Biometrics 00, 1–25 DOI: 10.1111/j.1541-0420.2005.00454.x
January 2012

Estimating the Average Treatment Effect on Survival Based on Observational Data and Using Partly Conditional Modeling

Qi Gong[1] and Douglas E. Schaubel[2]

[1]Gilead Science Inc., 333 Lakeside Dr, Foster City, CA, 94404, U.S.A.

[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109, U.S.A.

*email: gongqi@gmail.com

**email: deschau@umich.edu

Summary: Treatments are frequently evaluated in terms of their effect on patient survival. In settings where randomization of treatment is not feasible, observational data are employed, necessitating correction for covariate imbalances. Treatments are usually compared using a hazard ratio. Most existing methods which quantify the treatment effect through the survival function are applicable to treatments assigned at time 0. In the data structure of our interest, subjects typically begin follow-up untreated; time-until-treatment and the pre-treatment death hazard are both heavily influenced by longitudinal covariates; and subjects may experience periods of treatment ineligibility. We propose semiparametric methods for estimating the average difference in restricted mean survival time attributable to a time-dependent treatment, the average effect of treatment among the treated, under current treatment assignment patterns. The pre- and post-treatment models are partly conditional, in that they use the covariate history up to the time of treatment. The pre-treatment model is estimated through recently developed landmark analysis methods. For each treated patient, fitted pre- and post-treatment survival curves are projected out, then averaged in a manner which accounts for the censoring of treatment times. Asymptotic properties are derived and evaluated through simulation. The proposed methods are applied to liver transplant data in order to estimate the effect of liver transplantation on survival among transplant recipients under current practice patterns.

Key words: Landmark analysis; Observational data; Partly conditional model; Proportional hazards regression; Time-varying covariates; Treatment effect.

1. Introduction

It is often of interest in biomedical settings to evaluate the benefit of a treatment on survival. In many clinical settings, treatment is not administered right at the time of diagnosis, such that a period of waiting time occurs for some (or perhaps all) patients. In cases where treatment is not randomized, it is often useful to assess the benefit of treatment under current treatment assignment patterns. Through the average effect-of-treatment-on-the-treated (ETT; Pearl, 2009), one can evaluate the benefit of treatment as currently practiced.

Survival probabilities are easily understood by health care professionals, as is the area under the survival curve (restricted mean lifetime). Various authors have proposed using Cox regression with the primary goal not being to estimate hazard ratios, but to compare differences in survival and/or restricted mean lifetime. For example, Zucker (1998) and Chen and Tsiatis (2001) proposed methods that involved averaging over fitted values from Cox models. Zhang and Schaubel (2011) extended the methods of Chen and Tsiatis (2001) to accommodate dependent censoring, then subsequently developed double-robust methods (Zhang and Schaubel, 2012). Each of the afore-described methods applies to treatments assigned at baseline, as opposed to time-varying treatments.

In the data structure of interest in this report, all patients begin follow-up untreated, with some eventually receiving treatment and others dying beforehand. Pre-treatment mortality and treatment assignment rates are dependent on longitudinal covariates (including periods during which a subject is declared treatment-ineligible), such that a patient's pre-treatment death is dependently censored by the receipt of treatment. Post-treatment survival is dependent on a subject's condition at the time of treatment, and the duration of pre-treatment follow-up time. Our objective is to contrast two scenarios: (a) treatment is never assigned (b) treatment is assigned according to current practice patterns.

The proposed methods are motivated by the end-stage liver disease (ESLD) setting. The number of available deceased-donor livers is always less than the number of patients in need of liver transplantation. As a result, medically suitable patients are placed on a liver transplant waiting list. Patients typically begin follow-up on the wait list ('untreated'; i.e., not transplanted), such that transplantation can be viewed as a time-dependent treatment. In the U.S., chronic end-stage liver disease patients are sequenced in decreasing order of Model for End-Stage Liver Disease (MELD) score, a very strong predictor of pre-transplant mortality. Transplantation results in the dependent censoring of pre-transplant death, since MELD scores predict both wait list mortality and transplant rates. Note that patients may be removed from the wait list (or made inactive) and, in such cases, are permanently (or temporarily) ineligible to receive a transplant. In the setting of our interest, the effect of treatment on the treated is of greater interest than the average causal effect, due to the implausibility of all patients receiving treatment.

Our analysis in Section 5 is different from that in Gong and Schaubel (2013) since (i) the former only looked at pre-transplant survival; (ii) did not compare post- versus pre-transplant survival; (iii) reported contrasts only in terms of the hazard ratio; and (iv) did not exclude Status 1 (acute liver failure) patients and, in fact, focused on contrasting them with chronic ESLD patients.

We develop semiparametric methods to estimate the average effect-of-treatment-on-the-treated through partly conditional modeling. The proposed method involves averaging over the observed instances of treatment initiation, with the averaging accounting for the various complexities in data structure; e.g., treatment initiation times are subject to right censoring; patients may die before treatment is received; and patients cannot initiate treatment while ineligible. For each treated patient, we use the accrued history (up to the time of treatment initiation) to project out a survival curve for post-treatment residual lifetime. Based on

the same accrued pre-treatment history, we also project out the survival curve that would apply in the absence of treatment. This set-up lends itself well to partly conditional modeling (Zheng and Heagerty, 2005; Gong and Schaubel, 2013); see also the closely related concept of landmark analysis (Feuer et al., 1992; van Houwelingen, 2007; van Houwelingen and Putter, 2012; Parast, Tian and Cai, 2014). Gong and Schaubel (2013) developed methods for fitting partly conditional hazard regression models which apply to the absence-of-treatment setting in our set-up. We extend the ideas in Gong and Schaubel (2013) to estimate the average ETT through residual survival and restricted mean survival time. Although we focus on partly conditional modeling in this report, it should be noted that other pertinent methods exist, as described in Section 6.

The remainder of this article is organized as follows. In Section 2, we describe the proposed methods. Asymptotic properties are provided in Section 3 (for proofs, see Supplementary Materials), with finite-sample properties evaluated through simulation in Section 4. We apply the proposed methods to the motivating data set in Section 5. Concluding remarks are made in Section 6.

2. Proposed Methods

2.1 *Set-up and Notation*

We now formalize the ideas introduced in Section 1, in the absence of censoring. We remove subscripting, such that defined variates pertain to any hypothetical subject. We let T represent treatment time, with $T > 0$ since subjects begin follow-up untreated. Death time in the absence of treatment is denoted by D^0. Note that, consistent with the intent-to-treat principle, patients that initiate treatment are considered to be 'treated' thereafter. Let $\mathcal{E}(s) = 1$ if the patient is treatment-eligible (i.e., eligible to initiate treatment) at time s, and 0 otherwise. A patient may oscillate between the eligible and ineligible states before

time $D^0 \wedge T$, where $a \wedge b = \min(a, b)$. In particular, $\mathcal{E}(s) = 0$ for $s > D^0 \wedge T$, since a patient cannot *initiate* treatment more than once, and cannot initiate treatment after death. Note that a patient may only initiate treatment while eligible; i.e., $dI(T \leqslant s) = \mathcal{E}(s)dI(T \leqslant s)$.

Under the above-listed Scenario (a), $T = \infty$. Under Scenario (b), treatment only occurs when $T < D^0$, in which case D^0 is considered latent; D^0 serves as a competing risk for T. For a patient with treatment time $T = s$, D^1 is the death time, such that $(D^1 - s)_+$ is the residual post-treatment survival, with $a_+ = a \cdot I(a > 0)$ and $I(\cdot)$ being the familiar 0/1 indicator function. The quantity $(D^0 - s)_+$ then represents the residual survival beyond s that would have been observed in the absence of treatment. Note that if $D^0 < T$, then D^1 is undefined.

The covariate vector, which contains some time-varying elements, is denoted by $\boldsymbol{Z}^*(s)$. The patient's covariate and eligibility history up to time s is given by $\boldsymbol{\mathcal{H}}(s) = \{\boldsymbol{Z}^*(u), \mathcal{E}(u);\ 0 \leqslant u < s\}$. The above described set-up is illustrated in Figure 1. For a patient with treatment-initiation time $T = s$, we are interested in the average difference between $(D^1 - s)_+$ and $(D^0 - s)_+$ given $[\boldsymbol{\mathcal{H}}(s), T = s]$, with the average being taken with respect to the subdistribution function for T.

[Figure 1 about here.]

2.2 *Treatment Effect: Conditional and Average*

For a patient initiating treatment at time $T = s$, there are two death times of interest; the post-treatment residual death time, $(D^1 - s)_+$, and the residual death time that would have occurred in the absence of treatment, $(D^0 - s)_+$. At the time of treatment (e.g., $T = s$), we observe $\boldsymbol{\mathcal{H}}(s)$, and $\mathcal{E}(s) = 1$. Conditional on $[\boldsymbol{\mathcal{H}}(s), T = s]$, we contrast

$$S_1(t; s|\boldsymbol{\mathcal{H}}(s), T = s) = P\{(D^1 - s) > t|\boldsymbol{\mathcal{H}}(s), T = s) \quad (1)$$

$$S_0(t; s|\boldsymbol{\mathcal{H}}(s), T = s) = P\{(D^0 - s) > t|\boldsymbol{\mathcal{H}}(s), T = s) \quad (2)$$

the survival functions pertaining to the counterfactual variates $(D^1 - s)_+$ and $(D^0 - s)_+$, respectively. Note that, in both $S_1(t; s|\cdot)$ and $S_0(t; s|\cdot)$, the time index s represents conditioning time, while t refers to residual survival t time units beyond the conditioning time, s. That is, $S_j(t; s|\cdot)$ pertains to a *gap* of t units beyond time s, which equals *total* time $(s + t)$. We assume strong ignorability (Rubin, 1978), permitting inference on the counterfactuals $(D^1 - s)_+$ and $(D^0 - s)_+$, through observed data. The strong ignorability assumption is detailed in the Supplementary Materials. An implication this assumption is that $S_0(t; s|\boldsymbol{\mathcal{H}}(s), T = s) = S_0(t; s|\boldsymbol{\mathcal{H}}(s), \mathcal{E}(s) = 1)$, consistent with the counterfactuals $(D^1 - s)_+$ and $(D^0 - s)_+$ being independent of the receipt of treatment at time s.

