Published in final edited form as:

Biometrics. 2015 December ; 71(4): 950–959. doi:10.1111/biom.12361.

Time-Dependent Prognostic Score Matching for Recurrent Event Analysis to Evaluate a Treatment Assigned During Follow-up

Abigail R. Smith* and Douglas E. Schaebel**

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Summary

Recurrent events often serve as the outcome in epidemiologic studies. In some observational studies, the goal is to estimate the effect of a new or ‘experimental’ (i.e., less established) treatment of interest on the recurrent event rate. The incentive for accepting the new treatment may be that it is more available than the standard treatment. Given that the patient can choose between the experimental treatment and conventional therapy, it is of clinical importance to compare the treatment of interest versus the setting where the experimental treatment did not exist, in which case patients could only receive no treatment or the standard treatment. Many methods exist for the analysis of recurrent events and for the evaluation of treatment effects. However, methodology for the intersection of these two areas is sparse. Moreover, care must be taken in setting up the comparison groups in our setting; use of existing methods featuring time-dependent treatment indicators will generally lead to a biased treatment effect since the comparison group construction will not properly account for the timing of treatment initiation. We propose a sequential stratification method featuring time-dependent prognostic score matching to estimate the effect of a time-dependent treatment on the recurrent event rate. The performance of the method in moderate-sized samples is assessed through simulation. The proposed methods are applied to a prospective clinical study in order to evaluate the effect of living-donor liver transplantation on hospitalization rates; in this setting, conventional therapy involves remaining on the wait list or receiving a deceased-donor transplant.

Keywords

Recurrent events; prognostic score matching; treatment effect; proportional rates model

1. Introduction

Events that can occur repeatedly for the same subject are often of interest in clinical settings. Examples include repeated hospitalizations, post-surgery complications, viral infections, or myocardial infarctions. Many methods for modeling recurrent events have been described in the literature, including conditional models (e.g., Andersen and Gill, 1982; Pepe and Cai, 1993) and marginal approaches (Lawless and Nadeau, 1995; Lin et al., 2000; Pena,

* abbysmit@umich.edu

** deschau@umich.edu

Strawderman and Hollander, 2001). A comprehensive survey of methodology for recurrent events is provided by Cook and Lawless (2007).

It may be of interest to study recurrent events as a basis for evaluating a treatment, and a limited number of methods are available for this purpose (e.g., Cook et al., 2009; Schaubel and Zhang, 2010). For chronic conditions, it is often the case that treatment begins some time after the start of follow-up, such that treatment is time-dependent. Although methods exist for analyzing recurrent events in the presence of time-varying covariates (Chen et al., 2013), such methods were not designed for evaluating treatments. Most recurrent event methods assume that time-dependent elements are external (Kalbfleisch and Prentice, 2002), which essentially rules out their application to evaluate time-dependent treatments that are not randomized.

In the setting we consider in this report, treatment is time-dependent, in that subjects typically begin follow-up untreated, with some going on to receive treatment. Two forms of treatment are available, a new ‘experimental’ treatment and the standard treatment. The conventional course for a subject is to start follow-up untreated, then possibly receive the standard treatment. Our objective is to compare the recurrent event rate for subjects receiving the experimental treatment to that which they would have experienced under conventional therapy (which would be a combination of ‘untreated’ and ‘standard treatment’). In Figure 1, we depict the recurrent event experience of the same subject i under two scenarios. For the top time line, subject i receives the experimental treatment at follow-up time s ; after time s , the event course changes in relation to what it would have been on conventional therapy. The bottom time line pertains to the same subject, but under a scenario where experimental treatment does not exist. The comparison of interest is the ratio of the recurrent event rates t time units following time s . Note that the comparison of interest begins at the time of experimental treatment initiation, s , since the experimental treatment would clearly have no effect on subject i during the $[0, s)$ time interval. Ideally, we would like to compare each experimentally treated patient with their ‘ghost’, who remains on the conventional course; naturally, we never observe the latter if we observe the former.

Additional features of the data structure we consider are as follows. There are longitudinal measures available on each subject; i.e., time-dependent covariates. Moreover, the event history itself (e.g., up to the time of treatment) naturally provides information on the anticipated event rate (e.g., post-treatment). Treatment is not randomized, and generally depends on the time-varying covariates and the observed recurrent event experience. Finally, the experimental treatment effect may not be homogeneous and could be modified by the time-dependent entities.

The above-described data structure arises in the study of hospitalization rates after living donor liver transplantation (LDLT). The standard treatment for end-stage liver disease (ESLD) is deceased donor liver transplantation (DDLT). However, LDLT is a relatively new treatment option for ESLD that involves using a partial graft from a living donor in place of the traditional whole graft from a deceased donor. Our goal is to compare hospitalization rates for a patient who receives an LDLT relative to what that patient's hospitalization rate would have been had they remained on the waiting list and potentially received a DDLT.

