1} was a Normal variate and Z_{i2} followed a Bernoulli distribution. We set $\beta^T = (\beta_1, \beta_2) = (0.02, -0.2)$. We varied the sample size for each center as $n_j = 50, 100, 500$. The censoring percentages were approximately 20%. The Cox SMR was calculated at the end of the study period, while other estimators were calculated at various percentiles of observation time. Each data configuration was replicated 1000 times.

2.4.1 Setting 1: Hazards equal across centers

First, we considered a simulation setting where the hazard functions were equal across centers. Note that under this setting, the limiting values of all estimators equal 1. Whether or not the covariate or censoring distributions were center-dependent should have no influence on the results in this setting. The results with sample size $n_j = 100$ are displayed in Table 1. Not surprisingly, each of the measures was very close to 1. The variation of the Cox SMR was uniformly smaller than each of the alternative estimators. In terms of the asymptotic approximation, estimators based on indirect standardization were sufficiently well-behaved, in the sense that the average asymptotic standard errors (ASE) were generally close to the empirical standard deviations (ESD), while the empirical coverage probabilities (CP) were generally consistent with the nominal value 0.95. The performance of the SRR tended to improve as sample sizes increased: poor CP (i.e., in the 0.87-0.89 range) for sample size $n_j = 50$; moderate CP (0.91-0.93) for $n_j = 100$; good CP (0.93-0.95)

for $n_j = 500$.

2.4.2 Setting 2: Center-dependent hazards

For the second set of simulations, different values of α_j and γ_j were used, such that the hazard functions increased with increasing center number ($j = 1, \ldots, 10$). Both the censoring and covariate distributions were chosen to be center-independent, where the covariate Z_{i1} came from a Normal (50, 10) distribution; Z_{i2} followed a Bernoulli (0.5) distribution; and the censoring times were generated from a Uniform (3, 10) distribution. Results based on sample size $n_j = 100$ are given in Table 2. Three aspects are noteworthy here. First, we report the time-dependent estimators calculated at $t=5$, which was roughly the 70th percentile of the observation time distribution. We observed similar performance for these estimators calculated at other percentiles of the observation time distribution. Second, $\widehat{SMR}_j$ tended to differ notably from other estimators, consistent with the failure of the proportionality assumption. Third, the means for $\widehat{SMR}_j(t)$ and $\widehat{SMR}_j^\star(t)$ were quite close to each other in this setting, due to the covariate distributions being center-independent. Although the unstratified model omitting the center effects results in biased estimators of $(\beta_1, \beta_2)^T$, this bias turned out to be only moderate in this case.

We now focus on the finite-sample performance of the asymptotic results pertaining to $\widehat{SMR}_j$. For a center with hazard lower than the population average, the ASE of $\widehat{SMR}_j$ was too large compared to the ESD, leading to a CP much greater than 0.95. In contrast, for a center with hazard greater than the population average, the ASE was too small, leading to a notably low CP. These performances did not improve as we increased the sample size from $n_j = 100$ to $n_j = 500$. Clearly, the flaws in the variance estimation yield either an overestimated or underestimated

confidence interval for the Cox SMR based on an unstratified model. In contrast, the SE estimator for $\widehat{SMR}_j^\star$ was sufficiently accurate for this setting. Finally, the performance of the asymptotic SE for $\widehat{SRR}_j$ was sensitive to the sample size, with CP being similar to that from Setting 1.

2.4.3 Setting 3: Center-dependent censoring distribution

As we mentioned previously, even with the proposed modifications, estimators based on indirect standardization may be misleading if either the censoring or covariate distributions are center-dependent. To illustrate this point, we performed simulations with the following three conditions: (i) the hazard functions were center-dependent and increased with center index ($j = 1, \ldots, 9$), while center $j = 10$ had exactly the same hazard function as that of center $j = 9$; (ii) the censoring distributions for center $j = 9$ and $j = 10$ were substantially different. For center $j = 9$, the censoring times were generated from a Weibull distribution such that the censoring mainly occurred in the later stages, while for center $j = 10$ the censoring times were generated from a Weibull distribution for which the censoring tended to occur in the early stages; (iii) the covariate distributions were center-independent (same as those in setting 2). We compared the estimators for centers $j = 9$ and $j = 10$ in Table 3. The results showed that the SRR performed well, in the sense that empirical means for centers $j = 9$ and $j = 10$ were approximately equal. The limiting values of $\widehat{SRR}_9(t)$ and $\widehat{SRR}_{10}(t)$ were exactly the same. In contrast, each of the estimators based on indirect standardization tended to provide misleading results, since the limiting values and the corresponding estimators were notably different between centers $j = 9$ and $j = 10$.

2.4.4 Setting 4: Center-dependent covariate distributions

We also performed simulations under a setting in which the distribution of the covariate vector differed by center. The set up for hazard functions was exactly the same as for Setting 3. Censoring times were generated from Uniform (3, 10) distribution, while center $j = 9$ and $j = 10$ had substantially different covariate distributions. In center $j = 9$ the covariate Z_{i1} came from a Normal distribution with mean 30 and Z_{i2} followed a Bernoulli (0.8) distribution; while in center $j = 10$, the covariate Z_{i1} was derived from a Normal distribution with mean 70 and Z_{i2} followed a Bernoulli (0.2) distribution. To connect this to the data set that motivated our research, suppose that we are comparing two kidney transplant centers which differ greatly with respect to both mean age and the fraction of patients with diabetes. Patterns in the estimators paralleled those of Setting 3. Estimators based on indirect standardization tended to differ between these two centers. Only the $\widehat{SRR}_j$ performed well, from this perspective.

2.4.5 Test for constant center effect

We explored the performance of the supremum test introduced at the end of Section 3 to evaluate whether a particular center effect was constant over time. The values of γ_j were chosen to be equal across centers. The values of α_j were center-dependent and increased with center index j $(j = 1, \ldots, 10)$. The censoring times were generated from a center-dependent Weibull distribution. The covariate Z_{i1} was derived from a Normal distribution with center-dependent means and Z_{i2} followed a Bernoulli distribution with center-dependent probabilities. Under this setting, the center effects are constant over time. For each simulated data set, we obtained 1000 realizations of $(R_1, \ldots, R_n)$, where $R_i \sim N(0, 1)$ in order to compute $\widehat{T}_j$. Results

based on sample size $n_j = 100$ are given in Table 4. The proposed sup test appeared to yield the desired Type I error rate, with empirical significance levels being close to 0.05.

Table 2.1: Simulation Setting 1: hazards are center independent; censoring and covariate distributions are center independent; n_j=100; t=5; SMR$_j$=Cox SMR given in (2.2); SMR$_j(t)$=time-dependent SMR given in (2.9); SMR$_j^\star$=modification to SMR given in (2.10); SRR_j= standardized rate ratio, given in (2.11); only centers with even number ($j = 2, \ldots, 10$) are shown in this table)

Center	(γ_j, α_j)	Measure	True Value	BIAS	ESD	ASE	CP
$j = 2$	(0.90, 0.049)	SMR$_j$	1	0.006	0.137	0.138	0.96
		SMR$_j(t)$	1	0.011	0.161	0.162	0.96
		SMR$_j^\star(t)$	1	0.011	0.161	0.161	0.95
		SRR$_j(t)$	1	-0.003	0.158	0.157	0.93
$j = 4$	(0.90, 0.049)	SMR$_j$	1	0.005	0.142	0.138	0.95
		SMR$_j(t)$	1	0.012	0.169	0.162	0.94
		SMR$_j^\star(t)$	1	0.012	0.170	0.161	0.94
		SRR$_j(t)$	1	-0.003	0.167	0.156	0.92
$j = 6$	(0.90, 0.049)	SMR$_j$	1	0.012	0.140	0.139	0.96
		SMR$_j(t)$	1	0.023	0.164	0.163	0.95
		SMR$_j^\star(t)$	1	0.023	0.164	0.162	0.95
		SRR$_j(t)$	1	0.010	0.162	0.157	0.92
$j = 8$	(0.90, 0.049)	SMR$_j$	1	-0.001	0.143	0.138	0.95
		SMR$_j(t)$	1	0.009	0.168	0.162	0.95
		SMR$_j^\star(t)$	1	0.009	0.168	0.161	0.94
		SRR$_j(t)$	1	-0.006	0.166	0.156	0.92
$j = 10$	(0.90, 0.049)	SMR$_j$	1	0.008	0.140	0.138	0.95
		SMR$_j(t)$	1	0.015	0.166	0.162	0.96
		SMR$_j^\star(t)$	1	0.015	0.166	0.161	0.95
		SRR$_j(t)$	1	-0.000	0.163	0.157	0.93

2.5   Application to Kidney Transplant Data

We applied each of the methods considered to investigate the performance of Canadian transplant centers with respect to post kidney transplant survival. Data

Table 2.2: Simulation Setting 2 (hazards are center dependent; censoring and covariate distributions are center independent; n_j=100; t=5; SMR$_j$=Cox SMR given in (2.2); SMR$_j(t)$=time-dependent SMR given in (2.9); SMR$_j^\star$=modification to SMR given in (2.10); SRR_j= standardized rate ratio, given in (2.11); only centers with even number ($j = 2, \ldots, 10$) are shown in this table)

Center	(γ_j, α_j)	Measure	True Value	BIAS	ESD	ASE	CP
$j = 2$	(0.840, 0.029)	SMR$_j$	0.530	0.004	0.096	0.120	**0.98**
		SMR$_j(t)$	0.544	0.005	0.108	0.120	**0.98**
		SMR$_j^\star(t)$	0.543	0.004	0.108	0.107	0.94
		SRR$_j(t)$	0.546	0.002	0.108	0.106	0.92
$j = 4$	(0.920, 0.037)	SMR$_j$	0.778	0.003	0.120	0.129	0.96
		SMR$_j(t)$	0.778	0.002	0.136	0.141	0.96
		SMR$_j^\star(t)$	0.778	0.001	0.137	0.130	0.94
		SRR$_j(t)$	0.775	-0.003	0.135	0.128	0.92
$j = 6$	(1, 0.045)	SMR$_j$	1.085	0.006	0.140	0.140	0.93
		SMR$_j(t)$	1.065	0.005	0.164	0.154	0.94
		SMR$_j^\star(t)$	1.065	0.005	0.164	0.156	0.94
		SRR$_j(t)$	1.060	0.000	0.164	0.152	0.92
$j = 8$	(1.080, 0.054)	SMR$_j$	1.469	0.006	0.172	0.155	0.92
		SMR$_j(t)$	1.415	0.002	0.184	0.172	0.92
		SMR$_j^\star(t)$	1.417	0.004	0.185	0.187	0.95
		SRR$_j(t)$	1.419	-0.002	0.190	0.181	0.91
$j = 10$	(1.160, 0.062)	SMR$_j$	1.875	-0.001	0.202	0.174	**0.89**
		SMR$_j(t)$	1.794	-0.001	0.215	0.174	**0.92**
		SMR$_j^\star(t)$	1.799	0.003	0.217	0.219	0.95
		SRR$_j(t)$	1.826	-0.006	0.231	0.213	**0.91**

Table 2.3: Simulation Setting 3 (center $j = 9$ and $j = 10$ have same hazards but different censoring distributions) and Simulation Setting 4 (center $j = 9$ and $j = 10$ have same hazards but different covariate distributions); $n_j{=}100$; $t{=}5$

Setting	Measure	Center	True value	Mean	ESD
3	SMR_j	9	2.111	2.116	0.197
		10	1.735	1.747	0.215
	$\mathrm{SMR}_j(t)$	9	1.777	1.802	0.235
		10	1.664	1.679	0.220
	$\mathrm{SMR}_j^\star(t)$	9	1.867	1.894	0.235
		10	1.768	1.784	0.234
	$\mathrm{SRR}_j(t)$	9	**1,945**	**1.944**	0.262
		10	**1.945**	**1.940**	0.308
4	SMR_j	9	2.313	2.312	0.232
		10	1.960	1.957	0.172
	$\mathrm{SMR}_j(t)$	9	2.128	2.125	0.253
		10	1.896	1.892	0.177
	$\mathrm{SMR}_j^\star(t)$	9	2.102	2.097	0.326
		10	1.942	1.940	0.241
	$\mathrm{SRR}_j(t)$	9	**2.118**	**2.117**	0.343
		10	**2.118**	**2.112**	0.296

Table 2.4: Empirical significance levels (ESL) of supremum-based test for constant center effect, given in (2.13)

Center	(γ_j, α_j)	ESL
1	(1, 0.025)	0.050
2	(1, 0.027)	0.060
3	(1, 0.029)	0.055
4	(1, 0.031)	0.042
5	(1, 0.033)	0.059
6	(1, 0.035)	0.058
7	(1, 0.037)	0.053
8	(1, 0.039)	0.074
9	(1, 0.041)	0.044
10	(1, 0.043)	0.050

were obtained from the Canadian Organ Replacement Register (CORR), which is administered by the Canadian Institute for Health Information. CORR is a nation-wide and population-based organ-failure registry that includes all organ failure programs in Canada. The study population included patients receiving a kidney transplant in Canada between 1988 and 1997. Patients were followed from the date of transplantation to the time of graft failure, death or December 31, 1997. Graft failure was considered to occur when the transplanted kidney ceased to function. Patients younger than 18 and centers with sizes less than 50 were excluded. The final sample size was $n = 5143$ patients from $J = 20$ different centers. Of this sample, thirty-six percent of the patients experienced the event of interest (death or graft failure).