For fixed $L > 0$, restricted mean residual survival times are given by

$$\mu_1(L; s|\boldsymbol{\mathcal{H}}(s), T = s) = \int_0^L S_1(t; s|\boldsymbol{\mathcal{H}}(s), T = s)dt \tag{3}$$

$$\mu_0(L; s|\boldsymbol{\mathcal{H}}(s), T = s) = \int_0^L S_0(t; s|\boldsymbol{\mathcal{H}}(s), T = s)dt. \tag{4}$$

Conditioning on $[\boldsymbol{\mathcal{H}}(s), T = s]$, a pertinent contrast in survival functions is then

$$\delta(t; s|\boldsymbol{\mathcal{H}}(s), T = s) = S_1(t; s|\boldsymbol{\mathcal{H}}(s), T = s) - S_0(t; s|\boldsymbol{\mathcal{H}}(s), T = s), \tag{5}$$

while a contrast in restricted mean residual lifetime is defined as

$$\Delta(L; s|\boldsymbol{\mathcal{H}}(s), T = s) = \mu_1(L; s|\boldsymbol{\mathcal{H}}(s), T = s) - \mu_0(L; s|\boldsymbol{\mathcal{H}}(s), T = s). \tag{6}$$

Average survival functions are then defined as

$$\begin{aligned} S_1(t) &= E[S_1(t; T|\boldsymbol{\mathcal{H}}(T), T)] \\ S_0(t) &= E[S_0(t; T|\boldsymbol{\mathcal{H}}(T), T)], \end{aligned} \tag{7}$$

where, in each case, the expectation is taken with respect to the joint distribution of $[\boldsymbol{\mathcal{H}}(T), T)]$ over the identifiable range of T which would in practice be capped by the

maximum follow-up time. Correspondingly, average restricted mean residual lifetimes are:

$$\begin{aligned} \mu_1(L) &= E[\mu_1(L;T|\boldsymbol{\mathcal{H}}(T),T)] = \int_0^L S_1(t)dt \\ \mu_0(L) &= E[\mu_0(L;T|\boldsymbol{\mathcal{H}}(T),T)] = \int_0^L S_0(t)dt. \end{aligned} \tag{8}$$

The ETT can then be defined in terms of mean difference in survival probability as

$$\delta(t) = E[\delta(t;T|\boldsymbol{\mathcal{H}}(T),T)] = S_1(t) - S_0(t) \tag{9}$$

and in terms of average difference in residual mean survival time, by

$$\Delta(L) = E[\Delta(L|\boldsymbol{\mathcal{H}}(T),T)] = \mu_1(L) - \mu_0(L) = \int_0^L \delta(t)dt. \tag{10}$$

Having specified the treatment effect of interest, the remaining subsections in Section 2 describe the proposed methods for estimating $\delta(t)$ and $\Delta(L)$.

2.3 *Observed data: Notation and set-up*

We let D_i denote the death time for subject i ($i = 1, \ldots, n$). The time of treatment is given by T_i, with $T_i = \infty$ when $D_i < T_i$. Treatment and death times are subject to independent right censoring, C_i, intended to represent the combination of administrative censoring and random loss to follow-up. Observation time is then given by $X_i = D_i \wedge C_i$. We define counting processes for death, treatment and censoring, as $N_i(t) = I(D_i \leqslant t \wedge C_i)$, $N_i^T(t) = I(T_i \leqslant t \wedge D_i \wedge C_i)$ and $N_i^C(t) = I(C_i \leqslant t \wedge D_i)$, respectively. Recall that $\mathcal{E}_i(u)$ equals 1 if patient i is eligible for treatment at time u, and 0 otherwise. Note that $N_i^T(t) = \int_0^t \mathcal{E}_i(u)dN_i^T(u)$, since treatment can only be initiated for an eligible subject. The covariate vector, observed longitudinally, is denoted by $\boldsymbol{Z}_i^*(t)$. The covariate and treatment-eligibility history for subject i as of time t is denoted by $\boldsymbol{\mathcal{H}}_i(t) = \{\boldsymbol{Z}_i^*(u), \mathcal{E}_i(u);\ u \in [0,t)\}$. Covariate information is assumed to not be available after treatment is assigned, such that the total observed history for subject i is given by $\boldsymbol{\mathcal{H}}_i(X_i \wedge T_i)$; such data are not required by the proposed methods.

2.4 *Assumed Models and Estimation Methods*

We now describe the assumed models for $(D_i^1 - T_i)_+$, $(D_i^0 - T_i)_+$, T_i and C_i. As implied by (7) and (8), our target ETT implies averaging over the observed $[T_i, \boldsymbol{\mathcal{H}}_i(T_i)]$ distribution. Per (1) and (2), we achieve this by working with $[(D_i^1 - s)_+|\boldsymbol{\mathcal{H}}_i(s), T_i = s]$ and $[(D_i^0 - s)_+|\boldsymbol{\mathcal{H}}_i(s), T_i = s]$ directly, after which we will then average explicitly. We model the partly conditional hazard function for $[(D_i^1 - s)_+|\boldsymbol{\mathcal{H}}_i(s), T_i = s]$, which uses in the covariate vector all pertinent information in the history prior to the time of treatment, $\boldsymbol{\mathcal{H}}_i(T_i)$. The model is partly conditional since the covariate is not updated after the time treatment is initiated. The covariate is not updated after time T_i since we want to project residual survival from T_i onward, and a survival projections based on traditional time-dependent model would require a model for $\boldsymbol{\mathcal{H}}_i(s+t)$. In many cases, a model for $\boldsymbol{\mathcal{H}}_i(s+t)$ is complicated to fit accurately, and is of little inherent interest to the investigators.

2.4.1 *Post-Treatment Survival.* We let $\lambda_1(t; s|\boldsymbol{\mathcal{H}}(s), T = s)$ denote the hazard function corresponding to $S_1(t; s|\boldsymbol{\mathcal{H}}(s), T = s)$ from (1). We assume the following proportional hazards model,

$$\lambda_1(t; s|\boldsymbol{\mathcal{H}}_i(s), T_i = s) \;=\; \lambda_{01}(t) \exp\{\boldsymbol{\beta}_1' \boldsymbol{Z}_{i1}(s)\}, \tag{11}$$

where the covariate $\boldsymbol{Z}_{i1}(s)$ is chosen to summarize the pre-treatment history, $\{\boldsymbol{\mathcal{H}}_i(s), T_i = s\}$, pertinent to predicting post-treatment survival. Typically, time until treatment, T_i, would be represented parametrically in the covariate vector, $\boldsymbol{Z}_{i1}(s)$. Note that the $\boldsymbol{Z}_{i1}(s)$ covariate is fixed at treatment time $T_i = s$, reflecting the partly conditional (Zheng and Heagerty, 2005; Gong and Schaubel, 2013) nature of (11), which uses time-dependent data 'frozen' at time of treatment. This could also be considered a 'landmark' analysis (e.g., van Houwelingen, 2007), with landmark times being customized to each subject and set to T_i.

We assume that treatment times are independently censored by C_i. Assuming that $(D_i - T_i)_+$ is independently censored by $(C_i - T_i)_+$ given $[\boldsymbol{Z}_{i1}(T_i), T_i]$, parameter estimation for

model (11) proceeds through unweighted partial likelihood. We denote the resulting estimators for model (11) by $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\Lambda}_{01}(t)$, with the latter being the Breslow-Aalen (1972) estimator. We estimate $S_1(t; s|\boldsymbol{\mathcal{H}}_i(s), T_i = s)$ by $\widehat{S}_1(t; s|\boldsymbol{Z}_{i1}(s)) = \exp\{-\widehat{\Lambda}_1(t; s|\boldsymbol{Z}_{i1}(s))\}$, where $\widehat{\Lambda}_1(t; s|\boldsymbol{Z}_{i1}(s)) = \widehat{\Lambda}_{01}(t)\exp\{\widehat{\boldsymbol{\beta}}_1'\boldsymbol{Z}_{i1}(s))\}$, and $\mu_1(L; s|\boldsymbol{\mathcal{H}}_i(s), T_i = s)$ by $\widehat{\mu}_1(L; s|\boldsymbol{Z}_{i1}(s)) = \int_0^L \widehat{S}_1(t; s|\boldsymbol{Z}_{i1}(s))dt$.

2.4.2 *Survival in the Absence of Treatment.* We begin by describing the assumed hazard model for survival in the absence of treatment. We then outline the proposed data augmentation, which involves selecting calendar date cross-sections. Next, we detail fitting the model through an inverse weighted and stratified log rank estimating function.