Note that a traditional analysis (e.g., proportional rates model) featuring a time-dependent experimental treatment indicator 0/1 would generally produce a biased estimate of the contrast of interest (described in the preceding paragraph). This is largely due to the comparison groups having not been constructed appropriately (i.e., in a manner which respects the timing of the treatment assignments across subjects). We illustrate this phenomenon in Section 4 through the data set which motivated our work.

Since it is not possible to observe the treatment-free event process for subjects that undergo treatment, a potential alternative is to compare those who receive a certain time-dependent treatment at time s to other similar subjects who were eligible to receive the treatment at time s but do not. The averaged event rates after s of matched subjects are then compared to the post-treatment event rate of the index subject. Methodology for this type of analysis currently exists for the univariate survival setting (Schaubel et al., 2009). In the current report, we extend this method to the recurrent event setting.

There are several key differences in the methods we propose in this report and those of Schaubel et al. (2009). First, the outcome is a recurrent event and, hence, more complex than a survival time. Second, we use a two-stage modeling approach. The first stage features a conditional rate model intended, not for interpretation, but to track each patient's recurrent event history and projected prognosis. The linear predictor from the Stage 1 model is used at Stage 2 to caliper-match each patient receiving the experimental treatment to yet-untreated patients as opposed to the hard matching and covariate adjustment used by Schaubel et al. (2009). The final model for the recurrent event rate then consists only of the treatment effect and a distance measure which accounts for any small differences in the pre-treatment event trajectories between the experimental patient and the matched controls. The matching we propose in this report aims to create counterfactual 'ghosts' to recover the conventional therapy experience of the experimentally treated patients.

The remainder of this report is organized as follows. In Section 2, we introduce the notation and describe the proposed methods, including parameter estimation and inference. A simulation study is presented in Section 3. In Section 4, the proposed methods are applied to data from the Adult-to-Adult Living Donor Liver Transplantation Cohort Study (A2ALL) to estimate the impact of LDLT on hospital admission rates relative to conventional therapy (remaining on the wait list and potentially receiving a DDLT). Some concluding remarks are offered in Section 5, including areas of future work and work currently in progress.

2. Methods

2.1 Notation and Data Set-up

In the development that follows, i represents subject ($i = 1, \dots, n$). The time-dependent covariate for subject i is denoted $\mathbf{Z}_i^*(t)$. Subjects are generally untreated at time 0, with treatment time given by $T_i = T_i^E \wedge T_i^S$, where T_i^E represents the time of 'experimental' treatment initiation, T_i^S denotes time of 'standard' treatment initiation and $a \wedge b = \min(a, b)$. Note that at most one of T_i^E and T_i^S occurs. Treatment (be it experimental or standard) is assumed to be non-reversible, in the sense that patients are considered 'treated' from the

time of treatment initiation forward. The true number of events for subject i in $[0, t]$ is defined as $N_i^*(t) = \int_0^t dN_i^*(u)$. Event and treatment times are subject to independent right censoring by C_i . The number of observed events is given by $N_i(t) = \int_0^t I(C_i > u) dN_i^*(u)$. Note that right censoring is administrative in this setting.

To fix ideas, we make the distinction between ‘treatment’ and ‘therapy’. In particular:

- There are two forms of *treatment*: *standard* and *experimental*
- *Conventional therapy* involves the treatment course that all patients would need to follow if the experimental version of the treatment did not exist. Under conventional therapy, patients begin follow-up *untreated*, with some going on to receive the *standard treatment*.

Our definition of *therapy* as being a sequence of *treatments* is not standard, but allows for convenient labeling. While untreated, a patient is actually following conventional therapy, with experimental treatment being initiated at time T_i^E . An untreated patient who later receives standard treatment is also following conventional therapy.

We define several event counting processes. The number of pre-treatment events in $(0, t]$ is given by

$$N_i^0(t) = \int_0^t I(T_i > u) dN_i^*(u).$$

Consider patient i at follow-up time s under two scenarios. In the first scenario, patient i receives the experimental treatment at time s ; i.e., $T_i^E = s$. The pertinent post-experimental treatment event counter is defined as

$$N_i^*(t; s) = I(T_i^E = s) \int_s^{s+t} dN_i^*(u). \quad (1)$$

Note that it will be our convention that $N(t; s)$ refers to the interval of length t , but starting at time s ; a single time index, as in the previously-defined $N_i^0(t)$, pertains to the $(0, t]$ time interval. Correspondingly, we define an event counter representing the events that would have been experienced in the absence of experimental treatment, also beginning at time s ,

$$N_i^{CT}(t; s) = I(T_i \geq s) \int_s^{s+t} I(T_i^E > u) dN_i^*(u), \quad (2)$$

where the superscript *CT* denotes ‘conventional therapy’. Parallelling $N_i^*(t; s)$, the counter $N_i^{CT}(t; s)$ tracks the patient on $(s, s + t]$. To emphasize the correspondence between (1) and (2), consider a subject eligible to receive the experimental treatment at time s , such that $I(T_i^E = s) = 1$; note that $I(T_i \geq s) = I(T_i^E = s) \cup (T_i^E > s, T_i^s > s)$. Under the setting where $T_i^E = s$, the counting process (1) takes effect; had the experimental therapy not been an option, process