Cox regression was employed to investigate the mortality hazards. Adjustment covariates included age, diagnosis and donor type. Technically, the Cox SMR is simple to implement because it can be calculated based on martingale residuals. With an offset statement, the $\widehat{SMR}_j(t)$ and $\widehat{SMR}^\star_j(t)$ can also be easily calculated (e.g., SAS's PROC PHREG). The directly standardized estimator, $\widehat{SRR}_j(t)$, was computed using SAS IML. Center numbers are re-ordered by values of $\widehat{SRR}_j(t)$ at $t = 5$ years.

A comparison of the estimators across centers is displayed in Figure 1. The Cox SMRs for several centers were notably different from the time-dependent SMRs, which suggests that the center effects were not proportional. Note that the unstratified and stratified Cox models provided similar coefficients in this example because the differences in hazard, censoring and covariate distributions among centers were moderate. Therefore, the time-dependent $\widehat{SMR}_j(t)$, $\widehat{SMR}^\star_j(t)$ and $\widehat{SRR}_j(t)$ were generally quite similar.

Using the asymptotic normality of the proposed estimators, we constructed the

point-wise confidence intervals using SAS IML. Center $j = 12$ had a Cox SMR significantly larger than 1, while all other estimators for this center were non-significant. These findings suggest that center $j = 12$ offered good performance in the early stage of post-transplant follow-up, and thus, lower mortalities than the national average; however, poorer performance was observed in the later stages of follow-up. This trend was clear in Figure 2, in which the $\widehat{SRR}_{12}(t)$ and its smoothed counterpart, $\widetilde{SRR}_{12}(t)$, are illustrated.

The confidence interval of SMR* for all the centers at $t = 5$ years are depicted in Figure 3A and compared to the corresponding results from the SRR through Figure 3B. These figures demonstrate that $\widehat{SMR}_j^\star$ and $\widehat{SRR}_j$ were quite similar in this particular example.

2.6   Discussion

The aim of this study was to evaluate existing semi-parametric regression methods for comparing center-specific mortality and to propose alternative methods. The Cox SMR, calculated from an unstratified Cox model, is often used to estimate center effects. However, relatively little information is currently available in the literature regarding the assumptions and limitations associated with this method. Despite its ease of calculation, the Cox SMR does have some drawbacks. First, if the center-specific hazards cross, the Cox SMR, when calculated at the end of the follow up time, may fail to detect a difference. Second, the Cox SMR leads to invalid variance estimation and, hence, confidence intervals and hypothesis tests which may be inaccurate when the hazard functions are center-dependent. The first disadvantage can be addressed using a time-dependent version of the SMR. The second limitation can be remedied by a modified version of the SMR which

uses the regression parameter estimate from a center-stratified Cox model. Even with such modifications, estimators based on indirect standardization still converge to values which depend on center-specific censoring or covariate distributions. The SRR, directly standardized measure, does not share this drawback. The SRRs for two given centers will be unequal only because the center-specific mortality hazards differ, since direct standardization accounts for imbalance with respect to center-specific covariate and censoring distributions. We have seen, however, that larger sample sizes are typically required to implement the SRR.

Despite the fact that the SMR from the unstratified Cox model has been widely used, we do not generally recommend it as method for estimating center effects, given the afore-listed concerns. The modified SMR (based on a center-stratified Cox model) is preferred if the goal is to determine whether an individual center has higher or lower mortality rates than expected. If the goal of the study is to order centers based on performance, centers cannot be compared through methods based on indirect standardization (such as the SMR or its proposed modification) because the ordering of the SMRs across centers may be due to center-specific censoring and covariate distributions. Finally, for studies with moderate to large center sizes, the proposed direct standardization estimator SRR shares the SMR's ease of interpretation but rectifies its disadvantages, and hence is a more appropriate choice.

Analysis of multi-center studies has been the subject of much discussion. Existing methods of comparing center-specific mortality, whether based on fixed or random effect models, measure the average center effect and, thus, rely on the proportional hazards assumption; an assumption borne out of convenience, and one which will often fail in practice. A distinguishing feature of the methods we propose in this report is that center effects are estimated without assuming proportionality with respect to

center. The time-dependent measures we propose would be preferred in situations where center-specific hazard functions may cross. An alternative approach to accounting for non-proportional hazards a Cox model with a time-varying regression coefficient. This approach has drawbacks when applied in studies of center-specific mortality. For example, the choice of reference center is arbitrary, and different choices of the reference center may lead to different results. Another limitation is that the number of time-varying regression coefficients increase as the number of centers increase, which leads to difficulty in identifying the correct functional form for the the time-varying aspects. These two weaknesses are remedied by our proposed approaches, which do not require the specification of either a reference center or form for the time-dependence of center effects.

The material in this report could be extended in several useful directions. Perhaps most notably, it is often of interest to evaluate center effects in the setting where $N_i(t)$ counts an event that is recurrent (i.e., may occur multiple times per subject; e.g., hospitalizations). Although there would be some overlap among the key issues of concern in the recurrent event setting and those discussed in this report, there are likely many additional methodologic issues.

Figure 2.1: Evaluation of $J{=}20$ Canadian kidney transplant centers: $\widehat{SMR}_j$, $\widehat{SMR}_j(5)$, $\widehat{SMR}_j^\star(5)$, $\widehat{SRR}_j(5)$, where t represents years post-transplant. Center numbers are ordered by $\widehat{SRR}_j(5)$
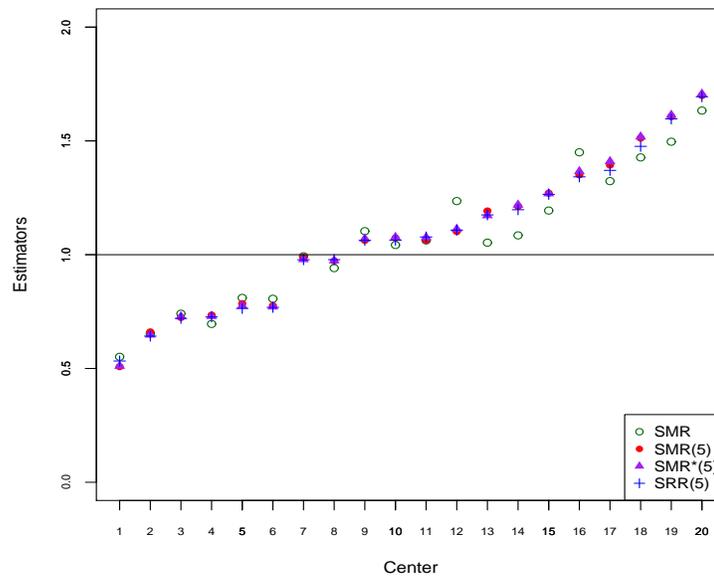
Figure 2.2: Evaluation of center j=12: $\widehat{SRR}_j(t)$ and its kernel-smoothed counterpart, $\widetilde{SRR}_j(t)$, where t represents years.

(a) Figure 2A
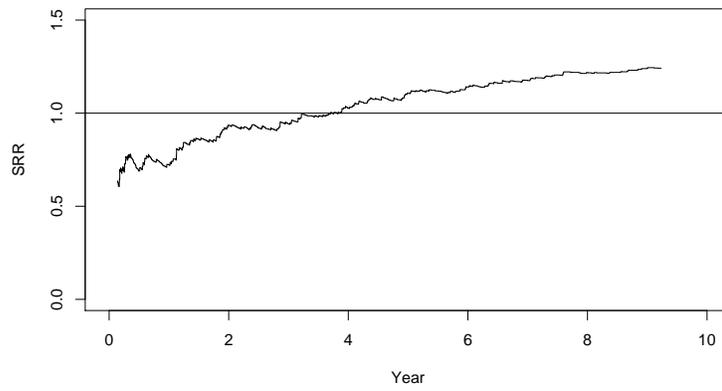
(b) Figure 2B
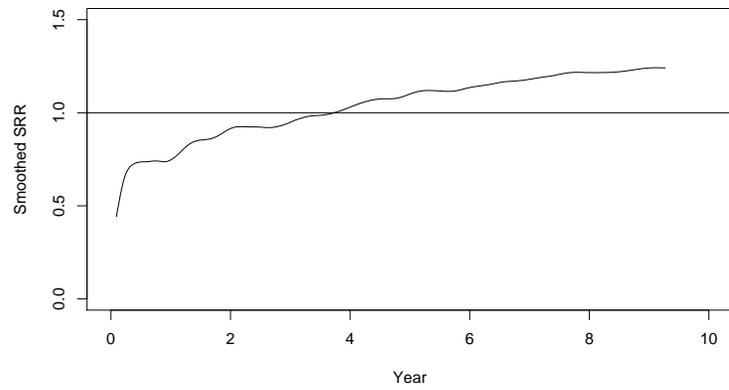
Figure 2.3: Evaluation of J=20 Canadian kidney transplant centers: Point estimates and 95% confidence interval of $\widehat{SRR}_j(t)$ and $\widehat{SMR}_j^{\star}(t)$ at $t = 5$ years. Center numbers are ordered by $\widehat{SRR}_j(5)$

(a) Figure 3A
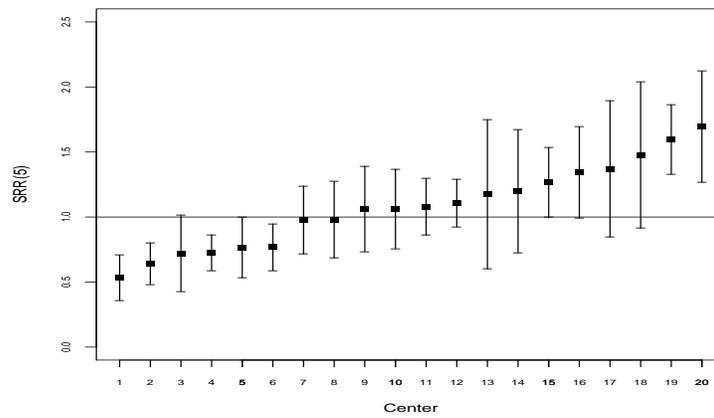
(b) Figure 3B
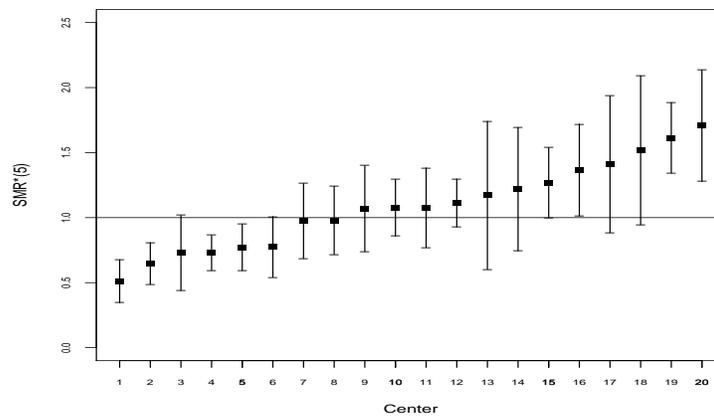
CHAPTER III

Evaluating Center-specific Long Term Outcomes Through Difference in Mean Survival Time

In the context of survival data, the difference in mean lifetime is arguably a more meaningful measure than the ratio of death rates. Existing methods based on restricted mean lifetime tend to focus on short- and medium-term clinical outcomes. Such measures may fail to capture the long-term effects that relate to the quality of follow-up care. Another challenging part of measuring center-specific performance is that multi-center studies tend to generate heterogeneous samples; e.g., the effect of certain important covariates can be unequal across centers. In this chapter, we propose a method that combines a log-normal frailty model and piece-wise exponential baseline rates to compare the mean survival time across centers. The proposed methods allow for estimation of mean survival time as opposed to restricted mean lifetime and, within this context, more robust profiling of long-term center-specific outcomes. Maximum likelihood based estimation is carried out using a Laplace approximation for integration. Direct standardization methods are applied to contrast mean survival time by center. Asymptotic properties of the proposed estimators are derived. Finite-sample properties are examined through simulation. The methods developed are then applied to national kidney transplant data.

3.1 Introduction

Survival analysis plays an important role in the analysis of multi-center studies. Often of particular interest is the determination of which medical centers have significantly better or significantly worse long term outcomes. For instance, the methods we propose in this chapter are motivated by the need to profile center-specific long-term kidney transplant outcomes in the United States. Kidney transplantation has a sophisticated and well-established program for center specific outcomes reporting and quality assurance. The data from this program are used by transplant centers for quality improvement, by payers and regulators to achieve quality assurance, and by referring physicians and patients to identify the appropriate center for treatment. Thus, valid measurements are crucial. In particular, the evaluation of center-specific outcomes is important to determine how therapy procedures at a given center affect patient outcomes.