We let $\lambda_0(t; s|\boldsymbol{\mathcal{H}}(s), T = s)$ denote the hazard function corresponding to (2). Under strong ignorability, note that $\lambda_0(t; s|\boldsymbol{\mathcal{H}}_i(s), T_i = s) = \lambda_0(t; s|\boldsymbol{\mathcal{H}}_i(s), \mathcal{E}_i(s) = 1)$, which we use in listing the assumed model,

$$\lambda_0(t; s|\boldsymbol{\mathcal{H}}_i(s), \mathcal{E}_i(s) = 1) \quad = \quad \lambda_{00}(t)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_{i0}(s)\}, \tag{12}$$

where $\boldsymbol{Z}_{i0}(s)$ is chosen such that $\lambda_0(t; s|\boldsymbol{\mathcal{H}}_i(s), \mathcal{E}_i(s) = 1) = \lambda_0(t; s|\boldsymbol{Z}_{i0}(s))$, implying that $\boldsymbol{Z}_{i0}(s)$ contains all elements of the history pertinent to predicting $(D_i^0 - s)_+$, including all appropriate functions of time-already-survived, s. Model (12) is partly conditional (Zheng and Heagerty, 2005; Gong and Schaubel, 2013) since, although the hazard at time $s + t$ is of interest, the model conditions on information which is 'frozen' at time s. In contrast, a typical (fully) conditional or 'time-dependent' model would condition on $\boldsymbol{\mathcal{H}}_i(s + t)$.

Partly Conditional Model: The motivation for using a partly conditional model is described at the start of Section 2.4. Generally, fitting a partly conditional model requires some form of data augmentation in which the records corresponding to each subject's observed data are restructured in order to facilitate fitting the assumed model. After such augmentation, each input record has a prior time survived (e.g., s_i) and corresponding prior history $\boldsymbol{\mathcal{H}}_i(s_i)$, with residual survival in the absence of treatment, $(D_i - s_i)_+$ then being analyzed. In fitting

the post-treatment residual survival model (11), there is an obvious choice for each treated subject's conditioning time, namely $s_i := T_i$. In accordance with (2), we actually need to project residual survival (in the absence of treatment) beyond this same conditioning time. Although the appropriate conditioning time for projecting (1) is clear, the nature of the data augmentation for fitting model (12) requires consideration.

Calendar Time Cross-sections: In landmark analysis, typically survival from a specific follow-up time point (or set of specific time points) is desired, with survival probability projected out after the chosen landmark time(s). In our case, since treatment can occur at any time point (e.g., $T = s$), we need to be able to project conditional survival forward from any conditioning time s. This suggests a partly conditional model which includes terms representing previous time survived, s. Variation in previous time survived is then required, which means that sampling component of the data augmentation should be based on something other then s itself. We choose to sample based on calendar time, since each calendar time cross-section will contain wide variation in previous time survived. As we later describe, we stratify the model by cross-section for computational savings, which is important in large data sets like that we analyze in Section 5. For instance, Gong and Schaubel (2013) developed a partly conditional model which chooses the conditioning times to be the s_i values observed on a randomly selected calendar date. For example, consider a particular calendar date (e.g., 07/01/2004); input records for fitting the model would consist of s_i (subject i's prior follow-up time as of 07/01/2004), the corresponding $\boldsymbol{\mathcal{H}}_i(s_i)$, and $(X_i - s_i)_+$ among subjects who (as of 07/01/2004) were alive, uncensored, yet-untreated, but eligible to initiate treatment; i.e., $\{i : X_i > s_i, \mathcal{E}_i(s_i) = 1\}$.

Method of Gong and Schaubel (2013): The estimation of $\boldsymbol{\beta}_0$ from model (12) was developed by Gong and Schaubel (2013). The essential ideas are presented here for continuity, and

because the authors only derived the properties of $\widehat{\boldsymbol{\beta}}_0$, but not those of $\widehat{S}_0(t; s|\boldsymbol{Z}_{i0}(s))$, $\widehat{\mu}_0(L; s|\boldsymbol{Z}_{i0}(s))$, $\widehat{S}_0(t)$ or $\widehat{\mu}_0(L)$.

To begin, we choose a set of K calendar dates, $\{CS_1, \ldots, CS_K\}$. Each cross-section date CS_k is intended to represent a calendar date at which a set of treatment-eligible patients (could have been but) was not treated; we model residual survival in the absence of treatment from this date forward. For calendar date CS_k, we select the cross-section of treatment-eligible patients who were not treated (on or before that day). For patient i, follow-up time (previous time survived) as of calendar date CS_k is denoted by s_{ik}. Hence, a patient selected into cross-section CS_k must, as follow-up time s_{ik} be: alive ($D_i > s_{ik}$), uncensored ($C_i > s_{ik}$), untreated ($T_i > s_{ik}$) and treatment-eligible $\mathcal{E}_i(s_{ik}) = 1$. Three remarks are important at this juncture. First, treatment-eligibility is a cross-section inclusion criterion, but not a censoring criterion; e.g., having been included in cross-section k and, hence, with $\mathcal{E}_i(s_{ik}) = 1$, patient i is not censored upon subsequently being deemed treatment-ineligible. Second, the covariate will be frozen at s_{ik}, such that the survival projection for the residual time $(D_i^0 - s_{ik})_+$ is based on $\boldsymbol{\mathcal{H}}_i(s_{ik})$. Third, a patient included in cross-section k is censored if treated; this induces dependent censoring. Each of these remarks is formalized shortly.

We now establish additional notation pertinent to model (12). Since survival time from cross-section is modeled, we define the following times-since-cross-section: $D_{ik} = (D_i - s_{ik})_+$, $T_{ik} = (T_i - s_{ik})_+$ and $C_{ik} = (C_i - s_{ik})_+$ as the death, treatment and censoring time respectively corresponding to the ith patient and measured from the kth cross section date. Figure 2 provides an illustration of how the treatment-free observation time is transformed into time-since-cross-section times. A modified counting and at-risk processes are also defined as $N_{i0k}(t) = N_i(s_{ik} + t)I(T_i > s_{ik} + t)$ and $Y_{i0k}(t) = I(D_{ik} \wedge C_{ik} \geqslant t)$, respectively.

[Figure 2 about here.]

Following Gong and Schaubel (2013), we estimate $\boldsymbol{\beta}_0$ through the stratified model,

$$\lambda_{0k}(t; s|\boldsymbol{\mathcal{H}}_i(s_{ik}), \mathcal{E}_i(s_{ik}) = 1) \quad = \quad \lambda_{00k}(t) \exp\{\boldsymbol{\beta}_0' \boldsymbol{Z}_{i0}(s_{ik})\}, \tag{13}$$

where $\boldsymbol{\beta}_0$ is the same parameter in the unstratified model of interest, (12). Model (13) is quite flexible. Non-proportionality can be accommodated by replacing $\boldsymbol{\beta}_0$ with $\boldsymbol{\beta}_0(t)$, a parametric function on t. The parameter vector could also be allowed to be a parametric function of previous time survived; e.g., $\boldsymbol{\beta}_{0k}$, or $\boldsymbol{\beta}_0(s_{ik})$. Moreover, interactions between s_i and elements of $\boldsymbol{\mathcal{H}}_i(s_i)$ are also possible. Alternatively, van Houwelingen and Putter (2015) suggested a stopped Cox model to avoid non-proportionality, with artificial censoring at $t = L$. By breaking the stratification on k, one could also model the effect of calendar time.

Inverse weighting: Model (13) conditions on $\boldsymbol{\mathcal{H}}_i(s_{ik})$. However, we anticipate that $\boldsymbol{\mathcal{H}}_i(s_{ik}+t)$ would be predictive of both the treatment hazard and the pre-treatment death hazard at time $(s_{ik} + t)$. The mutual association, even conditional on $\boldsymbol{\mathcal{H}}_i(s_{ik})$, between pre-treatment death after s_{ik}, the probability of treatment after s_{ik} and $\boldsymbol{\mathcal{H}}_i(s_{ik} + t)$ sets up dependent censoring of $(D_i - s_{ik})_+$ by $(T_i - s_{ik})_+$. The potential bias due to such dependent censoring can be corrected through a variant of Inverse Probability of Censoring Weighting (IPCW; e.g., Robins and Rotnitzky, 1992) which requires a model for the treatment-initiation hazard. We fit the following two treatment hazard models:

$$\lambda_i^T(t|\boldsymbol{\mathcal{H}}_i(t), \mathcal{E}_i(t)) \quad = \quad \mathcal{E}(s_{ik})\ \mathcal{E}_i(t)\lambda_0^T(t) \exp\{\boldsymbol{\theta}_0' \boldsymbol{Z}_i(t)\}, \tag{14}$$

$$\mathcal{E}(s_{ik}\ \lambda_{ik}^\dagger(t; s_{ik}|\boldsymbol{Z}_{i0}(s_{ik}), \mathcal{E}(s_{ik}) \quad = \quad \mathcal{E}(s_{ik})\lambda_{0k}^\dagger(t) \exp\{\boldsymbol{\theta}_1' \boldsymbol{Z}_i(s_{ik}))\}, \tag{15}$$

with model (14) assumed to be the correct model; model (15) is expected to be misspecified, but is only used to provide a weight stabilizer. We assume no-unmeasured-confounders for treatment, $\lambda_i^T(t|\boldsymbol{\mathcal{H}}_i(t)) = \lambda_i^T(t|\boldsymbol{\mathcal{H}}_i(D_i), D_i)$, and that $\lambda_i^T(t|\boldsymbol{\mathcal{H}}_i(t)) = \lambda_i^T(t|\boldsymbol{Z}_i(t))$. Note that $\lambda_i^T(t|\boldsymbol{\mathcal{H}}_i(t), \mathcal{E}_i(t))$ in (14) uses (total) follow-up time t (measured from time 0) as the time axis, conditions on information on $[0, t)$, while $\lambda_{ik}^\dagger(t; s_{ik}|\boldsymbol{Z}_{i0}(s_{ik}), \mathcal{E}(s_{ik}) = 1))$ in (15) uses (residual)

time since s_{ik} and conditions on the history over $[0, s_{ik}]$ given $[\mathcal{E}_i(s_{ik}) = 1]$. Parameters in (14) and (15) are estimated through standard partial likelihood (Cox, 1975).