(2) takes effect. An important point is that the patient is untreated on $(0, s)$ under either scenario to which (1) and (2) pertain.

Finally, we define a 0/1 process for being observed to receive the experimental treatment,

$$N_i^E(t) = \int_0^t I(C_i > u) dI(T_i^E \leq u). \quad (3)$$

As shown in Section 2.2, $N_i^E(t)$ is used to fit a model needed for inverse weighting.

2.2 Proposed Models and Estimation Methods

As described above, the goal of this method is to compare the post-experimental treatment recurrent event mean to the corresponding event mean under conventional therapy (CT). For a subject with $T_i^E = s$, the comparison would be on $(s, s + t]$ for $t > 0$, such that the pertinent counting processes are given in (1) and (2). We denote the mean of (1) by

$$\mu_i^*(t; s) = E \left[\int_0^t N_i^*(du; s) | T_i^E = s, \mathbf{H}_i(s) \right], \quad (4)$$

where $\mathbf{H}_i(s) = \{Z_i^*(u), N_i(u), I(T_i > u), I(C_i > u); 0 \leq u < s\}$ represents the observed pre-treatment history for subject i on $[0, s)$. As we describe shortly, we avoid explicitly modeling (4). Note that (4) represents a partly conditional model (Pepe and Couper, 1997; Zheng and Heagerty, 2005; Gong and Schaubel, 2013) since it conditions on the history only up to time s , not $s + t$. We make this choice for the sake of interpretation, but it is not without consequences with respect to parameter estimation, as detailed shortly. This idea of conditioning on only part of the history is connected to landmark analysis (van Houwelingen, 2007; van Houwelingen and Putter, 2008; 2012).

Having defined the post-experimental recurrent event mean in (4), the corresponding quantity (in the absence of experimental treatment) is given by

$$\mu_i^{CT}(t; s) = E \left[\int_0^t N_i^{CT}(du; s) | \mathbf{H}_i(s), T_i > s, T_i^E > u \right], \quad (5)$$

which, like (4), we avoid modeling directly. The model of chief interest is then given by

$$\mu_i^*(t; s) = \mu_i^{CT}(t; s) \exp\{\beta_\star\}, \quad (6)$$

which can equivalently be expressed in terms of a rate function by

$$\mu_i^*(dt; s) = \mu_i^{CT}(dt; s) \exp\{\beta_\star\}. \quad (7)$$

In this model $\mu_i^{CT}(t; s)$, the experimental treatment-free mean number of events is scaled up or down by $\exp\{\beta_\star\}$ if subject i received the experimental treatment at time s . The mean

number of post-experimental treatment events is then compared to the mean number of experimental treatment-free events after time s .

More general versions of model (7) are possible, including

$$\mu_i^*(dt;s) = \mu_i^{CT}(dt;s) \exp\{\beta_*(t;s)\}, \quad (8)$$

where $\beta_*(t; s)$ is a parametric function of both time of experimental treatment initiation (given by s) and time since $T_i^E = s$, given by t . Model (8) is more flexible, in the sense that the effect of receiving the experimental treatment, on the event rate, can depend on the time until and time since treatment was received.

In reality, once a subject receives the experimental treatment, we can no longer observe the subject's experience receiving conventional therapy. Therefore, a patient experimentally treated at time s will be compared to similar patients who did not initiate experimental treatment at follow-up time s but were eligible to do so. Similar to Schaubel et al. (2009), we use the concept that each experimental treatment time initiates an “experiment”, in which the recipient of the experimental treatment is compared to ‘similar’ treatment-eligible candidates. Note that ‘similar’, in this context, refers to current status (i.e., at time s) and history on $[0, s)$. Eligibility for the comparison is defined as

$$e_i(s) = I(T_i^E = s) + I(T_i > s)$$

i.e., at time s , patient i either received the experimental treatment or remained untreated.