A variety of methodological approaches have been used to evaluate centers. The most frequently used methods rely on the comparison of observed versus expected outcomes. The estimator for the difference in mean survival time is one such example. Although mean survival time is often of interest, the quantity may not be well estimated in the presence of censoring, since the estimated survival function need not drop to zero. To address this concern, a commonly used alternative measure is the restricted mean lifetime, interpreted as the expected number of time units survived, out of a pre-determined upper limit L. To incorporate covariates, Karrison (1987) compared the restricted mean lifetime between two groups using a piecewise exponential model. Zucker (1998) extended Karrison's approach based on the stratified Cox model, in which the relationship between groups and hazard are left arbitrary.

Chen and Tsiatis (2001) proposed estimators for the average causal treatment difference in restricted mean lifetime between treatment groups, which accounts for the treatment imbalance in prognostic factors. Zhang and Schaubel (2011) extended the method to accommodate dependent censoring. Due to the use of the truncation time, L, restricted mean lifetime is a useful measure for short- and medium-term clinical outcomes. However, for patients receiving kidney transplantation, the majority of death and graft failures occur in the long-term (Wolfe et al. 1999). In such cases, measures for short- and medium-term outcomes may fail to capture the long-term effects that reveal the overall quality of care provided by the center. For instance, in our motivating example, the longest term outcome available in existing center-specific reports is $3-$year survival (OPTN/SRTR). The events after 3 years are ignored and the loss of information corresponding to the later events may be substantial. Long term center-specific profiling may offer important insight into variations in processes and intensity of care. Thus, in order to motivate and investigate long-term kidney transplant outcomes, valid metrics of long term function are needed. With these concerns, restricted mean lifetime may not be a good substitute and mean survival time are often of interest.

The purpose of this chapter is to propose a valid method to profile centers with respect to mean survival time. Within this context, we have developed novel methods to profile center-specific measures of transplant utility. These methods allow for a reliable measurement of the expected mean survival time for each transplant, and within this context robust profiling of long-term center specific outcomes. Considering such measures may provide new perspectives on the relationship between short-term outcomes, long-term outcomes, intensity of care, and costs of kidney transplant care and how all these impact the overall utility of kidney transplantation.

The remainder of this chapter is organized as follows. The data sources and study population are described in the next section. In Section 3.3, we summarize some important issues in the comparison of long-term center-specific outcomes. In Section 3.4, we describe our proposed model and method to estimate the difference of mean survival time. Finite-sample properties are examined in Section 3.5 through simulations. The proposed methods are applied to national kidney transplant data in Section 3.6. The article concludes with a discussion in Section 3.7, which includes a comparison with pertinent methods in the existing literature.

3.2 Data sources and study population

Data were obtained from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donors, wait-listed candidates, and transplant recipients in the United States, as submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), of the U.S. Department of Health and Human Services, provides oversight to the activities of the OPTN and SRTR contractors.

Included in the analysis were adult patients ($\geq$ 18 years of age at transplant) who underwent deceased donor kidney transplantation between January 1990 and December 2008. Adjustment covariates in this study included age, race, gender, diagnosis, donation after cardiac death (DCD), Expanded Criteria Donor (ECD), BMI, duration of pre-transplant dialysis, indicator of previous kidney transplant, and cold ischemic time. Transplant centers with sample size smaller than 10 and patients who received living donor transplants were eliminated from additional analysis. The final sample size was $n = 146,248$ from $J = 282$ centers across the United States. The sample sizes varied considerably across centers and some centers have only a

few patients. The mean age at transplant was 48 and mean donor age was 36. The median survival time over the study period was 10.8 years across centers. Graft failure was considered to occur when the transplanted kidney ceased to function, such that failure time (recorded in years) was defined as the time from transplantation to graft failure or death, whichever occurred first.

3.3   Assumed model

In this section, we describe the most important issues in the analysis of long-term center-specific outcomes for kidney transplant centers. After some requisite preliminaries, we discuss the model we choose for center-effect measures.

3.3.1   Piece-wise exponential baseline rates

Estimating mean survival time is generally difficult to carry out in the presence of censoring. We propose a method based on piece-wise exponential baseline rates with log-normal random effect. The main reason that we select a parametric hazard model instead of the Cox model is that the later allows for inference only on the $(0, \tau]$ intervals where τ is the maximum observation time. In the Cox regression, the baseline hazard function is completely unknown and estimation involving the baseline may not be stable at later follow-up times when the risk set is small. Therefore, although the Cox model is a reasonable choice in the context of estimating restricted mean lifetime, it may not be a suitable choice when mean survival time is of interest. Another advantage of piecewise exponential baseline rates is that we can make use of the link between mixed models and the log-normal random effect model. Holford, T. R. (1980) noted that the piece-wise exponential model was equivalent to a Poisson log-linear regression model with count either 0 or 1 for each combination of individual and interval, where the death indicator is the response and the log of exposure

time enters as an offset. Such connections lead to both computational and theoretical simplification. For instance, existing generalized linear mixed model software, such as PROC GLIMMIX in SAS, can be applied to estimate the parameters. This is especially important for our motivating example, since estimating center effects can be a computationally intense task in nation-wide studies with large numbers of patients and centers. Finally, the choice of this parametric baseline hazard functions leads to fully parametric marginal likelihood, such that we can make use of maximum likelihood techniques to estimate the parameters. In contrast, for Cox models with log-normal random effects, the asymptotic properties of the penalized partial likelihood approach are not yet well established. We will discuss the motivation for using the log-normal distribution for the random effects in the next subsection.

3.3.2 Log-normal random effect

Another important consideration in center effect studies is that the comparison of center-specific outcomes should be based on risk-adjusted models that account for center-specific heterogeneity with respect to the types of patients treated. For such purposes, a fixed effect model with centers as indicator variables is a simple choice. Despite its ease of implementation, a common limitation for this model is the assumption of equal coefficients across centers, which is often violated in practice. For instance, variations may exist among centers in both the baseline risk and the effectiveness of follow-up therapy. In other words, just as it is reasonable to think that baseline survival may differ across centers, it is possible that center-specific covariate effects may also be unequal. Evaluations of such variations are important to determine how a particular therapy should be administered. This concern can be addressed using a center-specific fixed effect model. However, this approach is

attractive only if the sample sizes across centers are large relative to the number of centers. In our motivating example, as is often seen in nation-wide studies, there are many centers with relatively few patients. Thus, fixed effect models are not feasible in our setting because the estimated regression coefficients may be quite imprecise. For this reason, alternative methods for estimating center effects are needed. An alternative to a fixed effects approach is the random effect or frailty model, in which subjects in a center share some common effects, with center-specific effects treated as a sample from a specific probability distribution. Under this framework, the number of parameters does not increase with the number of centers. Note that our motivating example includes not only a large number of centers, but also some centers with small sample sizes. A random effect approach is appealing from this perspective, since such an approach is more robust to small center sizes and allows all such centers to be included in the analysis.

A wide variety of random effect models have been studied in survival analysis. Among them, the gamma frailty model [Clayton (1978); Clayton and Cuzick (1985)] and log-normal frailty model [McGilchrist (1993); McGilchrist and Aisbett (1991); Yamaguchi and Ohashi (1999)] are the most extensively studied approach. In particular, the log-normal random effect model is applied in our study, the reason being that interactions between covariate and random effects can be readily implemented. For instance, in our motivating example, we have interest in both the center-specific follow-up therapy effects as well as the baseline risks.

3.4   Estimation

First, we provide the notation to be used in this article. Let T_i and C_i represent the survival and censoring time, respectively, for the i'th patient, where $i = 1, ..., n$.

Observation times are denoted by $X_i = T_i \wedge C_i$, where $a \wedge b = \min\{a, b\}$. Correspondingly, we set the observed death indicators to $\Delta_i = I(T_i \leq C_i)$. Let J be the number of centers, with the total number of subjects denoted by $n = \sum_{j=1}^{J} n_j$, where n_j is the number of subjects in center j. Each subject is characterized by a time-constant covariate vector, Z_i. Correspondingly, let β be the fixed effect coefficient vector for Z_i. Let G_i denote the center for subject i. Let $a_i(t)$ be a time-dependent covariate for attained-age; β_a is the fixed effect coefficient for attained-age. Let $b = (b_0, b_1)$ be an vector of random effect, where $b_0 = (b_{01}, \cdots, b_{0J})$ and $b_1 = (b_{11}, \cdots, b_{1J})$ are vectors of random intercepts and random slopes, respectively, assumed to follow independent normal distributions with $b_0 \sim N(0_{J \times 1}, \theta_0 \mathbf{I}_{J \times J})$ and $b_1 \sim N(0_{J \times 1}, \theta_1 \mathbf{I}_{J \times J})$.

Our proposed estimation procedure involves two stages. In Stage 1, we obtain parameter estimates under the assumed parametric model, then compute subject- and center-specific fitted survival functions. In Stage 2, we estimate center-specific mean survival time by averaging across the marginal covariate distribution using fitted values based on parameters obtained from Stage 1. Details are discussed in the following subsections.

3.4.1 Stage 1: Model and parameter estimation

To approximate the baseline hazard by piecewise constants, we divide the observed time period into K follow-up time intervals with cutpoints $0 = t_0 < t_1 < \cdots < t_K = \infty$. Within each of the intervals, the hazard function, $\lambda_0(t)$, is assumed to be:

$$\lambda_0(t) = \lambda_k, \ t \in \Omega_k \equiv [t_{k-1}, t_k), \ k = 1, \cdots, K$$

Thus, the baseline hazard is constant within each interval, but is allowed to vary across intervals. The choice of intervals is arbitrary. However, as shown by Lawless

(1998), models using a piecewise constant baseline hazard with suitable intervals often yield accurate estimates for fixed effect and frailty parameters.

One advantage for piecewise exponential model is that it can easily accommodate time-varying covariates provided that they change values only at interval boundaries. In our study, the only time-varying covariates is attained-age. To accommodate the change of attained-age, we further split each of the time interval Ω_k into multiple subintervals Ω_{kq}, where $q = 1, \cdots, k_q$ and k_q is the number of subintervals within Ω_k. The values of attained-age is then defined as the age at the left boundary of each subinterval. The boundaries of subintervals are chosen as either the cut points for original interval or the time scales corresponding to an integer value of attained-age. Note that by this way each subinterval get its own measure of exposure and its own death indicator, but all subintervals would be tagged as belonging to the same interval, so they would get the same baseline hazard. This procedure may be best explained by way of example. For individual i with age-at-transplant equal to 18.3 and with 3 year follow-up time, we can split the particular intervals $[0, 1)$, and $[1, 3)$ hazard functions into subinterval $[0, 0.7), [0.7, 1), [1, 1.7), [1.7, 2.7), [1.7, 2.7)$ and $[2.7, 3)$, and the corresponding values of attained-age are recorded as step functions: $a_i(t) = 18.3, 19, 19.3, 20, 21$ and 21.3 for each sub-interval.

The assumed proportional hazards frailty model is formulated as

$$(3.1) \qquad \lambda_{ij}(t) = \lambda_k \exp \left\{ Z_i' \beta + b_{0j} + a_i(t)(\beta_a + b_{1j}) \right\},$$

for $t \in \Omega_k$, where $\lambda_{ij}(t) \equiv \lambda(t|Z_i, a_i(t), G_i = j, b_j)$ denotes the hazard function for the subject i in time interval k conditional on $Z_i, a_i(t), G_i = j$, and $b_j = (b_{0j}, b_{1j})$. Based on the assumed model, $\lambda_{ij}(t)$ is a piece-wise function of follow-up, t. Note that the attained-age are step functions over time: $a_i(t) = a_{ikq}$ for $t \in \Omega_{kq}$.

The main reason we consider a time-dependent age (i.e. attained age) instead of time-constant age (age at transplant) is that age is one of the most important risk factors for long-term studies, and separating contributions of attained age and follow-up time is important to determine how therapy procedures affect patient outcomes. This is particularly relevant in long-term follow-up studies where there may be considerable changes in the mortality pattern. For example, in the context of our motivating example, preliminary analysis indicated that the trends associated with follow-up time diminish considerably as follow-up time increases. In contrast, the effect of attained age plays a more important role as age increases (details will be provided in Section 3.5). Since one of the concepts motivating our work is that trends over time may differ by center, it then makes sense to allow for the effect of attained age to differ by center. We therefore allow attained age, $a_i(t)$, to be associated with both fixed effect and random effects. Finally, it should be noted that the variable $a_i(t)$, although inherently time-dependent, is a fixed-path covariate and, hence, an external time-dependent covariate (Kalbfleisch and Prentice 2002). If internal covariates were being considered (i.e. $Z_i(t) \neq Z_i$, $Z_i(t)$ measurable only if $T_i \geq t$), then the methods we propose in this report do not apply.