As derived in Gong and Schaubel (2013), an appropriate weight function is given by

$$W_{ik}^A(t) \;=\; Y_{i0k}(t)\exp\{\Lambda_i^T(s_{ik}+t) - \Lambda_i^T(s_{ik})\}, \tag{16}$$

where $\Lambda_i^T(t) = \int_0^t \mathcal{E}_i(u)\lambda_0^T(u)\exp\{\boldsymbol{\theta}_0'\boldsymbol{Z}_i(u)\}du$. The quantity $W_{ik}^A(t)$ can be thought of as the inverse of the conditional probability of remaining untreated at time $(s_{ik}+t)$, given that the subject was untreated and treatment-eligible at time s_{ik}. Gong and Schaubel (2013) suggest the following stabilized inverse weight,

$$W_{ik}^B(t) \;=\; Y_{i0k}(t)\frac{\exp\{\Lambda_i^T(s_{ik}+t) - \Lambda_i^T(s_{ik})\}}{\exp\{\Lambda_{ik}^\dagger(t)\}}. \tag{17}$$

Note that artificially censoring subjects at $t = L$ would be an alternative to the stabilizer.

Parameter Estimation for Model (12): An estimator for $\boldsymbol{\beta}_0$, denoted by $\widehat{\boldsymbol{\beta}}_0$, is obtained through solving the following inverse-weighted score function,

$$\boldsymbol{U}_0(\boldsymbol{\beta}) \;=\; \sum_{i=1}^n\sum_{k=1}^K\int_0^{\tau_{0k}} \mathcal{E}_i(S_{ik})\{\boldsymbol{Z}_{i0}(s_{ik}) - \overline{\boldsymbol{Z}}_{0k}(t;\boldsymbol{\beta},W)\}W_{ik}^B(t)dN_{i0k}(t), \tag{18}$$

with $\overline{\boldsymbol{Z}}_{0k}(t;\boldsymbol{\beta}_0) = \boldsymbol{R}_{0k}^{(1)}(t;\boldsymbol{\beta}_0)/R_{0k}^{(0)}(t;\boldsymbol{\beta}_0)$ and $\boldsymbol{R}_{0k}^{(d)}(t;\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \mathcal{E}_i(s_{ik})W_{ik}(t)\boldsymbol{Z}_{i0}(s_{ik})^{\otimes d}\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_{i0}(s_{ik})\}$ with $d = 0, 1, 2$ and where τ_{0k} satisfies $P\{Y_{i0k}(\tau_{0k}) = 1\} > 0$, and can in practice be set to the largest X_{ik} among subjects with $\mathcal{E}_i(s_{ik}) = 1$. A Breslow-Aalen estimator pooled across strata is obtained as

$$\widehat{\Lambda}_{00}(t;\widehat{\boldsymbol{\beta}}_0) \;=\; n^{-1}\sum_{i=1}^n\sum_{k=1}^K\int_0^t R_0^{(0)}(u;\widehat{\boldsymbol{\beta}}_0)^{-1}\mathcal{E}_i(s_{ik})W_{ik}^B(u)dN_{i0k}(u) \tag{19}$$

for $t \in (0, L]$, where $R_0^{(0)}(u;\boldsymbol{\beta}_0) = \sum_{k=1}^K R_{0k}^{(0)}(u;\boldsymbol{\beta}_0)$.

2.4.3 *Conditional Treatment Effect.* Consider patient i, treated at follow-up time $T_i = s$ with covariate history $\boldsymbol{\mathcal{H}}_i(s)$. Post-treatment survival probability for this patient is predicted by $\widehat{S}_1(t; s|\boldsymbol{\mathcal{H}}_i(s), T_i = s)$, while predicted $L-$year restricted mean post-treatment lifetime is given by $\widehat{\mu}_1(L; s|\boldsymbol{\mathcal{H}}_i(s), T_i = s)$. Correspondingly, in the absence of treatment, predicted

survival and L-year restricted mean lifetime for subject i (from T_i onward) would be given by $\widehat{S}_0(t;s|\boldsymbol{\mathcal{H}}_i(s),T_i=s)$ and $\widehat{\mu}_0(L|\boldsymbol{\mathcal{H}}_i(s),\mathcal{E}_i(s)=1)=\int_0^L \widehat{S}_0(t|\boldsymbol{\mathcal{H}}_i(s),\mathcal{E}_i(s)=1)dt$, respectively. The treatment effect corresponding to treatment initiation by subject i at follow-up time T_i can then be estimated by

$$\widehat{\delta}(t;T_i|\boldsymbol{\mathcal{H}}_i(T_i),T_i) \;=\; \widehat{S}_1(t;T_i|\boldsymbol{\mathcal{H}}_i(T_i),T_i)-\widehat{S}_0(t;T_i|\boldsymbol{\mathcal{H}}_i(T_i),\mathcal{E}_i(T_i)=1,T_i), \tag{20}$$

in terms of survival probability, and

$$\widehat{\Delta}(L;T_i|\boldsymbol{\mathcal{H}}_i(T_i),T_i) \;=\; \widehat{\mu}_1(L;T_i|\boldsymbol{\mathcal{H}}_i(T_i),T_i)-\widehat{\mu}_0(L;T_i|\boldsymbol{\mathcal{H}}_i(T_i),\mathcal{E}_i(T_i)=1,T_i) \tag{21}$$

in terms of restricted residual mean survival time.

2.4.4 *Average Treatment Effect.* Having established how to estimate the treatment effect for a subject treated at $T_i=s$ with covariate history $\boldsymbol{\mathcal{H}}_i(s)$, we now describe how to estimate the quantities of chief interest, namely $\delta(t)=E[\delta(t|\boldsymbol{\mathcal{H}}_i(s),T_i=s)]$ and $\Delta(L)=\int_0^L\delta(t)dt$ from (9). In the absence of censoring, we could average with respect to the empirical distribution of $\{T_i,\boldsymbol{\mathcal{H}}_i(T_i)\}$ values. Right censoring of T_i values rules out using the sample mean, since this averaging would then generally depend on the C_i distribution. This implies inverse weighting the observed treatment assignments, such that the inverse weighted distribution reflects that which would have been obtained in the absence of censoring. We use the result,

$$E\left[\int_0^t \frac{dN_i^T(u)}{G_i(u)}\,\middle|\,\boldsymbol{\mathcal{H}}_i(u)\right] \;=\; F_i^T(t|\boldsymbol{\mathcal{H}}_i(t)), \tag{22}$$

where $F_i^T(t|\boldsymbol{\mathcal{H}}_i(t))=E[\int_0^t dI(T_i\leqslant u)|\boldsymbol{\mathcal{H}}_i(u)]$ is analogous to the cumulative incidence function for T_i (with D_i serving as a competing risk) and with $G_i(u)=P(C_i>u|\boldsymbol{Z}_i(0))$. We assume the following proportional hazards model for C_i,

$$\lambda_i^C(t) \;=\; \lambda_0^C(t)\exp\{\boldsymbol{\alpha}_0'\boldsymbol{Z}_i(0)\}. \tag{23}$$

Observed data used to fit model (23) include $\{X_i,I(C_i<D_i),\boldsymbol{Z}_i(0)\}$, with $\boldsymbol{\alpha}_0$ and $\Lambda_0^C(t)=\int_0^t\lambda_0^C(u)du$ estimated through unweighted Cox regression. Note that C_i is viewed in this report as administrative censoring, in which case (23) may not even depend on $\boldsymbol{Z}_i(0)$. If in

fact $\lambda_i^C(t)$ depended on the $\boldsymbol{\mathcal{H}}_i(t)$, model (23) could easily be enriched to accommodate such dependence, with little subsequent modification to the procedures next described.

Finally, estimators of $\delta(t)$ and $\Delta(L)$ are given by

$$\widehat{\delta}(t) = \frac{\sum_{i=1}^n \int_0^\tau \widehat{\delta}(t;u|\boldsymbol{\mathcal{H}}_i(u), T_i = u)\, \widehat{G}_i(u)^{-1} dN_i^T(u)}{\sum_{i=1}^n \int_0^\tau \widehat{G}_i(u)^{-1} dN_i^T(u)}, \tag{24}$$

$$\widehat{\Delta}(L) = \int_0^L \widehat{\delta}(t)dt \tag{25}$$

respectively, where $\widehat{G}_i(u) = \exp\{-\widehat{\Lambda}_i^C(u)\}$, and with τ satisfying $P(X_i \geqslant \tau) > 0$ and typically chosen to be the maximum observed follow-up time.

3. Asymptotic Properties

We assume that the random vectors $\{X_i, N_i(X_i), N_i^T(X_i), \boldsymbol{\mathcal{H}}_i(X_i \wedge T_i)\}$ are independent and identically distributed for $i = 1 \ldots n$, with all elements of $\boldsymbol{\mathcal{H}}_i(t)$ bounded for $t \in (0, \tau]$. A complete list of regularity conditions is provided in the Supplementary Materials document.