Our method of estimating β_* from (7) involves a stratified analysis, as will be formalized shortly. Each experimental treatment patient generates a stratum, which will include the index patient as well as similar treatment-eligible patients. The most relevant definition of similar, here, is with respect treatment-eligibility at s , denoted by $e_i(s)$, and accumulated covariate and recurrent event history on $(0, s]$, represented by $\mathbf{H}_i(s)$. In order to quantify each subject's history, we use a prognostic score (Hansen, 2008) based on the pre-treatment event rate, modeled using a time-dependent proportional rates model,

$$d\mu_i^0(t) = E[dN_i^*(t) | \mathbf{H}_i(t), T_i > t] = \exp\{\boldsymbol{\alpha}_0^T \mathbf{Z}_i(t)\} d\mu_0(t), \quad (9)$$

where the covariate $\mathbf{Z}_i(t)$ is chosen to capture the pertinent components of the history, $E[dN_i^*(t) | \mathbf{H}_i(t), T_i > t] = E[dN_i^*(t) | \mathbf{Z}_i(t), T_i > t]$. Model (9) resembles the Lin et al. (2000) model, but is more accurately interpreted as an Andersen-Gill (1982) model, due to the explicit dependence on the prior event history, a property avoided by Lin et al. (2000). The regression parameter $\boldsymbol{\alpha}_0$ from (9) can be computed by solving the unweighted Cox (1972) score equation. Due to the dependence on internal covariates (Kalbfleisch and Prentice, 2002), elements of $\boldsymbol{\alpha}_0$ are difficult to interpret. This is not a problem, in our case, since we do not care to interpret, let alone carry out inference on $\boldsymbol{\alpha}_0$. Note that we focus on untreated experience since, at any time s , the patient generating the stratum (by initiating the

experimental treatment) and the potential matches are all necessarily untreated with either experimental or standard treatment on $[0, s)$.

The purpose of the prognostic score is to match patients that have similar pre-treatment event rates, the rationale being that previous event rate is the most important predictor of the current event rate. Unlike a propensity score, which uses the treatment event rate to match subjects with similar probabilities of being treated, the prognostic score aims to compare the effect of treatment on the event rate among subjects that were on the same trajectory with respect to their pre-treatment event rate. The use of prognostic scores in conjunction with, or as an alternative to, propensity scores has been considered in several reports (e.g., Rubin and Thomas, 2000; Stuart, Lee and Leacy, 2013; Leacy and Stuart, 2014; Li, Schaubel, and He, 2014) and will be addressed further in Section 5. Once the prognostic scores have been estimated, caliper matching is used to assign untreated control subjects to a subject receiving the experimental treatment at time s . Caliper matching requires that the prognostic scores of matched subjects be within a certain radius of the prognostic score of the index subject. Appropriate selection of the caliper involves balancing the need for homogeneity within stratum with the need to have an adequate number of matches for each index subject. We propose that prior to data analysis the caliper size be chosen such that the median number of matches is between 10 and 20. It is important to verify that some level of balance in the event rate at time s is achieved with the chosen caliper. Methods for checking balance in propensity matching have been discussed by Hansen (2008) and Harder, Stuart, and Anthony (2010). This will be explored further in the application to living donor liver transplant described in Section 4. The discrepancy between prognostic scores for experimental subject j and control subject i can be quantified through the subject-pair specific rate ratio,

$$\psi_{i,j}(s) = \frac{d\mu_i^0(s)}{d\mu_j^0(s)} = \exp\{\alpha_0^T [\mathbf{Z}_i(s) - \mathbf{Z}_j(s)]\}.$$

Subject i is considered to have a path on $[0, s)$ sufficiently similar to subject j if $|\log \psi_{ij}(s)| \leq \varepsilon$ where $\varepsilon > 0$ is a pre-determined constant.

Combining the eligibility indicators and prognostic scores, patient i is included in the stratum generated by patient j if $m_{ij}(s) = 1$, where

$$m_{ij}(s) = e_i(s)I(T_i > s)e_j(s)I(T_j^E = s)I(|\log \hat{\psi}_{ij}(s)| \leq \varepsilon),$$

with $\hat{\psi}_{ij}(s) = \exp\{\hat{\alpha}_0^T [\mathbf{Z}_i(s) - \mathbf{Z}_j(s)]\}$. In order to account for the residual difference between patients i and j , we propose to adjust for $\log \hat{\psi}_{ij}(s)$ in the final model. Incorporating the eligibility indicator and the prognostic score distance, the final fitted model for the event mean for stratum j is then

$$\mu_{ij}^*(t;s) = m_{ij}(s) \mu_i^{CT}(t;s) \exp\{\beta_\star I(T_i^E = s) + \beta_\psi \log \hat{\psi}_{ij}(s)\}. \quad (10)$$

To clarify the representation in model (10), j is the stratum (generated by patient j through $T_i^E = s$) and i is the patient within stratum. The model governs the experimental patient through the indicator $I(T_i^E = s)$, which equals 1 if $i = j$. The vector of parameters to be estimated and the corresponding covariates are given by

$$\beta_{\star\psi} = \begin{bmatrix} \beta_\star \\ \beta_\psi \end{bmatrix} \quad \mathbf{Z}_i^*(s) = \begin{bmatrix} I(T_i^E = s) \\ \log \hat{\Psi}_{ij}(s) \end{bmatrix}, \quad (11)$$

such that model (10) can be re-written as $\mu_{ij}^*(t;s) = m_{ij}(s) \mu_i^{CT}(t;s) \exp\{\beta_{\star\psi}^T \mathbf{Z}_i^*(s)\}$.