Under independent and noninformative censoring, Holford (1980) and Laird and Oliver (1981), noted that the piece-wise exponential model is equivalent to a certain Poisson regression model. Moreover, the conditional frailty likelihood of center j (i.e., conditional on b_j) has been shown to be proportional to the conditional likelihood of a Poisson regression model with count either 0 or 1 by Andersen et al. (1989). As a consequence, the conditional likelihood is given by

$$L(\lambda_1, \cdots, \lambda_K, \beta; b_j) \propto \Pi_{i=1}^{n_j} \Pi_{k=1}^{K} \Pi_{q=1}^{K_q} \left\{ \lambda_k t_{ikq} \exp(X_i'\beta + b_{0j} + a_{ikq}(\beta_a + b_{1j})) \right\}^{\Delta_{ikl}}$$

$$\exp\left\{ -\lambda_k t_{ikq} \exp(X_i'\beta + b_{0j} + a_{ikq}(\beta_a + b_{1j})) \right\},$$

where t_{ikq} is the total time for subject i in subinterval Ω_{kq} and Δ_{ikq} is the event indicator for subject i in subinterval Ω_{kq}. Assuming that the J centers are independent, the conditional likelihood across all the centers is given by

$$L(\lambda_1, \cdots, \lambda_K, \beta, b) = \prod_{j=1}^{J} L(\lambda_1, \cdots, \lambda_K, \beta, b_j).$$

Then, the marginal likelihood has the form

$$L(\lambda_1, \cdots, \lambda_K, \beta, \theta) = \prod_{j=1}^{J} \int \exp\{g(\lambda_1, \cdots, \lambda_K, \beta, \theta)\} db_j,$$

where

$$g(\lambda_1, \cdots, \lambda_K, \beta, \theta) = \ell(\lambda_1, \cdots, \lambda_K, \beta, b_j) - \frac{1}{2} b_j' D^{-1}(\theta) b_j - \log\left\{ \frac{1}{\sqrt{(2\pi)^d |D(\theta)|}} \right\}.$$

The objective function for numerical minimization is twice the negative of the corresponding log likelihood approximation. For given $\widehat{\theta}$, the estimating equations come from setting the derivative of $\ell(\lambda_1, \cdots, \lambda_K, \beta, \widehat{\theta})$, with respect to the parameters $(\lambda_1, \cdots, \lambda_K, \beta)$, equals zero. One difficulty is that the marginal likelihood involves high dimensional integrals without closed forms. To simplify the computations, we use a Laplace approximation to approximate each such integral with a function that has a closed form. When the center-specific sample sizes are large, the Laplace approximation to the marginal log likelihood is as follows,

$$\prod_{j=1}^{J} \left\{ n_j g(\lambda_1, \cdots, \lambda_K, \beta, \widehat{\theta}) + \frac{R}{2} \log(2\pi) - \frac{1}{2} \log|-n_j g''(\lambda_1, \cdots, \lambda_K, \beta, \widehat{\theta})| \right\},$$

where R is the dimension of random effect and

$$g''(\beta, \lambda_1, \cdots, \lambda_K, \widehat{\theta}) = \left. \frac{\partial^2 g(\beta, \lambda_1, \cdots, \lambda_K, \theta)}{\partial \theta \partial \theta'} \right|_{\widehat{\theta}}.$$

To complete the singly iterative estimation process, θ satisfies the first-order condition:

$$\frac{\partial g(\beta, \lambda_1, \cdots, \lambda_K, \theta)}{\partial \theta} = 0,$$

for given values of β and λ (Vonesh et al. 2002). Further derivatives can be found in Feng et al. (2005). Vonesh et al. (1996) show that the maximum likelihood estimator based on the Laplace approximation is a consistent estimator with order rate of convergence depending on both the number of centers and the number of subjects within centers. The influences of J and n_j on our proposed estimators will be evaluated in Section 3.

3.4.2 Stage 2: Difference in mean survival time

It is well known that patient characteristics may vary significantly across centers. These factors can have a substantial impact on the expected center-specific survival outcomes. Accurate estimation of center effects must account for potential covariate imbalances across centers. To address this concern, we use a technique based on direct standardization. Specifically, we define the difference in mean survival time corresponding to the center j as

$$\delta_j = \mu_j - \mu.$$

where $\mu_j = E[E(T_i|Z_i, G_i = j)]$, with the outer expectation taken with respect to the marginal distribution of Z_i, and $\mu = E(T_i)$. An estimator for δ_j is given by

$$\widehat{\delta}_j = \widehat{\mu}_j - \widehat{\mu}.$$

Note that μ_j has the interpretation of mean survival time for the population, under the hypothetical scenario wherein all subjects were treated in center j. We estimate μ_j through

$$\widehat{\mu}_j = n^{-1} \sum_{i=1}^{n} \int_0^\infty \widehat{S}_{ij}(u)du,$$

where $S_{ij}(u) = P(T_i > u | Z_i, G_i = j)$ and

$$\widehat{S}_{ij}(u) = \exp\left\{ -\sum_{k=1}^{K} \sum_{q=1}^{k_q} \widehat{\lambda}_k * (t_{ikq} \wedge u) * \exp\{Z_i'\widehat{\beta} + \widehat{b}_{0j} + a_{ikq}(\widehat{\beta}_a + \widehat{b}_{1j})\} \right\}.$$

In contrast, we estimate μ using

$$\widehat{\mu} = n^{-1} \sum_{\ell=1}^{J} \sum_{i=1}^{n_\ell} \int_0^\infty \widehat{S}_{i\ell}(u)du,$$

which has the interpretation of estimated mean survival time for the population, where $\widehat{S}_{i\ell}(u)$ is the estimated survival for a subject i in center ℓ. Note that based on the definition of mean survival time, the integral of survival function should go to ∞. However, in practice, it is more reasonable to calculate this integral until patient achieve certain age. We will provide more detail in Section 5 about how we choose this upper limit for our analysis. $\widehat{\delta}_j = \widehat{\mu}_j - \widehat{\mu}$ amounts to direct standardization, since the averaging in each of $[\delta_1, \cdots, \delta_J]$ is with respect to the marginal covariate distribution. Therefore, this approach accounts for potential covariate imbalances across centers and allows for valid comparison between centers.

3.4.3 Variance estimation

The proposed Difference of Mean Survival Time, $\widehat{\delta}_j = f(\widehat{\beta}, \widehat{b})$, a function of fixed effects and random effects. Therefore, the Delta method can be used to obtain a variance estimator for $\widehat{\delta}_j$ based on the variance-covariance matrix for fixed effects and random effects $[\widehat{\beta}, \widehat{b}]'$. Specifically, let the matrix, $C = \widehat{V}([\widehat{\beta}, \widehat{b}]')$, be the variance

matrix for $[\widehat{\beta}, \widehat{b}]'$, and let M be the vector of derivative of Difference of Mean Survival Time with respect to $[\widehat{\beta}, \widehat{b}]'$. Using Delta method, $M'CM = \widehat{V}(\widehat{\delta}_j)$, is the desired variance estimator for $\widehat{\delta}_j$.

3.4.4 Implementation of proposed methods

To fit the model and obtain the parameter estimators in Stage 1, we implement the Laplace approximation using the SAS procedure GLIMMIX, with the option METHOD=LAPLACE for the required integral approximation. The mean survival time in Stage 2 is a numerical integration of scalar functions in one dimension over connected finite intervals. We programmed this estimation process using SAS function QUAD. The QUAD call is an adaptive global-type integrator that produces a quick, rough estimate of the integration result and then refines the estimate until achieving the prescribed accuracy.

With some simple data manipulations, the proposed variance estimators can be obtained from the SAS procedure GLIMMIX, which forms the approximation prediction variance matrix, C, for $[\widehat{\beta}, \widehat{b}]'$. However, this variance covariance matrix, however, cannot be outputted directly. To solve this problem, we make use of the ESTIMATE statement and propose the following estimation procedure. With a constant-specifications $\langle$ fixed effect values ...$\rangle$ $\langle$ | random effect values ...$\rangle$, the ESTIMATE statement of SAS GLIMMIX procedure constructs the vector $M' = [M_1' M_2']$, where M_1 is a vector of derivative of Difference in Mean Survival Time with respect to the fixed effect parameters, β, and M_1 is a vector of derivative of Difference in Mean Survival Time with respect to the G-side random effect parameters, b. Then the $M'CM$ output from the ESTIMATE statement is the desired variance estimator for $\widehat{\delta}_j$.

3.5 Simulation Study

The finite-sample properties of the estimators described in previous section were evaluated through a series of simulation studies. We considered $J = 10$, $J = 25$, and $J = 50$ centers. Death times were generated from the piecewise exponential model. To mimic the motivating example, we divided the whole follow-up period into the following intervals: $[0, 1]$, $(1, 2]$, $(2, 3]$, $(3, 5]$, $(5, 10]$, $(10, \infty]$. For the kth interval, the hazard function for subject i in center j is given by formula (3.1), where X_i followed a Bernoulli distribution. We set $\beta_0 = -0.1$; $a_i(t) = a_{0i} + t$, where a_{0i} came from a Normal distribution with constant variance and center-dependent means, and t is the follow-up time. Random effects b_{0j} and b_{1j} $(j = 1, \cdots J)$ were generated from independent Normal distributions $N(0, \theta_0)$ and $N(0, \theta_1)$, respectively, with $\theta_0 = 0.25$ and $\theta_1 = 0.005$. We varied the sample size for each center as $n_j = 25, 50, 100$. The censoring percentages were approximately 20%. Each data configuration was replicated 500 times. We used 95% prediction interval in all simulations. We choose the same upper limit as that in Section 5 to calculate the mean survival time for each observation (until the age 85).

In Table 1, we list the results for fixed effect and random effect parameters. The fixed effect parameter estimators were sufficiently well-behaved, in the sense that the bias was quite small, and the average asymptotic standard errors (ASE) were generally close to the empirical standard deviations (ESD), while the empirical coverage probabilities (CP) were generally consistent with the nominal value 0.95. The performance of the random effect parameters depended on sample size. If the sample size was small, we failed to obtain the random effect parameters in certain number of replicates, and hence, the corresponding CP is poor. It is worth noting that, for such

replicates, even the random effect variance estimators could not be computed, we still obtained the point estimators b_{0j} and b_{1j} for $j = 1, \cdots J$, which lead to reliable estimation for our proposed measures (Table 2). We reported the performance of proposed estimators in Table 2. The ASE estimators for were sufficiently accurate and the CP were generally consistent with the nominal value 0.95. The results for setting with $J = 10$ or $J = 25$ centers are not shown. In general, for fixed J, as n_j increase, the CP increase. On the other hand, for fixed n_j, as J increase, the CP also increases. Overall, the simulation results indicate that the method is performing in a reasonable manner.

3.6   Application to Kidney Transplant Data

To separately examine the effects of follow-up time and attained age, we first fitted two preliminary piece-wise exponential models. In the first model, the effect of attained age was approximated as piece-wise linear (Figure 1A). In the second model, attained age was treated as time axis and follow-up time was considered as categorical covariate. The corresponding effect of follow-up time was shown in Figure 1B, which suggests that the hazards associated with follow-up stabilized as follow-up time increased. This finding motived us to assume that the hazard functions remains to be constant after the maximum follow-up time. Note that this choice evolves extrapolation outside of oberservation range and is an arbitrary choice for our particular dataset. In a more general situation, we may use linear function for extrapolation. Another preliminary analysis we performed is to choose a reasonable upper limit for calculating the mean survival time. The maximum observed value for attained-age in our data is 96. However, analysis based on person-year indicated that only a few observations had attained age larger than 85 ($< 0.005\%$). Therefore, we calculate the mean survival time for each patient until they achieve age 85.

In the next step, we performed a log-normal frailty model with piece-wise exponential baseline rate. Specifically, we choose 6 intervals for time. A random effect

Table 3.1: Simulation results: Performance of fixed effect and random effect parameters.