Theorem 1*: Under certain regularity conditions,* $n^{1/2}\{\widehat{\delta}(t) - \delta(t)\}$ *and* $n^{1/2}\{\widehat{\Delta}(L) - \Delta(L)\}$ *each converge asymptotically to zero-mean Gaussian processes with covariance functions* $E[\xi_j(t)^2]$ *and* $E[\eta_j^2]$*, respectively, where* $\{\xi_1(t), \ldots, \xi_n(t)\}$ *and* $\{\eta_1(L), \ldots, \eta_n(L)\}$ *are i.i.d. with mean 0 asymptotically. Expressions for* $\xi_i(t)$ *and* $\eta_i(L) = \int_0^L \xi_i(t)dt$*, which are quite lengthly, are provided in the Supplementary Materials.*

Variance estimators for $\widehat{\delta}(t)$ and $\widehat{\Delta}(L)$ are given by $n^{-2}\sum_{i=1}^n \widehat{\xi}_i(t)^2$ and $n^{-2}\sum_{i=1}^n \widehat{\eta}_i(L)^2$, respectively; where $\widehat{\eta}_i(L)$ and $\widehat{\xi}_i(t)$ are computed by replacing all limiting values by their empirical counterparts. A proof of Theorem 1 is given in the Appendix. The essence of the proof is demonstrating that, asymptotically, $n^{1/2}\{\widehat{\delta}(t) - \delta(t)\} = n^{-1/2}\sum_{i=1}^n \xi_i(t) + o_p(1)$ through a sequence of Taylor series expansions and applications of empirical process results.

The proof is provided for the weight, $\widehat{W}_{ik}^A(t)$. In practice, the stabilized weight, $\widehat{W}_{ik}^B(t)$ would often be preferred. As implied by Theorem 1, the computation of the variance is quite involved, and such computation becomes more complicated when a stabilizer is incorporated.

Such concerns motivate a computationally simpler form for the variance estimator, resulting from taking $\widehat{G}_i(t)^{-1}$ and $\widehat{W}_{ik}^A(t)$, or $\widehat{W}_{ik}^B(t)$ as the case may be, as fixed. Variance estimators for $\widehat{\delta}(t)$ and $\widehat{\Delta}(L)$ then simplify considerably. We evaluate the performance of these simplified variance estimators through simulation in Section 4.

4. Simulations

We generated treatment-free survival to follow the assumed partly conditional model using methods from Gong and Schaubel (2013). First, subject i enters the study on calendar date, B_i, which is generated from a Uniform$(0, b)$ distribution. We then generate a single binary (0,1) group indicator Z_{ia}, taking the value 1 with probability 0.5. A longitudinal covariate, $Z_i(s_{ik})$, is then created and assumed to be measured at a common set of cross-section dates: $CS_1, CS_2, \ldots, CS_K$. To generate data $\{D_i, Z_{ia}, Z_{ib}\}$ where $\boldsymbol{Z}_{ib} = vec\{Z_i(s_{ik})\}$, we first let $Z_{ib0} = b_i + \sum_{k=1}^{K} \log(V_{ik})/\gamma_2$, where $b_i \sim N(\mu, \sigma^2)$ and $V_{ik} \sim P(\rho)$, independent positive stable random variables with index ρ. A pre-treatment death time, D_i^0, is then generated with hazard $\lambda_{i0}(t) = V_{i0}^{1/\rho}\lambda_0(t)\exp\{\gamma_1 Z_{ia} + \gamma_2 Z_{ib0}\}$, where $V_{i0} \sim P(\rho)$ and is independent of V_{ik}, with $\Lambda_0(t) = (t/a)^{1/\rho^2}$ and a is a constant. Setting $Z_i(s_{ik}) = Z_{ib0} - \log(V_{ik})/\gamma_2$, the pre-treatment death hazard can then be written as $\lambda_{i0}(t) = V_{i0}^{1/\rho}\lambda_0(t)\exp\{\gamma_1 Z_{ia} + \gamma_2 Z_i(s_{ik}) + \log(V_{ik})\}$. Treatment time, T_i, is generated from the proportional hazards model, $\lambda_i^T(t) = \lambda_0^T(t)\exp\{\theta_{01} Z_{ia} + \theta_{02} I(R_i > t)\}$, where $\lambda_0^T(t) = d_3$ and $\boldsymbol{\theta}_0' = (\theta_{01}, \theta_{02})$ and the time of treatment-ineligibility, R_i, is generated with hazard $\lambda_i^R(t) = \lambda_0^R(t)\exp\{d_1 V_{i0}\}$, where $\lambda_0^R(t) = d_2$. Thus, R_i and D_i are positively correlated, which is consistent with the data which motivated the proposed methods. Independent censoring time, C_i, is generated from hazard $\lambda_i^C(t) = \lambda_0^C(t)\exp\{\alpha_0 Z_{ia}\}$, where $\lambda_0^C(t) = d_4$. Note that treatment time and pre-treatment death time, T_i, and D_i are dependent since both depend on treatment-ineligibility time, R_i. However, the independent censoring time C_i is independent of D_i conditional on Z_{ia}.

After obtaining the pertinent survival function, transforming the time scale to represent

time since cross-section (setting $t_k = t - s_{ik}$), then averaging, we obtain

$$\lambda_i(t_k|Z_{ia}, Z_i(s_{ik}), D_i > s_{ik}) = \frac{\lambda_0(t_k + s_{ik})\rho^2\{\Lambda_0(t_k + s_{ik})\}^{(\rho^2-1)}}{\cos(\pi\rho/2)^{(\rho+1)}} \exp\{\rho^2\gamma_1 Z_{ia} + \rho^2\gamma_2 Z_i(s_{ik})\}.$$

Setting $\Lambda_0(t) = (t/a)^{1/\rho^2}$ and $\lambda_0(t_k + s_{ik})\rho^2\{\Lambda_0(t_k + s_{ik})\}^{(\rho^2-1)} = 1/a$ yields

$$\lambda_i(t_k|Z_{ia}, Z_i(s_{ik}), D_i > s_{ik}) \;\; = \;\; \exp\{\rho^2\gamma_1 Z_{ia} + \rho^2\gamma_2 Z_i(s_{ik})\}/[a\cos(\pi\rho/2)^{(\rho+1)}].$$

If we define $\lambda_{i0k}(t; s_{ik}) = \lambda_i(t_k|Z_{ia}, Z_i(s_{ik}), D_i > s_{ik})$, $\lambda_{00k}(t) = [a\cos(\pi\rho/2)^{(\rho+1)}]^{-1}$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}) = (\rho^2\gamma_1, \rho^2\gamma_2)$, then the proportional hazards model for pre-treatment death time is given by $\lambda_{i0k}(t; s_{ik}) = \lambda_{00k}(t)\exp\{\boldsymbol{\beta}_{01}Z_{ia} + \boldsymbol{\beta}_{02}Z_i(s_{ik})\}$.

For patients who received treatment prior to dying ($D_i > T_i$), a post-treatment death time $(D_i^1 - T_i)_+$, is then generated via the hazard, $\lambda_{i1}(t; T_i) = \lambda_{01}(t)\exp\{\boldsymbol{\beta}_{11}Z_{ia} + \boldsymbol{\beta}_{12}Z_i(T_i)\}$, where t represents time from treatment and $\boldsymbol{\beta}_1' = (\beta_{11}, \beta_{12}) = (\rho^2\gamma_1, \rho^2\gamma_2)$. We set $\lambda_{01}(t) = a_1$.

The complexity in the data generator is necessary to induce the partly conditional structure of the pre-treatment survival model. The positive stable frailty has become a common choice in the simulation of multivariate survival set-ups due to its preservation of the proportional hazards assumption both conditionally and marginally. Analogous set-ups were used by Zheng and Heagerty (2005) and Gong and Schaubel (2013).

We used $K = 10$ cross section dates, with $CS_k = 100 \times k$. For the simulation results presented, parameter specifications were as follows: $b = 500$, $(\theta_{01}, \theta_{02}) = (-1, -1)$, $\mu = 18$, $\sigma = 1$, $(\gamma_1, \gamma_2) = (-1, -0.5)$, $d_1 = d_2 = d_3 = d_4 = 0.001$, and $\rho = 0.8$, which implies $(\beta_{01}, \beta_{02}) = (\beta_{11}, \beta_{12}) = (-0.64, -0.32)$; We varied a from $a = 2000$, to $a = 5000$ and $a = 7000$, which led to treatment initiation rates of 10%, 15% and 20%, respectively; with similar independent censoring rates in each case. Each data configuration was replicated 1000 times, with $n = 500$ subjects per replicate.

We present settings where treatment has no effect ($\delta(t) = \Delta(L) = 0$), for which $a_1 = [a\cos(\pi\rho/2)^{(\rho+1)}]^{-1}$. We also list results for a setting with a positive treatment effect ($\delta(t) > 0$, $\Delta(L) > 0$) induced by specifying $a_1 = 0.5 \times 10^{-4}$. In developing appropriate parameter

settings, we conceptualized the time scale as representing days. For reporting purposes, time is recorded in years, with results presented for $\widehat{\delta}(1)$, $\widehat{\delta}(2)$, $\widehat{\delta}(3)$ and $\widehat{\Delta}(3)$. The weight $\widehat{W}_{ik}^{B}(t)$ was used throughout, with the simplified variance estimators applied.