Subjects that are matched to the experimental subject enter the experiment without receiving any treatment, but could subsequently receive either the experimental treatment or standard treatment. If a patient receives the standard treatment after time s , follow-up in the experiment continues since the goal is to compare experimental treatment to conventional therapy (i.e., beginning untreated and potentially later receiving standard treatment). However, if a matched subject receives the experimental treatment after time s they are censored from all experiments in which they serve as controls and begin their own experiment as the index subject. This generally results in dependent censoring since, although $dI(T_i^E \leq s+t)$ can be considered random given $\mathbf{H}_i(s+t)$, the model for $\mu_{ij}^*(t;s)$ from (10) only conditions on $\mathbf{H}_i(s)$, the pre-treatment history up to time s . Since $\{\mathbf{H}_i(s+t); t > 0\}$ can affect $dI(T_i^E \leq s+t)$ and, by definition, affects $N_i(t; s)$, the censoring of matched controls via experimental treatment initiation constitutes dependent censoring.

We apply a variant of Inverse Probability of Censoring Weighting (IPCW; Robins and Rotnitzky, 1992; Miloslavsky et al., 2004) to account for this dependent censoring by modeling the censoring mechanism (experimental treatment) using the following model,

$$\lambda_i^E(t) = \lambda_0^E(t) \exp[\beta_E^T \mathbf{Z}_i(t)],$$

fitted through standard partial likelihood (Cox, 1975).

2.3 Parameter Estimation

We now describe the estimation of the parameter of interest, $\beta_\star$, through the estimation of $\beta_{\star\psi}$. To begin, we define the pertinent risk set indicator for stratum j ,

$$Y_{ij}(t;s) = m_{ij}(s) I(C_i > s+t) \{I(T_i^E = s) + I(T_i^E > s+t)\}$$

For now, suppose that, subsequent to being matched to an index experimental treatment patient j at time s , matched controls were randomly selected to receive experimental treatment given $H_i(s)$. In this restrictive case, the following process,

$$m_{ij}(s) \int_s^{s+t} Y_{ij}(u; s) \{N_i(du; s) - \mu_i(du; s)\} \quad (12)$$

would have mean zero.

As mentioned above, we will use IPCW to account for censoring of matched patients who receive the experimental treatment in each experiment. IPCW involves weighting using the inverse probability of being censored or, in this case, receiving the experimental treatment. Since the experiment begins at time s , the weight only accounts for the probability of receiving the experimental treatment during $[s, s+t)$. Correspondingly, we define the quantity

$$G_i(t; s) = \frac{G_i(s+t)}{G_i(s)} = \exp \left\{ - \int_s^{s+t} \lambda_i^E(u) du \right\}$$

Where

$$G_i(t) = \exp \left\{ - \int_0^t \lambda_i^E(u) du \right\}$$

The weight function can then be defined as

$$w_{ij}(t; s) = Y_{ij}(t; s) G_i(t; s)^{-I(T_i^E > s)},$$

such that the index experimental treatment patient is appropriately self-weighting; censoring after T_i^E can only occur independently via C_i .

Applying the weight function to (12) produces the zero-mean process,

$$m_{ij}(s) \int_0^{\tau-s} W_{ij}(u; s) M_{ij}(du; s), \quad (13)$$

where we define $M_{ij}(du; s) = Y_{ij}(u; s) \{N_i(du; s) - \mu_{ij}(du; s)\}$, with τ chosen to satisfy $P(C_i \leq \tau) > 0$ and often set to $\max\{C_1, \dots, C_n\}$. Aggregating across subjects for the experiment occurring at time s produces the set of zero mean processes,

$$\sum_{i=1}^n m_{ij}(s) \int_0^t W_{ij}(u; s) M_{ij}(du; s) \quad (14)$$

$$\sum_{i=1}^n m_{ij}(s) \int_0^t \mathbf{Z}_i^*(s) W_{ij}(u;s) M_{ij}(du;s). \quad (15)$$