J, n_j	Parameters	TRUE	BIAS	ESD	ASE	CP
$J = 50, n_j = 100$	λ_0	0.160	0.002	0.020	0.019	0.94
	λ_1	0.130	0.002	0.017	0.016	0.95
	λ_2	0.100	0.001	0.013	0.012	0.93
	λ_3	0.090	0.001	0.012	0.011	0.93
	λ_5	0.085	0.001	0.011	0.011	0.93
	λ_{10}	0.105	0.002	0.015	0.014	0.93
	β_1	0.100	-0.001	0.019	0.018	0.94
	β_2	-0.100	0.001	0.032	0.033	0.97
	θ_1	0.250	-0.02	0.079	0.078	0.92
	θ_2	0.005	-0.000	0.002	0.002	0.95
$J = 50, n_j = 50$	λ_0	0.160	0.003	0.023	0.024	0.95
	λ_1	0.130	0.002	0.020	0.020	0.94
	λ_2	0.100	0.001	0.013	0.012	0.93
	λ_3	0.090	0.002	0.014	0.014	0.96
	λ_5	0.085	0.002	0.013	0.014	0.96
	λ_{10}	0.105	0.003	0.018	0.018	0.95
	β_1	0.100	-0.000	0.023	0.024	0.96
	β_2	-0.100	-0.000	0.049	0.047	0.96
	θ_1	0.250	0.002	0.097	0.097	0.94
	θ_2	0.005	-0.002	0.003	0.003	0.95
$J = 50, n_j = 25$	λ_0	0.160	0.004	0.032	0.032	0.94
	λ_1	0.130	0.003	0.028	0.027	0.93
	λ_2	0.100	0.002	0.021	0.022	0.95
	λ_3	0.090	0.003	0.020	0.020	0.93
	λ_5	0.085	0.002	0.019	0.019	0.93
	λ_{10}	0.105	0.003	0.026	0.025	0.93
	β_1	0.100	0.000	0.033	0.033	0.97
	β_2	-0.100	-0.000	0.067	0.068	0.96
	θ_1	0.250	-0.005	0.121	0.124	0.92
	θ_2	0.005	0.000	0.004	0.004	0.88
$J = 50, n_j = 10$	λ_0	0.160	0.009	0.054	0.052	0.94
	λ_1	0.130	0.007	0.045	0.044	0.94
	λ_2	0.100	0.005	0.036	0.035	0.95
	λ_3	0.090	0.005	0.032	0.031	0.94
	λ_5	0.085	0.004	0.032	0.030	0.93
	λ_{10}	0.105	0.010	0.044	0.043	0.94
	β_1	0.100	0.000	0.053	0.053	0.95
	β_2	-0.100	0.004	0.113	0.111	0.95
	θ_1	0.212	-0.048	0.151	0.187	0.86

Table 3.2: Simulation results: Performance of center effect estimators; $\bar{\bar{\delta}}$ = average of δ_j over $j = 1 \cdots J$.

J, n_j	Parameters	TRUE	BIAS	ESD	ASE	CP
$J = 50, n_j = 100$	δ_1	-5.273	-0.108	0.212	0.234	0.93
	δ_5	-4.180	0.074	0.296	0.322	0.98
	δ_{15}	-2.533	0.146	0.473	0.439	0.93
	δ_{25}	-0.819	0.020	0.613	0.565	0.93
	δ_{35}	1.219	-0.049	0.782	0.737	0.94
	δ_{45}	5.046	-0.164	1.112	1.045	0.94
	δ_{50}	11.284	-0.252	1.396	1.372	0.94
	$\bar{\bar{\delta}}$	0.000	-0.000	0.740	0.636	0.94
$J = 50, n_j = 50$	δ_1	-5.225	-0.084	0.330	0.333	0.94
	δ_5	-4.110	0.081	0.426	0.455	0.97
	δ_{15}	-2.500	0.168	0.616	0.615	0.96
	δ_{25}	-0.866	0.088	0.786	0.784	0.96
	δ_{35}	1.193	-0.052	1.099	1.012	0.93
	δ_{45}	5.021	-0.255	1.456	1.422	0.96
	δ_{50}	11.025	-0.261	1.806	1.862	0.96
	$\bar{\bar{\delta}}$	-0.003	0.003	1.006	0.878	0.95
$J = 50, n_j = 25$	δ_1	-5.168	-0.036	0.449	0.482	0.96
	δ_5	-4.040	0.131	0.590	0.649	0.98
	δ_{15}	-2.479	0.239	0.872	0.859	0.95
	δ_{25}	-0.797	0.035	1.164	1.069	0.94
	δ_{35}	1.232	-0.104	1.481	1.373	0.93
	δ_{45}	4.485	-0.257	1.974	1.892	0.94
	δ_{50}	10.510	-0.258	2.535	2.463	0.96
	$\bar{\bar{\delta}}$	0.000	-0.001	1.384	1.197	0.94
$J = 50, n_j = 10$	δ_1	-5.273	-0.108	0.212	0.234	0.93
	δ_5	-3.854	0.315	1.034	1.030	0.95
	δ_{15}	-2.533	0.146	0.473	0.439	0.93
	δ_{25}	-0.643	-0.022	1.786	1.533	0.93
	δ_{35}	1.219	-0.049	0.782	0.737	0.94
	δ_{45}	4.438	-0.308	2.711	2.550	0.94
	δ_{50}	11.284	-0.252	1.396	1.372	0.94
	$\bar{\bar{\delta}}$	0.013	-0.013	2.010	1.692	0.93

for intercept was chosen for the heterogeneity in baseline risk, and a random effect for attained age was estimated for the heterogeneous effect of age at transplant and follow up time. We then estimate the mean survival time across centers. We also constructed the standard error estimator for the difference of mean survival time based on Delta method described in section 2.4.

We ordered centers based on difference in mean survival time and we provided the plots of confidence intervals in Figure 2. Centers with confidence intervals higher than 0 were identified as significantly better than the national average, while centers with confidence intervals lower than 0 were identified as significantly worse than the national average. A total of 50 centers had mean survival time significantly lower than the expected, while 29 centers were significantly above the average.

3.7 Discussion

Survival analysis plays an important role in the analysis of multi-center studies. Often of particular interest is the determination of which medical centers have significantly better or significantly worse long term outcomes. For instance, the methods we propose in this report are motivated by the need to profile center-specific long-term kidney transplant outcomes in the United States. Existing measures of center-specific performance tend to focus on short- and medium-term outcomes. However, long term center-specific comparisons are often of great interest. In this report, we propose methods which combine a log-normal frailty model and piece-wise exponential baseline hazard to compare mean survival time across centers. The Laplace approximation for integration is applied in order to obtain maximum likelihood estimations in a computationally tractable fashion. Our proposed methods can be reliably used for multi-center studies with a large number of centers, even some centers have small sample sizes. The distinguishing feature of our methods is that they allow for prediction of the overall mean survival time for each center and, within this context, consistently profiling of long-term center-specific outcomes.

The Cox model has been widely used in survival analysis setting. In this report,

we did not use this model because the mean survival time may not be well estimated, due to the fact that the survival function may not drop to zero in the presence of censoring. In contrast, we predict mean survival time using parametric baseline hazard functions. For simplicity, we suggest to use a pre-specified number of time intervals for the piece-wise exponential model. An alternative method would be to choose an optimized grid points based on tree methods (Demarqui et al. 2008; Huang et al. 1998). This approach is not applied in this report because the motivating data set is extremely large, meaning that numerical calculation would be very cumbersome.

Random effect models have received a lot of attention for the analysis of multi-center survival data. One of the main problems in the application of random effect models to real data is the limited availability of standard software. Thus, of special interest in this report is how existing standard software and well established large-sample theory can be applied to real data. In this study, we take advantage of the connection between the log-normal frailty model and the generalized linear mixed model. The proposed estimators and their corresponding variance estimators can be easily computed through the SAS procedure GLIMMIX. The major difference between our proposed method compared to existing applications of frailty models is that previous studies mainly focus on the appropriate evaluation of main effect (e.g., treatment effect) in the presence of center effects. In other words, the main interest of these studies is to remove the potential confounding effect of centers from the analysis. There are not so many studies which estimate the center effects themselves using random effects. Although random effect estimators of baseline risk and their exponential transformations have been used to describe the relative center effects, these estimator are intuitively difficult to understand for non-statisticians. Moreover, in addition to the variation in the baseline risk, there may also exist variation in some important covariates across centers. Such variation should also be taken into consideration to accurately estimate center effects. For instance, in our motivating example, centers may use a variety of therapy procedures. Therefore, it is desirable to determine how the collection of therapy procedure affects patient outcomes. To

our knowledge, how to make use of the combination of these multivariate random effect estimators to evaluate center effects has not been addressed by previous studies. In this report, we illustrated both theoretically and through simulations that log-normal frailty model with piece-wise exponential baseline rates provides a simple and interpretable summaries to evaluate center effects directly. Furthermore, since a large number of centers will be examined, some centers may be identified as outlier centers by chance alone, due to multiple-testing issues. With respect to this concern, random effect model has advantages comparing to fixed effect models since the previous methods are shrunken estimator closer to the mean level and, thus, are more conservative and less likely to classify centers as outliers.

Finally, it is worth noting that another technique to compare the mean survival time across center is the indirect standardization method, in which the difference of the mean survival time for each center is defined as the observed minus the expected mean survival time, where the expected mean survival time for a center is calculated assuming the subjects from this center have the same hazards and center-specific slopes as the average. This estimator averages with respect to the covariate distribution within each center and hence, it is sensitive to center-specific covariate distributions. Given this property, indirect standardization is a useful tool to quantify center performance relative to the national average. However, if the goal is to order centers based on performance, direct standardization is a more appropriate measure.

One of the key issues with the piece-wise exponential model involves determining the appropriate number of time intervals to be used. The number of time intervals is something that must be determined by the analyst. Although any number of time periods can be chosen, it is important to recognize that there is always a tradeoff to be made. If one chooses a large number of time periods, then we will get a better approximation of the unknown baseline hazard but we will have to estimate a larger number of coefficients and this may cause problems. Alternatively, if one chooses a small number of time periods, then there will be fewer estimation problems but

the approximation of the baseline hazard will be worse. A key requirement when choosing the number of time periods is that there should be units that fail within each of the different time intervals. If this is not the case, then one will not obtain sensible estimates.

Figure 3.1: Confidence Interval Plots.
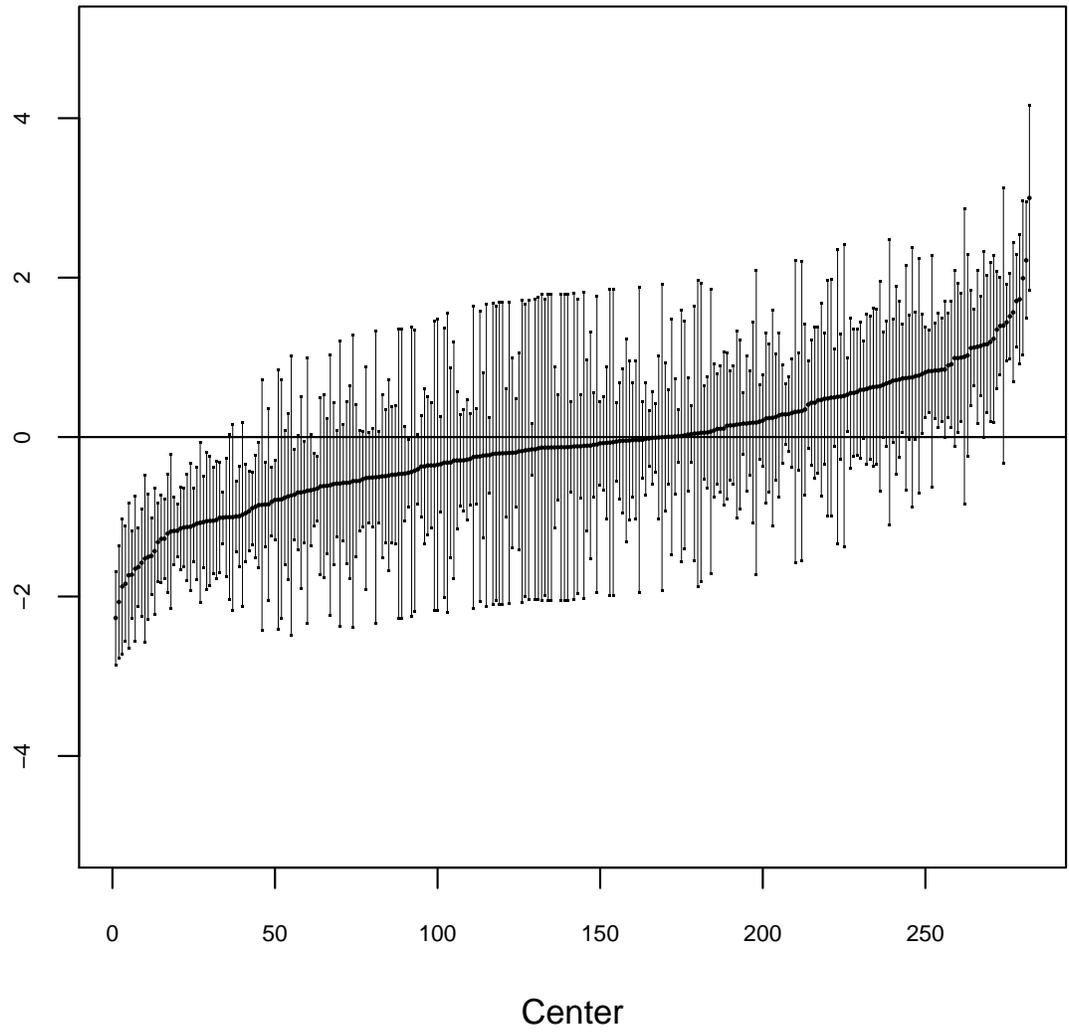
CHAPTER IV

Semi-parametric Methods for Relative Risk Center Effect Measures

Although multiplicative hazard models are used most frequently in biomedical studies, additive hazard models are available and may be preferable in certain settings. In the third chapter, we develop methods for evaluating center-specific survival using a center-stratified additive hazards model. We propose to estimate the relative center effects by the ratio of baseline survival functions. The proposed measure is a semiparametric generalization of the relative risk, which is often used in clinical studies. An attractive property of our proposed method is that the ratio of survival functions is invariant to the adjustment covariate. That is, the ratio of baseline survival functions represents the contrast between subject i at center j versus subject i at the hypothetical center with baseline hazard function equal to the national average; where subject i can have any covariate value. We derive the asymptotic properties of the proposed estimators, and assess finite-sample characteristics through simulation. The proposed methods will be applied to national kidney transplant data.