Table 1 presents simulation results for settings with $\Delta(L) = 0$ and $\Delta(L) > 0$. The quantity $\Delta(L)$, with $L = 3$, can be interpreted as the difference of 3-year restricted mean survival time due to treatment, among the treated. The proposed estimators appear to be approximately unbiased, with coverage probabilities close to the nominal 95% level. Some degree of under-coverage is observed, which is due to the approximation of the results from Section 3 by treated the (random) weights as fixed. The under-coverage is not in unacceptable amounts, particularly relative to the great reduction in complexity and hence computational burden associated with the approximation.

[Table 1 about here.]

We examined the performance of the proposed methods under various degrees of model misspecification (see Supplementary Materials). The methods generally perform adequately, although some bias is introduced, and increases with increasing model misfit. The method appears to be most sensitive to misspecification of the treatment initiation hazard.

5. Application to Liver Transplant Data

We applied the proposed methods to estimate the average effect of liver transplantation among the transplanted, by Model for End-stage Liver Disease (MELD) score. This study used data 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 U.S., submitted by the members of the Organ Procurement and Transplantation Network (OPTN), and has been described elsewhere. The Health Resources and Services Adminis-

tration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors.

The study population included patients age $\geqslant$ 18 wait listed between 03/01/2002 and 12/31/2009. We excluded patients who were Status 1 (acute liver failure) or previously transplanted. Cross-section dates were chosen every 7 days, 30 days or 90 days from 03/01/2002 to 12/31/2009, which led to K=409, 96, or 32 cross sections respectively. The transplant hazard model, $\lambda_{ir}^{T}(t) = \mathcal{E}_i(t)\lambda_{0r}^{T}(t)\exp\{\boldsymbol{\theta}_0'\boldsymbol{Z}_i(t)\}$, was stratified by United Network for Organ Sharing (UNOS) Region ($r = 1, \ldots, 11$). The covariate, $\boldsymbol{Z}_i(t)$, included MELD score, albumin, age, gender, race, diagnosis of Hepatitis C, body mass index, diabetes, hospitalization, blood type, dialysis within prior week, encephalopathy, ascites and serum creatinine.

The pre-transplant death model, $\lambda_{i0kr}(t) = \lambda_{00kr}(t)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i(s_{ik})\}$, was also stratified, where $k = 1, \ldots, K$ stands for cross section and r again denotes UNOS Region. The covariate, $\boldsymbol{Z}_i(s_{ik})$, included MELD score, albumin, age, gender, race, diagnosis, body mass index, diabetes, hospitalization status at listing, previous dialysis, malignancy, time on wait-list (i.e., s_{ik} itself), slope of MELD score over $[0, s_{ik}]$, slope of albumin, percentage of time spent in inactive status, and percent of time receiving dialysis. In the post-transplant death model, $\lambda_{i1}(t; T_i) = \lambda_{01}(t)\exp\{\boldsymbol{\beta}_1'\boldsymbol{Z}_{i1}(T_i)\}$, $\boldsymbol{Z}_{i1}(T_i)$ included terms for T_i, MELD score, albumin, age, gender, race, diagnosis, body mass index, diabetes, hospitalization status at listing, previous dialysis and malignancy and Donor Risk Index (DRI; Feng et al., 2006).

The pre-transplant study sample consisted of $n = 66,884$ patients, of which 34,539 were observed to receive a deceased-donor liver transplant. For the MELD 30-40 subgroup, weekly cross section dates were chosen. For MELD 18-29 cross sections were drawn monthly. For MELD 6-17, cross sections were drawn every 3 months. Note that, we also tried weekly cross section dates for MELD 6-29 patients, which yielded almost identical results. The analysis was based on the weight, $W_{ik}^{B}(t)$.

Figure 3 shows the estimated survival curves for MELD groups 6-8, 15-17, 20-22 and 36-40. Note that the MELD score categories refer to MELD at transplant. Within a MELD category, $\widehat{S}_1(t)$ can be interpreted as the average survival probability, with t representing residual time post-transplant. Analogously, $\widehat{S}_0(t)$ can be interpreted as the average survival that would have resulted in the absence of liver transplantation, among patients who received a liver transplant. For the MELD 6-8 group, survival in the absence-of-transplantation exceeds post-transplant survival until approximately $t = 2$ years post-transplant. However, $\widehat{S}_1(t) > \widehat{S}_0(t)$ for $t > 2$ years, with the distance between the curves widening as t increases. The early survival advantage (absence-of-transplant versus with a transplant) for patients in the MELD 6-8 group is the combination of relatively mortality in this subgroup, combined with the risk of surgery-related mortality (not faced unless transplantation occurs). The early survival advantage without transplant is even observed in MELD 15-17 patients, but is much less pronounced and very short-lived. In fact, $\widehat{S}_1(t) > \widehat{S}_0(t)$ for $t > 0.25$ years in this subgroup. For MELD 36-40 group, the absence-of-transplant survival curve drops dramatically during the first couple of months, then steadily declines thereafter. Note that $\widehat{S}_1(t)$ curves are quite similar across MELD subgroups, with $\widehat{S}_0(t)$ decreasing strongly as MELD increases.

[Figure 3 about here.]

In Table 2, we list estimates of the difference in survival probability, $\widehat{\delta}(t)$ for $t = 1, 3, 5$ years, as well as $\widehat{\Delta}(5)$, the difference in 5-year restricted mean residual lifetime. The group that benefits the most from liver transplantation is clearly MELD 36-40, with an average gain in residual survival time of $\widehat{\Delta}(5) \approx 2.4$ years. The next greatest gain is observed in the MELD 30-35 group, with $\widehat{\Delta}(5) = 1.4$ years. For MELD scores between 15 and 30, there is little difference in the gain in 5-year restricted mean residual survival time, with $\widehat{\Delta}(5)$ fluctuating about 1 year across the MELD 26-29, 23-25, 20-22, 18-19 and 15-17 subgroups. Only for the MELD 6-8 group is $H_0 : \Delta(5) = 0$ not rejected.

In the Supplementary Materials, we provide results based on the Sequential Stratification method (Schaubel, Wolfe and Port, 2006; Schaubel et al., 2009), which features inverse weighted time-dependent stratification to create customized comparisons groups for each subject receiving the time-dependent treatment. Comparing our results in Table 2 to those based on Sequential Stratification, the main difference is in the MELD 6-8 group; the models from Sharma et al (2015) report a hazard ratio of 2.04 ($p < 10^{-4}$), indicating that liver transplant is associated with a doubling of the mortality hazard in this subgroup. In the presence of non-proportionality (which is clear in Figure 3, particularly for this subgroup), the hazard ratio and difference in restricted mean do not have to agree.

[Table 2 about here.]

Additional analysis is presented in the Supplementary Materials. For each MELD category, multiplying the number of transplants by the $\widehat{\delta}(5)$ yields the number of life-years saved via liver transplantation (considering only the first 5 post-transplant). The largest number of transplants was in the MELD 15-17 category (5,028), but the greatest number of life-years saved (7,649) was in the MELD 36-40 group. We estimate that 34,757 years of life were spared based on the liver transplants observed in this analysis. The Supplementary Materials also present plots of pre-transplant MELD profiles over time, the baseline pre-transplant mortality hazard, the liver transplant baseline hazard, and cumulative incidence of transplantation.

6. Discussion

In this report, we develop methods for estimating the average effect on the treated of a time-dependent treatment. The methods can be used to evaluate the benefit, in terms of patient survival, of a treatment under current treatment assignment practices. The methods were applied to quantify the survival benefit of deceased-donor liver transplantation among the transplanted, by Model for End-stage Liver Disease (MELD) score.

The proposed methods are not intended to guide treatment decisions. For example, the fact that we estimate a larger treatment effect for MELD 36-40 than for 30-35 does not imply that a patient with MELD=32 should wait until his/her MELD score increases to $\geqslant$ 36 before they agree to be transplanted. The proposed methods cannot generally be used to compare treatment effects, since each treatment effect is averaged differently. For example, the difference in the treatment effect between patients transplanted at MELD 15-17 ($\widehat{\Delta}(5) = 1.00$) and MELD 12-14 ($\widehat{\Delta}(5) = 0.59$) is partly attributable to the former group being transplanted with higher quality donor livers.

There are now many methods available for evaluating a time-dependent treatments. Marginal Structural Models (MSM; e.g., Hernán, Brumback and Robins, 2000; Robins, Hernán and Brumback, 2000) are not well-suited to our set-up due to the potential for treatment to interact with time-varying covariates. Structural Nested Failure Time Models (SNFTMs; e.g., Robins, 1988; Joffe et al., 1998; Keiding et al., 1999; Hernan et al., 2005; Taubman et al., 2009; Vock et al., 2013) are an alternative. In particular, the method of Vock et al. (2013) was motivated by the lung transplant setting. Versions of Sequential Stratification, which involves stratified and inverse weighted Cox regression, have been used to evaluate the benefit of kidney transplantation (Schaubel, Wolfe and Port, 2006) and liver transplantation (Schaubel et al., 2009). An advantage the proposed method over SNFTMs and Sequential Stratification is the avoidance of any parametric assumptions regarding the treatment effect. SNFTMs assume that treatment alters the time scale through a constant, while Sequential Stratification assumes proportionality of the pre- and post-treatment hazard functions. A further advantage of our proposed methods over SNFTMs relates to implementation. Although explicit coding would be required for either approach, the 'core' models in our method merely involve Cox regression and, therefore, can be fitted using standard statistical software (SAS, R) after modifying the input data appropriately.

In estimating the ETT, we consider the absence of treatment; i.e., $T_i = \infty$. In setting where this is found to be too ambitious a goal (e.g., lack of sufficiently long follow-up, in a setting where treatment is inevitable), one could change $[T_i = \infty]$ to $[T_i > L]$ in describing the absence-of-treatment scenario.