We reorganize this system to solve implicitly for the baseline mean, $\mu_0^{CT}(u;s)$ in (14), then substitute into (15). Then, aggregating across all experiments yields the final estimating function for $\beta_{\star\psi}$

$$U(\beta) = \sum_{j=1}^n \sum_{i=1}^n \int_0^t m_{ij}(s) \int_0^{\tau-s} \{\mathbf{Z}_i^*(s) - \bar{\mathbf{Z}}_{\star}(u;s)\} \hat{W}_{ij}(u;s) N_i(du;s) dN_j^E(s), \quad (16)$$

where $\hat{W}_{ij}(t;s) = Y_{ij}(t;s) \exp\{[\hat{\Lambda}_i^E(s+t) - \hat{\Lambda}_i^E(s)]I(T_i^E > s)\}$, and with

$$\bar{\mathbf{Z}}_{\star}(u;s) = \frac{\sum_{\ell=1}^n m_{\ell j}(s) \hat{W}_{\ell}(u;s) \mathbf{Z}_{\ell}^*(s) \exp\{\beta_{\star\psi}^T \mathbf{Z}_{\ell}^*(s)\}}{\sum_{\ell=1}^n m_{\ell j}(s) \hat{W}_{\ell}(u;s) \exp\{\beta_{\star\psi}^T \mathbf{Z}_{\ell}^*(s)\}}.$$

Since $U(\beta)$ from (16) behaves asymptotically like a zero-mean estimating function, the solution to $U(\beta) = \mathbf{0}$, denoted by $\hat{\beta}_{\star\psi}$ should yield a consistent estimator of $\beta_{\star\psi}$.

2.4 Variance Estimation, Asymptotic Properties, and Inference

Due to the complex nature of the data structure in this setting we investigated the use of both the robust sandwich estimator and the bootstrap estimator for variance estimation and propose the use of the bootstrap estimator (Efron, 1979). In moderate sized samples approximately 100 bootstrap samples should be used. If the normal approximation is a reasonable, 100 bootstrap samples should be testing and interval estimation. If the normal approximation is in question, then upwards of 10,000 may be necessary to estimate the 2.5th and 97.5th percentiles of the distribution for hypothesis testing. A straightforward pre-test method for determining the number of necessary bootstrap samples is described by Davidson and MacKinnon (2000).

3. Simulation Study

Simulation studies were conducted to assess the performance of the method described above. For each subject who received the experimental treatment, the subsequent (counterfactual) experience (i.e., in the absence of experimental treatment receipt) was also generated. This counterfactual experience was not used in the modeling, but was used to determine target values of $\beta_{\star}$, which are difficult to pre-specify.

First, for each subject i , a frailty variate was generated from a Gamma (1, 0.5) distribution, bounded at 2, the 90th percentile. Next, two independent Bernoulli (0.5) variates, Z_{i1} and Z_{i2} were created. Pre-treatment recurrent event experience was then generated through a frailty model with rate parameter $\gamma_i d\mu_0 \exp\{a_1 Z_{i1} + a_2 Z_{i2}\}$. The time of standard treatment receipt, T_i^S , was then generated to follow the hazard $\lambda_0^S \exp\{\beta_{s1} Z_{i1} + \beta_{s2} Z_{i2} + \beta_{s3} \log(N_i(t^-) + 1)\}$. The recurrent event times post-standard treatment were generated from rate parameter

$\gamma_i d\mu_0^s \exp\{\delta_1 Z_{i1} + \delta_2 Z_{i2}\}$. Thus, as mentioned in the first paragraph of this subsection, we initially generated the conventional therapy (pre-treatment plus post-standard treatment) recurrent event experience for all subjects. We then generated T_i^E from hazard function, $\lambda_0^E \exp\{\beta_{E1} Z_{i1} + \beta_{E2} Z_{i2} + \beta_{E3} \log(N_i(t^-) + 1)\}$. The experimental treatment was only considered to actually be received if $T_i^E < T_i^S$ and, in such cases, the post-experimental recurrent event experience then followed rate model $\gamma_i d\mu_0^E \exp\{\phi_1 Z_{i1} + \phi_2 Z_{i2}\}$.

For each data set generated, we modeled the hazard for T_i^E using the model, $\lambda_i^E(t) = \lambda_0^E(t) \exp\{\beta_{E1} Z_{i1} + \beta_{E2} Z_{i2} + \beta_{E3} \log(N_i(t^-) + 1)\}$, in order to obtain the estimated weights for IPCW. We then modeled the pre-treatment event rate using the model $d\mu_i^0(t) = \exp\{\alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \alpha_3 N_i^0(t^-)\} d\mu_0^0(t)$, in order to obtain prognostic scores. Subjects were matched if $|\log \psi_{ij}| \leq 0.1$, resulting in approximately 8 matches per stratum.