4.1 Introduction

In medical studies, when time to a failure event is of interest, the effect of the groups is often estimated by comparing the survival functions across groups. The survival function is more interpretable to non-statisticians than other measures, such as hazard function. In the setting of clinical trials, KaplanMeier estimator is a com-

monly used measure to compare survival functions across groups. In observation studies, the groups effect represents the impact of the factors that are center specific and are not accounted for the adjustment covariates. Therefore, an accurate comparison of groups needs to adequately account for the imbalance of risk factors across groups. It is well known that patient characteristics may vary considerably across groups. The inclusion of some high-risk patients at a given group can make that group's survival appear sub-standard. Thus, valid measures of group effects need to adjust for the covariate imbalance across groups.

The motivating example for this study is to evaluate the center-specific survival outcomes of kidney transplant centers. The appropriate analysis of center-specific outcome has been a subject of much discussion. Substantial variations may exist in patient survival among transplant centers. Centers whose observed mortality is significantly higher than expected are highlighted and face both public and professional scrutiny regarding the potential reasons for their deviation from expected performance. Accurate measures of center effect are needed.

There are mainly two types of models for studying center effects. The methods commonly used to evaluate centers are based on relative hazard models (e.g. Cox (1972, 1975)), for which the covariate effects are multiplicative. An alternative approach is the additive hazard model, in which covariate effects are expressed through hazard differences. Previous studies(e.g., Breslow and Day (1980); Cox and Oakes (1984); O'Neill (1986)) have shown that use of the proportional hazards model can result in serious bias when the additive hazards model is correct.

Additive hazards models have been studied by several authors. For instance, Aalen (1980, 1989) proposed nonparametric methods for additive hazard models. Lin and Ying (1994, 1995) proposed a semiparametric additive model and provided an explicit parameter estimators. Extensions of the additive hazards model which allow time-varying coefficients have been studied by several authors. Huffer and Mckeague (1991) studied weighted least squares estimation for a nonparametric additive risk model. McKeague and Sasieni (1994) developed a partly parametric additive hazards

model that includes both time-dependent and constant regression coefficients. Other extensions include marginal additive hazard models for multivariate survival data (Yin and Jianwen Cai 2004) and additive transformation models for alternative link function other than linear (Zeng and Cai 2010).

Lin and Ying (1994)'s models have received a lot of attention for the analysis of survival data. One reason for the popularity is that it is computational convenience and it is a natural analogue of the Cox model. Based on the Lin and Ying (1994) model, one approach that can be used to estimate center effects is to code center using indicator variables. This simple approach has drawbacks when applied in studies of center-specific mortality. For example, the choice of reference center is arbitrary. Different choice of the reference center may had a strong impact on the results, which would of course lead to controversy. Another limitation is that this simple fixed center effect models assume that the center effects are constant over time, which is often violated in practice. If the number of centers is small, we may apply the extensions of the additive hazards model with time-varying coefficients. However, the number of centers is often very large for multi-centers studies. The number of regression coefficients increases as the number of centers increases, in violation of the regularity conditions upon which the asymptotic theory for the model is based. Moreover, it is difficult to estimate a large number of time-dependent parameters simultaneously. To address these limitations, we propose to estimate center effects through a stratified additive hazard model.

The remainder of this chapter is organized as follows. The proposed model and method are described in the next section. Asymptotic properties are derived in Section 3. Finite-sample properties are examined in Section 4 through simulation studies. Section 5 applies the proposed methods to kidney transplant data from a national organ failure registry. We provide some discussion of the proposed and related methods in Section 6.

4.2 Proposed methods

First, we provide the notation to be used in this article. Let T_i and C_i represent the survival and censoring time, respectively, for the i'th patient, where $i = 1, ..., n$. 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 1 when condition A holds and 0 otherwise. The observed death indicators are denoted by $\Delta_i = I(T_i \leq C_i)$, and the death counting process is defined as $N_i(t) = \Delta_i I(X_i \leq t)$. 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, $(X_i, \Delta_i, G_i, Z_i)$, where Z_i is a vector of adjustment covariates.

4.2.1 Excess mortality model

Andersen and Væth (1989) studied an additive hazard (excess mortality) model, assuming the death intensity for individual i at center j, $\lambda_{ij}(t)$, is the sum of a known standard mortality hazard, $\lambda_i(t)$, and an unknown center-specific component, $\theta_j(t)$,

$$\lambda_{ij}(t) = \lambda_i(t) + \theta_j(t).$$

It is useful to generalize this model from two directions. First, it would be desirable to relax the assumption with a known covariate-specific standard mortality. For instance, if the standard mortality is treated as fixed rather than random, the variation of the standard mortality is not taken into consideration. Second, patient characteristics may vary considerably across centers. The inclusion of some high-risk patients in a center can make its survival appear poor. Therefore, a fair comparison of center effect needs to adequately account for the imbalance of risk factors among centers. To address these concerns, we propose a stratified additive hazard model.

4.2.2 Stratified additive hazards model

The stratified additive hazard model can be formulated as

$$(4.1) \qquad \lambda_{ij}(t) \equiv \lambda(t|Z_i, G_i = j) = \lambda_{0j}(t) + \beta^T Z_i,$$

where $\lambda_{0j}(t)$ is an unspecified center-specific baseline hazard function. Note that in formula (4.1), $\lambda_{0j}(t)$ is an unspecified center-specific baseline hazard function. Therefore, the model assumes only that the covariate effects are additive, but does not require any functional form for the relationship among the hazard functions across censers. If the baseline component is also additive,

$$(4.2) \qquad \lambda_{0j}(t) = \lambda_0(t) + \theta_j(t),$$

where the baseline hazard for center j, is the sum of unknown center common baseline and an unknown center-specific component $\lambda_{0j}(t) = \lambda_0(t)$, and an unknown center-specific component, $\theta_j(t)$. Integrating both sides of the formula (4.2), we obtain

$$(4.3) \qquad \Lambda_{0j}(t) = \Lambda_0(t) + \Theta_j(t),$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ and $\Theta_j(t) = \int_0^t \theta_j(u)du$.

4.2.3 Center effect measure

We then evaluate center effects using the ratio of baseline survival function $\psi_0 = [\psi_{10}, \cdots, \psi_{J0}]^T$, where

$$(4.4) \qquad \psi_{0j}(t) = \frac{S_{0j}(t)}{S_0(t)},$$

where $S_{0j}(t) = \exp\{-\Lambda_{0j}(t)\}$ is the baseline survival function for center j, and $S_0(t)$ is the baseline survival function for a hypothetical center with baseline survival function equal to the average across all centers:

$$(4.5) \qquad S_0(t) = \sum_{\ell=1}^{J} p_\ell S_{0\ell}(t),$$

where $p_l = P(G_i = \ell)$ is the probability that subject i belongs to censer j. A simple estimator for p_l can be defined as $\widehat{p}_l = n_\ell/n$.

One advantage for our proposed method is that, given the additive hazard structure, the ratio of survival function, for a particular subject reduces to the ratio of baseline survival functions. For instance, we can define $S_{ij}(t) = P(T_i > t | Z_i, G_i = j) = \exp\{-\Lambda_{0j}(t) - \beta^T Z_i t\}$, the survival function for subject i in center j; and define $S_{i0}(t) = P(T_i > t | Z_i, G_i = G_0) = \sum_{\ell=1}^{J} p_\ell \exp\{-\Lambda_{0\ell}(t) - \beta^T Z_i t)\}$, to be the survival function for subject i if this subject is treated at a hypothetical center, G_0, with baseline survival function equal to the average across all centers. Then, considering the ratio of baseline survival functions

(4.6)
$$\psi_{ij}(t) = \frac{S_{ij}(t)}{S_{i0}(t)},$$

terms involving subject-specific covariates will cancel out in the ratio. Hence, $\psi_{ij}(t) = \psi_{0j}(t)$, the ratio of survival function for a particular subject reduces to the ratio of baseline survival functions at the center level. In this light, the proposed measure has an interpretation at the individual level, which may be desirable for patients and perhaps investigators.

Consider formula (4.3),

$$\psi_{0j}(t) = \frac{\exp\{-\Lambda_0(t) - \Theta_j(t)\}}{\sum_{\ell=1}^{J} p_\ell \exp\{-\Lambda_0(t) - \Theta_\ell(t)\}},$$

the average baseline survival function defined in formula (4.5), lead to the following constrain for relative center effects

$$\sum_{\ell=1}^{J} p_\ell \exp\{-\Theta_\ell(t)\} = 1.$$

Thus, the proposed estimator and the relative center effect has the relationship:

$$\psi_{0j}(t) = \exp\{-\Theta_j(t)\}$$

Note that the excess cumulative hazard model can be transformed to relative survival function model. In other words, model (4.3) could be represented as

(4.7)
$$S_{0j}(t) = S_0(t)\rho_j(t),$$

where $\rho_j(t) = \exp\{-\Theta_j(t)\}$. Therefore, $\psi_j(t)$ is also an estimator for the relative survival function $\rho_j(t)$. We then have the following individual level relationship:

(4.8)
$$S_{ij}(t) = S_{i0}(t)\rho_j(t).$$

4.2.4 Estimation of center effects

Extending Lin and Ying (1994) estimation approach, we propose the following estimation function to estimate β,

$$(4.9) \qquad U(\beta) = \sum_{j=1}^{J} \sum_{i=1}^{n} \int_{0}^{\tau} \{Z_i - \overline{Z}_j\}\{dN_{ij}(t) - Y_{ij}(t)\beta^T Z_i dt\},$$

where

$$\overline{Z}_j = \frac{\sum_{i=1}^{n} Y_{ij}(t)Z_i}{\sum_{i=1}^{n} Y_{ij}(t)}$$

The resulting estimator for β takes the form

$$\widehat{\beta} = \widehat{A}^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n} \int_{0}^{\tau} \{Z_i - \overline{Z}_j\}dN_{ij}(t),$$

where

$$\widehat{A} = \frac{1}{n} \left[\sum_{j=1}^{J} \sum_{i=1}^{n} \int_{0}^{\tau} Y_{ij}(t)\{Z_i - \overline{Z}_j\}^{\otimes 2} dt \right]$$

In view of the counting process for each center,

$$N_{ij}(t) = M_{ij}(t) + \int_{0}^{t} Y_{ij}(u)\{d\Lambda_{0j}(u) + \beta^T Z_i dt\},$$

it is natural to estimate $\Lambda_{0j}(t)$ by

$$(4.10) \qquad \widehat{\Lambda}_{0j}(t) = \int_{0}^{t} \frac{\sum_{i=1}^{n}\{dN_{ij}(u) - Y_{ij}(u)\widehat{\beta}^T Z_i du\}}{\sum_{i=1}^{n} Y_{ij}(u)}.$$

Finally, to make sure that the cumulative hazard $\widehat{\Lambda}_{0j}(t)$ is monotone in t, a modification similar to Lin and Ying (1994) can be defined as

$$\widehat{\Lambda}_{0j}^{*}(t) = \max_{s \leq t} \widehat{\Lambda}_{0j}(\widehat{\beta}, s).$$

Note that, with this modification, the invariant-to-baseline property does not hold for finite samples. However, it is asymptotically true since as sample size increase, $\widehat{\Lambda}_{0j}^{*}(t)$ converges uniformly to $\Lambda_{0j}(t)$.

4.3 Asymptotic Properties

To derive the large-sample properties for the proposed measure, we impose the following regularity conditions throughout this chpater.

(a) $(X_i, \Delta_i, G_i, Z_i)$ are independent and identically distributed random vectors.

(b) $P(X_i \geq \tau) > 0$ where τ is a pre-specified time point.

(c) Z_{ik} have bounded total variation, i.e., $|Z_{ik}| < \kappa$ for all $i = 1, ...n$ and $k = 1, ..., p$, where κ is a constant and Z_{ik} is the kth component of Z_i.

(d) $\int_0^\tau \lambda(t) dt < \infty$.

(e) Positive-definiteness of matrix A, where A is the limit of $\widehat{A}$.

(f) $P(G_{ij} = 1 | Z_i) > 0$.