An alternative to the measures proposed in (9) and (10) would be to redefine $S_1(t)$ to be the population average survival (i.e., averaging over the current treated and untreated experiences), with $S_0(t)$ then representing the average population survival in the absence of treatment. Unless strong or unrealistic assumptions were made, the 'core' models for this approach would be quite similar to those in the proposed approach, except for the pre-treatment hazard model. The proposed averaging would be preferred in many practical settings (including the liver transplant setting which motivated our current work) since the absence of a treatment benefit among non-recipients is made explicit.

Supplementary Materials

Supplementary Materials, referenced in Sections 3, 4 and 5, are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This work was supported in part by National Institutes of Health Grant R01-DK070869. The data reported here have been supplied by the Minneapolis Medical Research Foundation (MMRF) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. The authors wish to thank the Associate Editor and two Reviewers, whose comments and suggestions led to considerable improvement of the manuscript. They also thank Min Zhang for her many thoughtful suggestions.

References

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10,** 1100–1120.

Breslow, N. E. (1972). Contribution to the discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society Series B* . **34,** 216–217.

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

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,** 269–275.

Feng, S., Goodrich, N.P., Bragg-Gresham, J.L., Dykstra, D.M., Punch, J.D., DebRoy, M.A., Greenstein, S.M. and Merion, R.M. (2006). Characteristics associated with liver graft failure: The concept of a donor risk index. *American Journal of Transplantation* **6,** 783–790.

Feuer, E.J., Hankey, B.F., Gaynor, J.J., Wesley, M.N., Baker, S.G. and Meyer, J.S. (1992). Graphical representation of survival curves associated with a binary non-reversible time dependent covariate. *Statistics in Medicine* **11,** 455–474.

Gong, Q. and Schaubel, D.E. (2013). Partly conditional estimation of the effect of a time-dependent factor in the presence of dependent censoring. *Biometrics*, **69,** 338–347.

Hernán, M.A., Cole, S.R., Margolick, J., Cohen, M., and Robins, J.M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and drug safety* **14**, 477–491.

Joffe, M.M., Hoover, D.R., Jacobson, L.P., Kingsley, L., Chmiel, J.S., Visscher, B.R. and Robins, J.M. (1998). Estimating the effect of Zidovudine on Karposi's sarcoma form observational data using a rank preserving structural failure time model. *Statistics in*

Medicine **17,** 1073–1102.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **282,** 457–481.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data: 2nd Edition*, New York: Wiley.

Keiding, N., Filiberti, M., Esbjerg, S., Robins, J.M. and Jacobsen, N. (1999). The graft versus leukemia effect after bone marrow transplantation: A case study using structural nested failure time models. *Biometrics* **57,** 23–28.

Parast, L., Tian, L. and Cai, T. (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of the American Statistical Association* **109,** 383–394.

Pearl, J. *Causality: Models, Reasoning, and Inference* (2009). New York: Cambridge.

Robins, J.M. (1988). The control of confounding by intermediate variables. *Statistics in Medicine* **8**, 679-701.

Robins, J. M., and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology - Methodological Issues*, N. Jewell, K. Dietz, and B. Farewell (eds), 297-331. Boston: Birkhäuser.

Rubin, D.B. (1978). Bayesian inference for causal effect: The role of randomization. *Annals of Statistics* **6,** , 34–58.

Schaubel, D.E., Wolfe, R.A. and Port, F.K. (2006). A sequential stratification method for estimating the effect of a time-dependent experimental treatment in observational studies. *Biometrics* **62,** , 910–917.

Schaubel, D.E., Wolfe, R.A., Sima, C.S. and Merion, R.M. (2009). Estimating the effect of a time-dependent treatment by levels of an internal time-dependent covariate. *Journal of the American Statistical Association* **104,** , 49–59.

Sharma, P., Schaubel, D.E., Goodrich, N.P. and Merion, R.M. (2015). Serum sodium and the survival benefit of liver transplantation. *Liver Transplantation* **21,** 308–313.

Van Houwelingen, H.C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34,** , 70–85.

Van Houwelingen, H.C. and Putter, H. (2012). *Dynamic prediction in clinical survival analysis*. New York: CRC Press.

Van Houwelingen, H.C. and Putter, H. (2015). Comparison of stopped Cox regression with direct methods such as pseudo-values and binomial regression. *Lifetime Data Analysis*, 21, 180-196.

Vock, D.M., Tsiatis, A.A., Davidian, M., Laber, E.B., Tsuang, W.M., Finlen Copeland, C.A. and Palmer, S.M. (2013). Assessing the causal effect of organ transplantation on the distribution of residual lifetime. *Biometrics* **69,** 820–829.

Zhang, M. and Schaubel, D.E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics* **67,** 740–749.

Zhang, M. and Schaubel, D.E. (2012). Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics* **68,** 999–1009.

Zheng, Y. Y. and Heagerty, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics*, **61,** 379–391.

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

Received January 2012. *Revised January* 2012. *Accepted January* 2012.

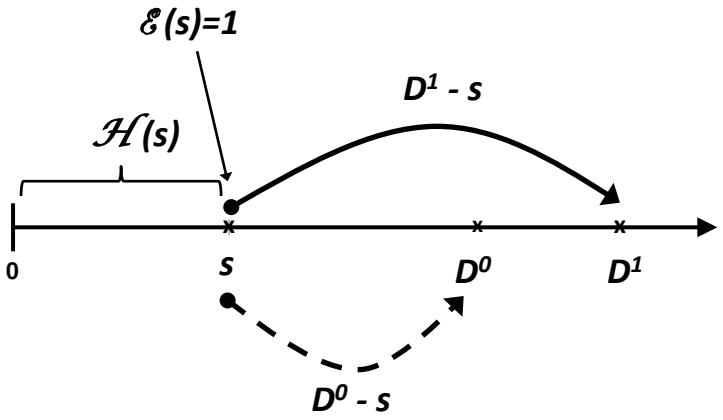
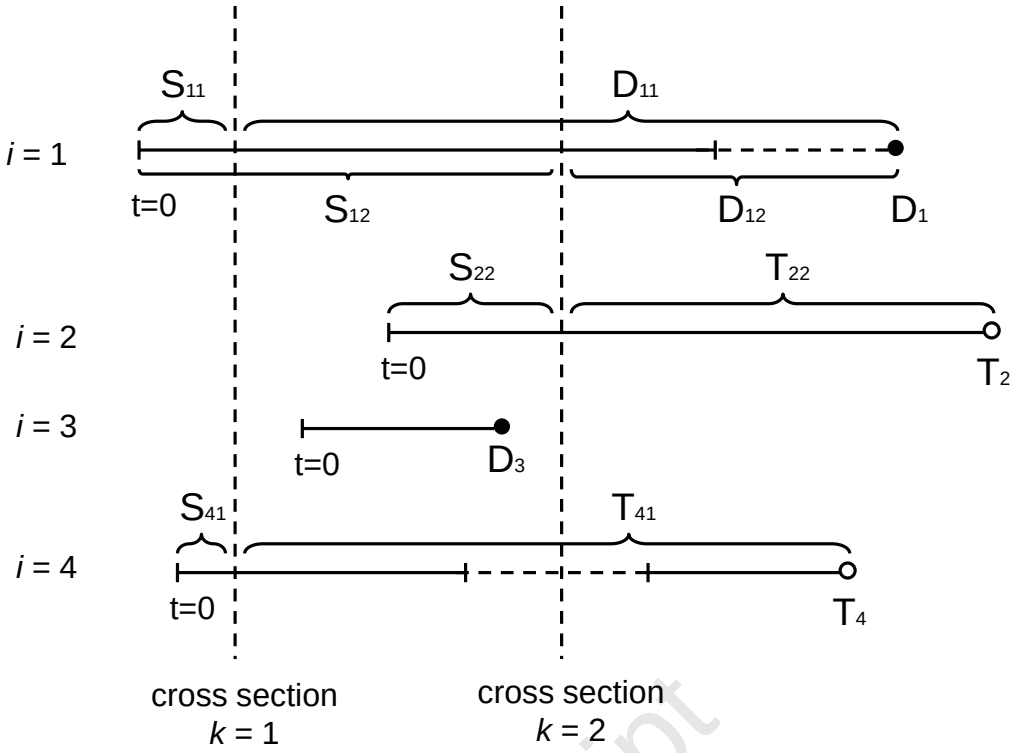


Figure 1. History on $[0, s)$ and residual survival beyond s under two scenarios. In each case, covariate (demographic, biological) and treatment-eligibility history $\boldsymbol{\mathcal{H}}(s)$ has accumulated, and the subject is treatment-eligible at time s, $\mathcal{E}(s) = 1$. Under Scenario (1), $T = s$ and $(D^1 - s)_+$ represents residual survival post-treatment. Under Scenario (0), $T = \infty$ since treatment is never available, such that death time is given by D^0 and residual survival beyond time s equals $(D^0 - s)_+$. Under the proposed methods, for each treated subject, partly conditional modeling is used to project $(D^1 - s)_+$ and $(D^0 - s)_+$ given $[\boldsymbol{\mathcal{H}}(s), T = s]$. The proposed effect-of-treatment-on-the-treated is then obtained after averaging over $[\boldsymbol{\mathcal{H}}(T), T]$.



Note: Vertical dashed lines denote cross-section dates, while horizontal dashed lines denote treatment-ineligible period.