Parameters used in the simulation studies are as follows. For the pre-treatment event rate, we set $d\mu_0 = 5$, $\alpha_1 = -0.3$, and $\alpha_2 = 0.7$. For the post-standard treatment event rate, $d\mu_0^S = 1$, $\delta_1 = -0.6$, and $\delta_2 = 0.2$. For the experimental treatment hazard, $\lambda_0^E = 0.5$, $\beta_{E1} = -0.7$, $\beta_{E2} = -0.5$, and $\beta_{E3} = 0.3$. For the standard treatment hazard, $\lambda_0^S = 0.5$, $\beta_{S1} = -0.8$, $\beta_{S2} = -0.7$, and $\beta_{S3} = 0.1$. Finally, for the post-experimental treatment rate, $d\mu_0^E$ was given values of 1, 1.5, 2.5, and 3, $\phi_1 = -0.7$, and $\phi_2 = 0.7$. This resulted in values of $\beta_\star$ of -0.901, -0.457, 0.082, and 0.261. In the simulated data 63% of the sample received experimental treatment, 36% received standard treatment, and 1% received no treatment. The mean number of events was 9.46 (sd = 0.72) in the no treatment state, 18.92 (sd = 1.49) in the post-conventional treatment state, and ranged from 16.41 to 24.96 in the post-experimental treatment state. These numbers describe the simulated data but are not directly related to the different treatment effects tested, due to the timing of experimental treatment. A total of $n = 200$ subjects were simulated 250 times, with 50 bootstrap samples used for variance estimation.

The results are shown in Table 1 for the estimate of the parameter for the experimental treatment effect. Direct matching and prognostic matching gave similar bias (range 0.022-0.029 and 0.015-0.030, respectively). However, prognostic score matching resulted in empirical standard errors that were 9-10% lower. Adjusting for $\log \psi_{ij}$ further reduces bias in the estimate of the experimental treatment effect (range 0.004-0.017 across the 4 scenarios). Figure 2 shows histograms of the parameter estimates centered on their target value with normal density curves. All estimates appear to follow a normal density centered at zero, supporting the theory that $\hat{\beta}_\star$ is asymptotically normal with mean $\beta_\star$.

4. Application to Living Donor Liver Transplant

The Adult-to-Adult Living Donor Liver Transplantation Cohort Study (A2ALL) is a multicenter NIH-sponsored study investigating the post-transplant morbidity and mortality of LDLT recipients and their donors. Potential living donor transplant recipients who had a donor evaluated between January 1, 1998 and August 31, 2009 were recruited at each of the 9 A2ALL centers beginning in the third quarter of 2004 and followed through August 31, 2010. These potential recipients may have ultimately received an LDLT, a DDLT, or

neither. Clinical data, including laboratory information, hospitalizations, and complications were collected based on a common protocol, and supplemented with data 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; these data are submitted by the members of OPTN and have been described elsewhere. The Health Resources and Services Administration (US Department of Health and Human Services) provides oversight for the activities of the OPTN and SRTR contractors. Data collection for pre- and post-transplant hospitalizations included date of admission and discharge, reason for hospitalization, and discharge diagnosis. Subjects could enroll either before or after transplant, and information prior to enrollment was collected via chart review.

LDLT is a technically complicated surgery. Since we essentially wish to evaluate whether LDLT results in a higher post-transplant hospitalization rate than conventional therapy, our objective is consistent with a non-inferiority test. The comparison of interest is between LDLT recipients and their counterparts who remain on the waiting list and potentially receive a DDLT. The time origin was date of first donor evaluation. This analysis includes 1467 liver transplant candidates, of which 714 went on to receive an LDLT, 455 received a DDLT, and 298 remained untransplanted.

The pre-treatment model for the event rate was adjusted for age, race, recipient diagnosis, Model for End Stage Liver Disease (MELD), event history (previous number of hospital admissions), and the following characteristics measured at the time of donor evaluation: hospitalization status, ascites, mechanical ventilation, dialysis status, and transplant center. Subjects with $|\log \psi_{ij}(s)| > 0.025$, who were untransplanted at the time of LDLT were matched. This resulted in a median of 20 matches (range 1-71, IQR 11-34). Of the 714 LDLTs, 692 (97%) were matched to at least one “control” subject. The mean $\log \hat{\psi}_{ij}(s)$ was 0.000033 (SD = 0.014), indicating that the index patients and their matched counterparts were very similar at the time of matching, with respect to prognosis.