With a simple algebraic manipulation, formula (4.9) yields

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

The strong representation theorem ((Pollard 1990), Theorem 9.4) or the lemma of Lin et al. (2000)) entails that $n^{-\frac{1}{2}} U(\beta_0)$ is essentially a sum of independent and identically distributed random variables, which entails that multivariate central limit theorem can be used to derive the asymptotical distribution. The asymptotic properties of the regression coefficient estimates are given in the following theorem. The proof in this chapter is a direct extension of that provided by Yin and Jianwen Cai (2004).

THEOREM 1: the random vector $n^{\frac{1}{2}} (\widehat{\beta} - \beta_0)$ converges weakly to a random vector with variance-covariance matrix, $\sigma_\beta = E[\xi_{\beta_i} \xi_{\beta_i}^T]$, where

$$\xi_{\beta_i} = A^{-1} U_i(\widehat{\beta}),$$

and A is a positive-definiteness of matrix defined previously. A consistent estimator of the covariance matrix is given by

$$\widehat{\sigma}_\beta = n^{-1} \sum_{i=1}^{n} \widehat{\xi}_{\beta_i} \widehat{\xi}_{\beta_i}^T,$$

with $\widehat{\xi}_\beta$ obtained by replacing limiting values in ξ_β with their empirical counterparts.

We now consider the asymptotic properties of the random vector $n^{\frac{1}{2}}(\widehat{\Lambda}_0 - \Lambda_0)$, where

$$n^{\frac{1}{2}}(\widehat{\Lambda}_0 - \Lambda_0) = n^{\frac{1}{2}}[\{\widehat{\Lambda}_{10} - \Lambda_{10}\}, \cdots, \{\widehat{\Lambda}_{J0} - \Lambda_{J0}\}]^T$$

THEOREM 2: the random vector $n^{\frac{1}{2}}(\widehat{\Lambda}_0 - \Lambda_0)$ converges weakly to a $J-$ dimensional random vector with variance-covariance matrix between $n^{\frac{1}{2}}(\widehat{\Lambda}_{0j}(s) - \Lambda_{0j}(s))$ and $n^{\frac{1}{2}}(\widehat{\Lambda}_{0k}(t) - \Lambda_{0k}(t))$ is $\sigma_{\Lambda jk}(s,t) = E[\xi_{\Lambda j}(s)\xi_{\Lambda k}(t)]$, where

$$\xi_{\Lambda ij(s)} = \int_0^s \frac{dM_{ij}(u)}{\sum_{i=1}^n Y_{ij}(u)} - \{\int_0^s \bar{z}_j(u;\beta)du\}A^{-1}\sum_{j=1}^J \int_0^\tau [Z_i - \bar{z}_j(u;\beta)]dM_{ij}(u;\beta),$$

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

Applying the Functional Delta method (Theorem 20.8, van der Vaart 1998), we have

$$n^{\frac{1}{2}}(\widehat{\psi}_0(t) - \psi_0(t)) = H(t)n^{\frac{1}{2}}\{\widehat{\Lambda}_{0j}(t) - \Lambda_{0j}(t)\} + o_p(1),$$

where $H(t) = [H_1(t), \cdots, H_J(t)]$,

$$H_j(t) = \left\{\frac{\exp\{-\widehat{\Lambda}_{0j}(t)\}}{n^{-1}\sum_{\ell=1}^J n_\ell \exp\{-\widehat{\Lambda}_{0\ell}(t)\}} - \frac{\exp\{-\Lambda_0(t)\}}{\sum_{\ell=1}^J p_\ell \exp\{-\Lambda_0(t)\}}\right\} *$$
$$\left\{\frac{-\exp\{-\Lambda_{0j}(t)\}\sum_{\ell=1}^J p_\ell \exp\{-\Lambda_{0\ell}(t)\} + p_j \exp\{-2\Lambda_{0j}(t)\}}{[\sum_{\ell=1}^J p_\ell \exp\{-\Lambda_{0\ell}(t)\}]^2}\right\}$$

THEOREM 3: the random vector $n^{\frac{1}{2}}(\widehat{\psi}_0 - \psi_0)$ converges weakly to a $J-$ dimensional random vector with variance-covariance matrix between $n^{\frac{1}{2}}(\widehat{\psi}_{0j}(s) - \psi_{0j}(s))$ and $n^{\frac{1}{2}}(\widehat{\psi}_{0k}(t) - \psi_{0k}(t))$ is $\sigma_{\psi jk}(s,t) = E[\xi_{\psi j}(s)\xi_{\psi k}(t)]$, where

$$\xi_{\psi ij}(s) = H_j(s)\xi_{\Lambda ij}(s)$$

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

4.4　Simulation Study

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

$$T_{ij} = \frac{-\log(U_i)}{\lambda_j + \beta^T Z_i}$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, 10$, where U_i is a Uniform(0,1) random variable; $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})^T$, Z_{i1} followed a Bernoulli (0.5) distribution, Z_{i2} followed a Bernoulli distribution with probability depend on Z_{i1}, and Z_{3i} came from a Normal distribution with constant variance 25 and mean depend on Z_{i1} and Z_{i2}. We set $\beta^T = (\beta_1, \beta_2, \beta_3) = (0.1, -0.1, 0.01)$. We varied the sample size for each center as $n_j = 50, 100$. The censoring time was generated from Uniform(1,10) distribution with percentages were approximately 15%. Each data configuration was replicated 1000 times.

In Table 1, we considered a simulation setting where the hazard functions and covariate distribution were equal across centers. In Table 2, the hazards functions are center-dependent and the covariate distribution are center-independent. In Table 3, both hazards and covariates are center-dependent. We list the results for performance of proposed model and methods. In all simulation settings, the coefficient estimators, cumulative hazard function and proposed estimator for center effects were sufficiently well-behaved, in the sense that the bias was quite small, and the average asymptotic standard errors (ASE) were generally close to the empirical standard deviations (ESD), while the empirical coverage probabilities (CP) were generally consistent with the nominal value 0.95.

4.5　Application to Kidney Transplant Data

We applied the proposed methods to evaluate U.S. kidney transplant centers with respect to mean post-transplant survival time. Data were 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

Table 4.1: Simulation Setting 1: hazards and covariate are center independent; $t=3$; Λ_j=cumulative hazard functions; $\psi_j(t)$=ratio of survival function; only centers with odd number are shown.

λ_j	Parameters	TRUE	BIAS	ESD	ASE	CP
	β_1	0.010	0.000	0.057	0.055	0.95
	β_2	-0.100	0.001	0.056	0.054	0.94
	β_3	0.100	0.000	0.064	0.062	0.95
0.8	$\Lambda_1(t)$	2.400	0.010	0.603	0.607	0.94
	$\psi_1(t)$	1	-0.000	0.307	0.307	0.92
0.8	$\Lambda_3(t)$	2.400	0.025	0.627	0.609	0.94
	$\psi_3(t)$	1.000	-0.013	0.308	0.305	0.93
0.8	$\Lambda_5(t)$	2.400	-0.002	0.605	0.606	0.94
	$\psi_5(t)$	1.000	0.011	0.303	0.308	0.92
0.8	$\Lambda_7(t)$	2.400	-0.001	0.616	0.697	0.93
	$\psi_7(t)$	1.000	0.008	0.300	0.308	0.95
0.8	$\Lambda_9(t)$	2.400	0.025	0.608	0.608	0.94
	$\psi_9(t)$	1.000	-0.014	0.305	0.304	0.93

Table 4.2: Simulation Setting 2: hazards are center dependent; covariate are center independent; $t{=}3$; Λ_j=cumulative hazard functions; $\psi_j(t)$=ratio of survival function; only centers with odd number are shown.

λ_j	Parameters	TRUE	BIAS	ESD	ASE	CP
	β_1	0.010	-0.001	0.049	0.048	0.94
	β_2	-0.100	0.001	0.048	0.048	0.94
	β_3	0.100	-0.003	0.056	0.055	0.94
0.5	$\Lambda_1(t)$	1.500	0.009	0.476	0.487	0.94
	$\psi_1(t)$	1.793	-0.019	0.312	0.324	0.95
0.6	$\Lambda_3(t)$	1.800	0.024	0.503	0.507	0.95
	$\psi_3(t)$	1.328	-0.020	0.294	0.291	0.94
0.7	$\Lambda_5(t)$	2.100	0.027	0.527	0.531	0.94
	$\psi_5(t)$	0.984	-0.007	0.254	0.258	0.93
0.8	$\Lambda_7(t)$	2.400	0.015	0.569	0.562	0.94
	$\psi_7(t)$	0.729	0.015	0.225	0.229	0.93
0.9	$\Lambda_9(t)$	2.700	0.046	0.601	0.606	0.95
	$\psi_9(t)$	0.540	0.006	0.195	0.199	0.93

Table 4.3: Simulation Setting 3: hazards and covariate are center dependent; $t=3$; Λ_j=cumulative hazard functions; $\psi_j(t)$=ratio of survival function; only centers with odd number are shown.

| λ_j	Parameters	TRUE	BIAS	ESD	ASE	CP
	β_1	0.010	0.000	0.048	0.049	0.96
	β_2	-0.100	-0.003	0.049	0.049	0.95
	β_3	0.100	0.002	0.053	0.052	0.96
0.5	$\Lambda_1(t)$	1.500	0.005	0.487	0.488	0.93
	$\psi_1(t)$	1.793	-0.022	0.316	0.326	0.95
0.6	$\Lambda_3(t)$	1.800	0.004	0.502	0.512	0.94
	$\psi_3(t)$	1.328	-0.004	0.287	0.293	0.94
0.7	$\Lambda_5(t)$	2.100	0.026	0.549	0.545	0.93
	$\psi_5(t)$	0.984	-0.011	0.256	0.260	0.94
0.8	$\Lambda_7(t)$	2.400	0.034	0.579	0.583	0.95
	$\psi_7(t)$	0.729	-0.003	0.235	0.230	0.93
0.9	$\Lambda_9(t)$	2.700	-0.009	0.604	0.623	0.94
	$\psi_9(t)$	0.540	0.032	0.204	0.207	0.95

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. Included in the analysis were adult patients ($\geq$ 18 years of age at transplant) who underwent kidney transplantation between January 2000 and December 2008. Adjustment covariates included age, race, gender, diagnosis, donation after cardiac death (DCD), Expanded Criteria Donor (ECD), BMI, dialysis time, indicator of previous kidney transplant, cold ischemic time, Patients with donor type other than deceased and centers with sample size less than 50 were eliminated from additional analysis. The final sample size was $n = 72,302$ from $J = 185$ centers across the United States. The median survival time over the study period was 3 years across centers.

We fit the stratified additive hazard model and estimate the baseline survival function for each center at year 3. We ordered centers based on the ratio of baseline survival function and we provided the plots of confidence intervals in Figure 1. Centers with lower confidence limit higher than 1 were identified as significantly better than the national average, while centers with upper confidence bound lower than 1 were identified as significantly worse than the national average. A total of 6 centers had ratio of survival function significantly worse than the expected, while 3 centers were significantly above the average.

4.6 Discussion

In this study, we develop methods for evaluating center-specific survival using a center-stratified additive hazards model. We propose to estimate each center effect by an ratio of baseline survival functions. The proposed measure is a semiparametric generalization of the relative risk, which is often used in clinical studies. Given the additive hazard structure, the ratio of survival function for a particular subject reduces to the ratio of baseline survival function. In this light, the proposed measure has an interpretation at the individual level, which is perhaps more relevant to a