Figure 2. Examples of the relationship between cross-section time and follow-up time. Four subjects ($i = 1, \ldots, i = 4$) and two cross sections ($k = 1, 2$) are shown. The four subjects begin follow-up at different calendar dates. For subject $i = 1$, failure times D_{11} and D_{12} correspond to cross sections $k = 1$ and $k = 2$, respectively. Note subject $i = 1$ is not censored at the treatment-ineligible time after cross section $k = 2$. Subject $i = 2$ is treated and, hence, dependently censored at time T_{22} following cross section $k = 2$. Subject $i = 3$ is excluded from cross-section $k = 1$ and $k = 2$ due to starting and finishing follow-up between CS_1 and CS_2. Subject $i = 4$ is included in cross section $k = 1$, but then becomes (and remains) treatment-ineligible until some a time after cross section $k = 2$. With respect to cross section $k = 1$, subject $i = 4$ is censored at treatment time T_{41}, as opposed to being censored earlier at the beginning of the treatment-ineligible period. Subject $i = 4$ is treatment-ineligible at cross section $k = 2$ and, hence, not included in CS_2.

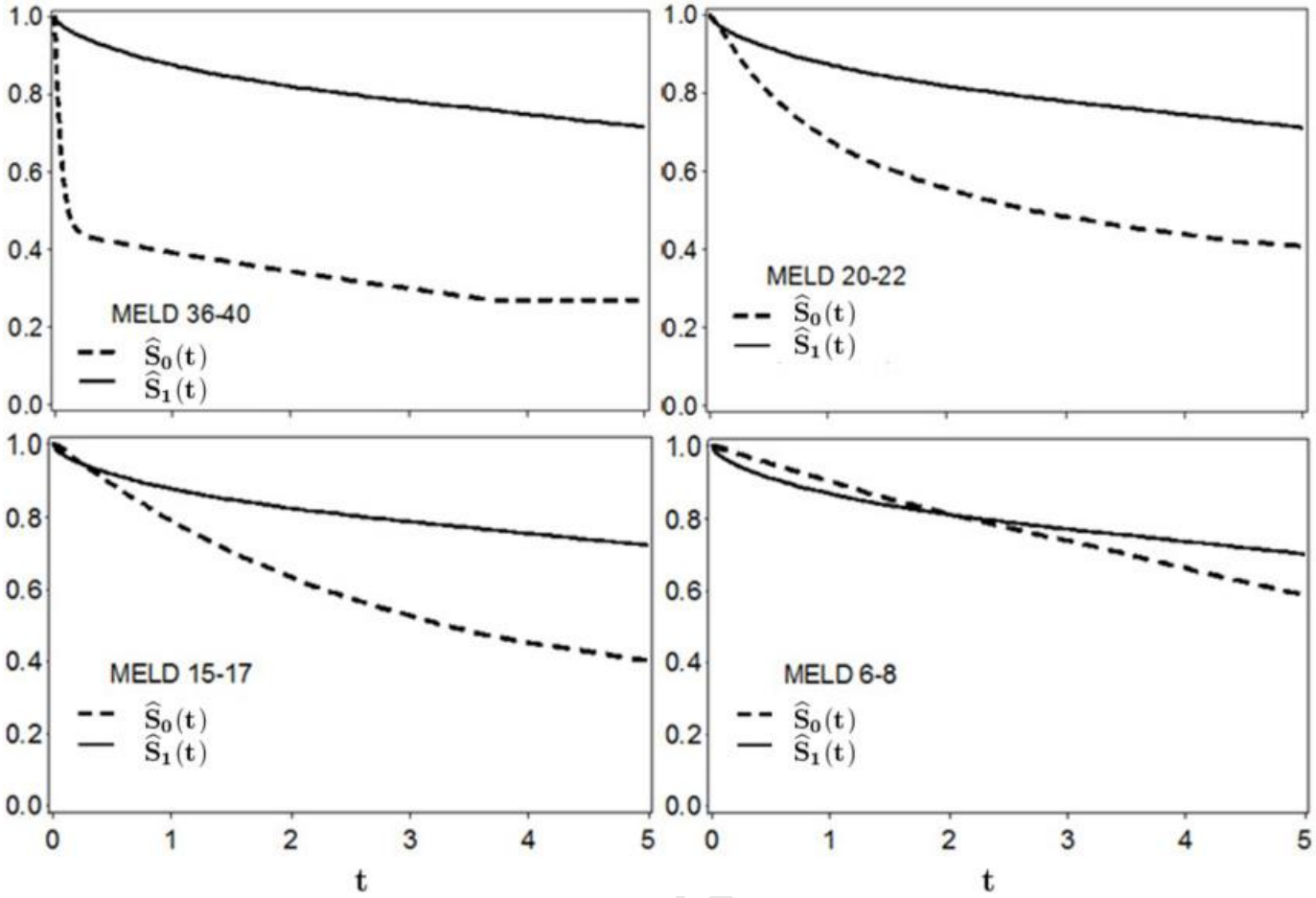


Figure 3. Analysis of SRTR data: Estimated survival curves after with a liver transplant (solid line) and in the absence of liver transplantation (dashed line) among liver transplant recipients. The time axis t is years post-transplant.

Table 1
Simulation results: $n = 500$, with weight function $W_{ik}^{B}(t)$

Setting	$E[N_i^C(\tau)]$	$E[N_i^T(\tau)]$	Parameter	True	BIAS	ESE	ASE	CP
1	0.10	0.10	$\Delta(3)$	0	0.040	0.204	0.190	0.92
			$\delta(1)$	0	0.012	0.089	0.082	0.92
			$\delta(2)$	0	0.016	0.092	0.085	0.93
			$\delta(3)$	0	0.022	0.094	0.082	0.91
2	0.15	0.15	$\Delta(3)$	0	0.022	0.164	0.154	0.93
			$\delta(1)$	0	0.007	0.065	0.061	0.93
			$\delta(2)$	0	0.010	0.077	0.072	0.93
			$\delta(3)$	0	0.010	0.083	0.077	0.91
3	0.20	0.20	$\Delta(3)$	0	0.009	0.144	0.141	0.94
			$\delta(1)$	0	0.001	0.056	0.054	0.93
			$\delta(2)$	0	0.004	0.067	0.066	0.94
			$\delta(3)$	0	0.005	0.074	0.073	0.94
4	0.10	0.10	$\Delta(3)$	0.87	0.030	0.204	0.190	0.92
			$\delta(1)$	0.29	0.009	0.088	0.074	0.92
			$\delta(2)$	0.35	0.009	0.100	0.088	0.92
			$\delta(3)$	0.35	0.008	0.110	0.097	0.92
5	0.15	0.15	$\Delta(3)$	0.61	0.017	0.150	0.145	0.94
			$\delta(1)$	0.19	0.006	0.054	0.052	0.94
			$\delta(2)$	0.25	0.008	0.070	0.068	0.94
			$\delta(3)$	0.28	0.005	0.082	0.077	0.92
6	0.20	0.20	$\Delta(3)$	0.43	0.020	0.135	0.133	0.94
			$\delta(1)$	0.13	0.006	0.048	0.048	0.94
			$\delta(2)$	0.18	0.009	0.064	0.062	0.93
			$\delta(3)$	0.20	0.006	0.077	0.072	0.93

ESE = empirical standard error; ASE = asymptotic standard error $CP = 95\%$ coverage probability; $E[N_i^C(\tau)]$ = proportion censored; $E[N_i^T(\tau)]$ = proportion treated; $\delta(t)$ and $\Delta(L)$ are as defined in (9) and (10), respectively.

Table 2
Analysis of SRTR data: Estimating the effect of liver transplantation on the transplanted (with 95% confidence interval in parentheses), by MELD score at transplant.

MELD Score	$\widehat{\delta}(1)$	$\widehat{\delta}(3)$	$\widehat{\delta}(5)$	$\widehat{\Delta}(5)$
6-8	-0.03 (-0.05, -0.01)	0.03 (-0.01, 0.05)	0.11 (0.07, 0.15)	0.11 (-0.03, 0.25)
9-11	-0.02 (-0.04, 0.00)	0.09 (0.07, 0.11)	0.17 (0.15, 0.19)	0.29 (0.15, 0.43)
12-14	0.02 (0.00, 0.04)	0.16 (0.12, 0.20)	0.23 (0.19, 0.27)	0.59 (0.43, 0.75)
15-17	0.09 (0.07, 0.11)	0.26 (0.22, 0.30)	0.32 (0.28, 0.36)	1.00 (0.80, 1.20)
18-19	0.15 (0.13, 0.17)	0.26 (0.24, 0.28)	0.27 (0.23, 0.31)	1.06 (0.90, 1.22)
20-22	0.19 (0.15, 0.23)	0.29 (0.23, 0.35)	0.30 (0.24, 0.36)	1.23 (0.95, 1.41)
23-25	0.19 (0.15, 0.23)	0.23 (0.19, 0.27)	0.26 (0.18, 0.34)	1.07 (0.79,1.35)
26-29	0.25 (0.17, 0.33)	0.19 (0.11, 0.27)	0.16 (0.06, 0.26)	0.99 (0.59, 1.39)
30-35	0.33 (0.25, 0.41)	0.27 (0.07, 0.47)	0.25 (0.01, 0.49)	1.45 (0.05, 2.85)
36-40	0.48 (0.40, 0.56)	0.48 (0.36, 0.60)	0.45 (0.33, 0.57)	2.38 (1.70, 3.06)

$\widehat{\delta}(t)$ and $\widehat{\Delta}(L)$ are as defined in (24) and (25) , respectively. The time scale represents years post-transplant.