Figure 4.1: Analysis of SRTR Data: Point Estimates $\widehat{\psi}_j$ ($j = 1, \cdots, 185$) at year 3 and 95% Confidence Intervals .
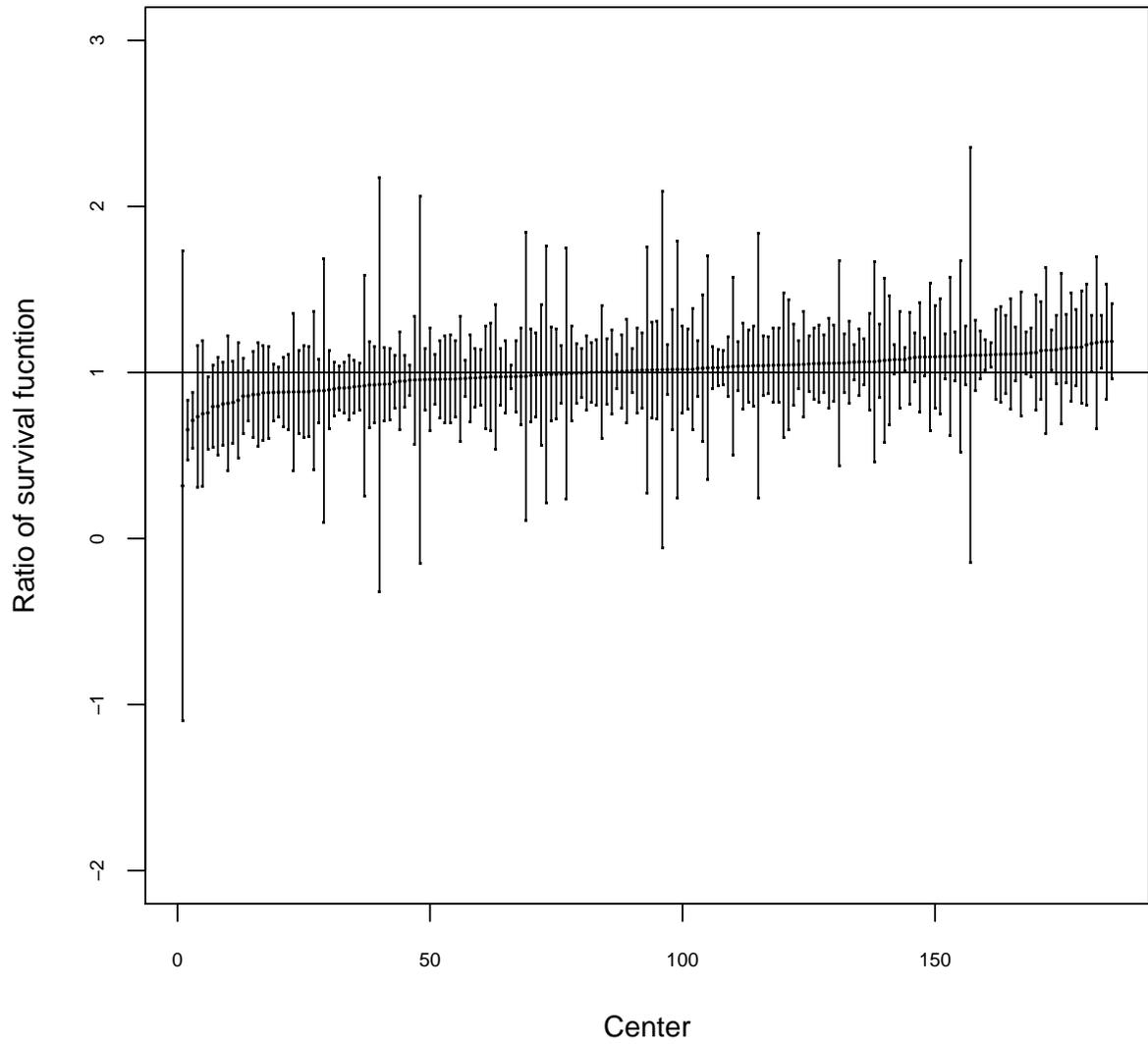
patient than estimators based on standardized or averaging. Under an additive hazard model, the ratio of baseline survival functions is invariant to the choice of baseline covariate level. That means, the ratio of survival functions represents the contrast between subject i at center j versus subject i at the hypothetical center with baseline hazard function equal to the national average; where subject i can have any covariate value. The proposed measure is a semiparametric generalization of the relative risk, which is often used in clinical studies. One advantage for our proposed method is that the ratio of survival functions is invariant to the choice of baseline covariate level. Therefore, the ratio of survival functions represents the contrast between subject i at center j versus subject i at the hypothetical center with baseline hazard function equal to the national average; where subject i can have any covariate value. To implement the proposed estimator, we propose to use a stratified additive model, which is a extension of the center-stratified Cox model. The major difference between this model and the commonly used survival models is that it allows the regression effect to be additive, in the meanwhile it also allow the baseline hazards be center-specific.

In this study, we focused on the additive hazard model. A more general hazard model can be defined as

$$(4.11) \qquad \lambda_{ij}(t) = g\{\lambda_0(t), \theta_j(t)\}h_1(\beta^T Z_i(t)) + h_2(\beta^T Z_i(t))$$

where g is an unknown function, and h is a known function (e.g., an exponential function, $exp(\beta^T Z)$; or an additive function, $\beta^T Z$).

When $h_1 = 1$ and h_2 is a linear link function, the hazard function reduces to the additive function, the stratified additive model and the ratio of baseline survival function proposed in this study can be used to evaluate center effect.

When h_1 is the exponential function and $h_2 = 0$, the hazard function reduces to the multiplicative hazard model

$$(4.12) \qquad \lambda_{ij}(t) = \lambda_{0j}(t)exp(\beta^T Z).$$

For such a model, the cumulative hazard ratio proposed by Wei and Schaubel (2008) can be used to evaluate the center effects. For instance, Wei and Schaubel (2008)

proposed measures to compare two treatment groups,for which they assume that there exist a natural (default) choice for the reference (e.g. placebo group).

$$\frac{\Lambda_{0j}(t)}{\Lambda_{0k}(t)}.$$

In our motivating example with multi-centers, it is not clear how to choose the reference center. Different choice of the reference center may had a strong impact on the results, which would of course lead to controversy. This concern can be solved by our proposed method with the reference defined as national average. A generalization of Wei and Schaubel (2008)'s method can then be applied to multi-groups (i.e. multi-centers) as following

(4.13)
$$\frac{\Lambda_{0j}(t)}{\Lambda_0(t)}$$

where $\Lambda_0(t) = \sum_{\ell=1}^{J} p_\ell \Lambda_0(t)$, with $p_\ell = \frac{1}{n} \sum_{\ell=1}^{J} n_\ell$.

CHAPTER V

Conclusion

In observational studies, patient characteristics may vary considerably across centers, and this can have substantial impact on the expected outcomes at a given center. An accurate evaluation of center effects needs to adequately account for the imbalance of patient characteristics. In this dissertation, we propose three novel statistical methods to compare the performances of kidney transplant centers in the presence of censored data.

The standardized mortality ratio (SMR) based on a Cox regression model is often used to evaluate center-specific mortality. However, the asymptotic properties and finite-sample behavior of the Cox SMR are not well-studied. Chapter II evaluates some strong limitations of the Cox SMR that relate to its underlying assumptions. To address these limitations, modifications to the Cox SMR has been developed based on a stratified Cox model. In addition, since center effects computed through indirect standardization are not comparable, semiparametric generalization of direct standardization are proposed.

In the context of survival data, the difference in mean lifetime is arguably a more meaningful measure than the ratio of death rates. In Chapter III, we propose a method which combines a log-normal frailty model and piece-wise exponential baseline rates to compare the mean survival time across centers. The proposed methods allow for valid estimation of mean survival time as opposed to the restricted mean lifetime and, within this context, robust profiling of long-term center-specific outcomes. Maximum likelihood based estimation is carried out using a Laplace approximation

for integration. The proposed methods work well with a large database and can accommodate a large number of centers.

Chapter IV develops methods for evaluating center-specific survival using a center-stratified additive hazards model. The relative center effects are estimated using the ratio of baseline survival functions. The proposed measure is a semiparametric generalization of the relative risk, which is often used in clinical studies. One attractive feature for the proposed method is that the ratio of survival function for a particular subject reduces to the ratio of baseline survival function, and such ratio of baseline survival functions is invariant to the choice of baseline covariate level.

The direct standardization methods in Chapter II could be extended in several useful directions. Perhaps most notably, it is often of interest to evaluate center effects in settings with recurrent events. Further generalizations of these problems include issues of competing risks and the incorporation of additive hazard models. For instance, a directly standardized cumulative incidence function could be used to estimate center effects in the competing risk setting. Furthermore, the direct standardization method is a version of the G-computation method for average causal effects, for which we need to correctly specify a model relating survival and covariates. If such a model was incorrectly specified, then the resulting center effects would be biased. Another approach that has been advocated for estimating the average causal effect is through the use of inverse probability weighted estimators. In this approach, both the censoring distribution and the probability of receiving one of the treatment (propensity score) are modeled as functions of the covariates. If these models were incorrectly specified, then the inverse probability weighted estimators would be biased. To address these concerns, a double robust estimator based on augmented inverse probability weighting may provide better solutions. The double robust estimator would be consistent if either the model for survival or the models for censoring and treatment were correctly specified.

Bibliography

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Mathematical Statistics and Probability Theory. Lecture Notes in Statist. Springer, New York.* **2**, 1–25.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907–925.

Andersen, P. K., Klein, J. P., and Zhang, M.-J. (1999). Testing for center effects in multi-centre survival stuides: a monte carlo comparison of fixed and random effects. *Statistics in Medicine* **18**, 1489–1500.

Andersen, P. K. and Væth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45**, 523–535.

Andersen, P. K., et al. (1985). A cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics* **41**, 921–932.

Andersen, P. K., et al. (1989). Estimation of variance in cox's regression model with shared gamma frailties. *Biometrics* **53**, 1475–1484.

Andersen, P. K., et al. (1993). *Statistical Models Based on Counting Processes.* Springer Series in Statistics.

Berry, G. (1983). The analysis of mortality by the subject-years method. *Biometrics* **39**, 173–184.

Bilias, Y., Gu, M., and Ying, Z. (1997). Towards a general asymptotic theory for cox model with staggered entry. *Annals of Statistics* **25**, 662–682.

Breslow, N. E. (1972). Contribution to the discussion on the paper bt d. r. cox, regression and life table. *Journal of the Royal Statistical Society, Series B* **34**, 216–217.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, 1, The design and Analysis of Case-Contril Studies.* Lyon: IARC.

Breslow, N. E. and Day, N. E. (1985). The standardized mortality ratio. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences. The Bernard G. Greenberg Volume Edited by P.K. Sen.* pages 55–74.

Breslow, N. E. and Langholz, B. (1987). Nonparametric estimation of relative mortality functions. *Journal of Chronic Diseases* **40 (suppl.2)**, 89S–99S.

Breslow, N. E., et al. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association* **78**, 1–12.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* **57**, 1030–1038.

Clayton, D. and Cuzick, J. (1985). Multivariate generalization of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, series A* **148**, 82–117.

Clayton, D. G. (1978). A model for association in bivariate life table and its application in epidemiological studies of familiar tendency in chronic disease incidence. *Biometrics* **65**, 141–151.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–200.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* New York: Chapman.

Demarqui, F., Loschi, R. H., and Colosimo, E. A. (2008). Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis* **14**, 333–356.

Feng, s., Wolfe, R. A., and F.K., P. (2005). Fraility survival model analysis of the national deceased donor kidney transplant dataset using poisson variance structures. *Journal of the American Statistical Association* **100**, 728–735.

Hazel, I. (2005). Standardization methods. *Encyclopedia of Biostatistics* **7**, 5151–5163.

Huang, X., Chen, S., and Soong, S.-j. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics* **54**, 1420–1433.

Huffer, F. W. and Mckeague, I. W. (1991). Weighted least squares estimation for aalen's additive risk model. *Journal of the American Statistical Association* **95**, 238–248.

Karrison, T. (1987). Restricted mean lifetime with adjustment for covariates. *Journal of the American Statistical Association* **82**, 1169–1176.

Keiding, N. (1987). The method of expected number of deaths, 1786-1886-1986. *International Statistical Review* **55**, 1–20.

Lacson, E., et al. (2001). Limitations of the facility-specific standardized mortality ratio for profiling health care quality in dialysis. *American Journal of Kidney Diseases* **37**, 267–275.

Lin, D. Y., Fleming, R., T, and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards models. *Biometrika* **81**, 73–81.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Annals of Statistics* **23**, 1712–1734.

Lin, D. Y., et al. (2000). Semiparametic regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62**, 711–730.

McGilchrist, C. A. (1993). Reml estimation for survival models with frality. *Biometrics* **49**, 221–225.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frality in survival analysis. *Biometrics* **47**, 461–466.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika* **81**, 501–514.

O'Neill, T. J. (1986). Inconsistency of the misspecified proportional hazards model. *Statistics & Probability Letters* **4**, 219–222.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics: Hayward, CA.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics* **11**, 453–466.

Simpson, C. H. (1951). The interpretation of interaction in contingence tables. *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

Vonesh, E. F., Chinchilli, V. M., and Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics* **52**, 572–587.

Vonesh, E. F., et al. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association* **97**, 271–283.

Wei, G. and Schaubel, D. E. (2008). Estimating cumulative treatment effects in the presence of nonproportional hazards. *Biometrics* **64**, 724–732.

Wolfe, R. A., et al. (1992). Using usrds generated mortality tables to compare local esrd mortality rates to national rates. *Kidney International* **42**, 991–996.

Wolfe, R. A., et al. (1999). Patient survival for waitlisted dialysis versus cadaveric renal transplant patients in the united states. *New England Journal of Medicine* **341**, 17251730.

Yamaguchi, T. and Ohashi, Y. (1999). Investigating centre effects in a muiti-centre clinical trial of superficial bladder cancer. *Statistics in Medicin* **18**, 1961–1971.

Yin, G. S. and Jianwen Cai, J. W. (2004). Additive hazards model with multivariate failure time data. *Biometrika* **91**, 801–818.

Zeng, D. L. and Cai, J. W. (2010). Additive transformation models for clustere failure time data. *Lifetime Data Analysis* **16**, 333–352.

Zhang, M. and Schaubel, D. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics* **57**, 740–749.

Zucker, D. M. (1998). Restricted mean life with covariates: Modeifcation and extension of a useful survival analysis method. *Journal of the American Statistical Association* **93**, 702–709.