Both traditional time-dependent models and models using the proposed method were fitted. The time-dependent model was adjusted for age, ethnicity, diagnosis, MELD, transplant center, diabetes and ascites at the time of donor evaluation, and event history defined as above. In the model using the proposed method 100 bootstrap samples were generated to estimate the variance of parameter estimates. Table 2 shows the results of two models, one using only a main effect for LDLT and another using an interaction between LDLT and MELD score, each fit for both methods mentioned above (a total of four models). In the time-dependent version of Model 1, LDLT recipients have a significantly higher rate of hospitalizations compared to the combination of waiting list patients and DDLT recipients (rate ratio (RR) = 1.22, $p < 0.0001$), consistent with Merion et al, 2010. However, using the proposed methods, there is no difference in hospitalization rates between subjects who receive LDLT and those receive conventional therapy (RR = 0.986, $p = 0.8245$). We also fit models with interactions between LDLT and MELD (Model 2). In the time-dependent model lower MELD subjects receiving LDLT had significantly higher hospitalization rates, with a rate ratio of 1.606 for subjects with MELD 6-11 and 1.351 for subjects with MELD 12-15 ($p < 0.0001$ in both cases). Subjects receiving LDLT in higher MELD categories did not have significantly different hospitalization rates relative to subjects waiting for DDLT.

Similar to the main effect model, the higher hospitalization rates in low MELD subjects receiving LDLT are not seen when the proposed method is used. Rate ratios range from 0.915-1.152 with p-values ranging from 0.3628-0.8978.

One potential reason for the differing results between the two methods is that in the proposed model patients are matched on pre-transplant event rate on the day of the index transplant, so the method compares the LDLT patient to patients with a similar pre-transplant event history. By contrast, in the traditional time-dependent analysis, pre- and post-transplant event rates are averaged over in both the LDLT and non-LDLT groups; patients in the comparison group may have differing event rates at the time a given subject receives an LDLT. To explore this further we investigated interactions with time and LDLT using the proposed method to explore potential differences in the rate ratios over time. Figure 3 shows that there is a sharp increase in the hospitalization rate immediately after LDLT. However, the RR declines quickly, to the point where it is significantly below 1 for a period, then continues to rise for the remainder of follow-up. The estimated RR actually rises above 1 (although non-significantly) towards the end of the follow-up period. The pronounced and steady rise in the RR (after the drop to significantly below 1) is due to the nature of the comparison groups. Specifically, as time progresses, a greater percentage of subjects in the conventional therapy groups are DDLT (as opposed to pre-transplant) patients. If the comparison were between experimental treatment and pre-transplant, then it is highly likely that the RR may have stabilized.

5. Discussion

Despite the wide array of methods available for modeling recurrent event data, there are relatively few such methods devoted specifically to estimating the effect of treatment. A limited number of methods have been developed to accommodate time-dependent covariates and informative censoring (e.g., Huang, Qin and Wang, 2010; Zhao et al, 2012). However, such methods apply under different data structures and assumptions. Our goal in this report was to develop methods to estimate treatment effects in a fairly complex setting: treatment is time-dependent, different forms of treatment are available, and the event rate in the absence of the treatment of interest is dependently censored.

One key aspect to this type of analysis is appropriate matching of control subjects to experimental subjects. In the method proposed prognostic scores were used to match index patients to control patients within a certain radius. It should be noted that this is slightly different than the usual prognostic score setting because all patients are observed in the untreated state for some period of time, not only the control patients, and thus all patients contribute to the event rate model used to generate the prognostic scores. An alternative would be to use k -nearest neighbor matching. However, this method appears to be less appropriate for two reasons. First, while some experimental patients may have many appropriate matches, requiring a pre-specified number of matches per experimental patient could result in large heterogeneity in some strata. Second, Abadie and Imbens (2008) have demonstrated that the bootstrap estimator is invalid in the setting of nearest neighbor matching. It could also be argued that propensity matching would be an appropriate matching method for this data structure. In that scenario the goal would be to simulate a

randomized trial by matching patients with similar probabilities of receiving experimental treatment. A time-dependent propensity score similar to that proposed by Lu (2005), may be appropriate.

A limitation to the method as described is that it does not take into account terminal events such as death. Many methods for recurrent/terminal event data have been proposed (e.g., Ghosh and Lin, 2002; Liu, Wolfe and Huang, 2004; Huang and Wang, 2004; Ye, Kalbfleisch and Schaubel, 2007; Kim et al., 2012, Kalbfleisch et al., 2013) and could be applied to the method described above. Extension of the proposed methods to incorporate terminal events is currently underway.

Acknowledgments

The authors thank the Scientific and Data Coordinating Center project area team at Arbor Research Collaborative for Health, and the investigators, coordinators, and other staff that participate in the A2ALL Study. This study was supported by the National Institute of Diabetes and Digestive and Kidney Diseases through a cooperative research agreement (U01-DK62498), and by National Institutes of Health Grant R01 DK070869. Some data reported here were supplied by the Minneapolis Medical Research Foundation as the contractor for the Scientific Registry of Transplant Recipients. 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 Scientific Registry of Transplant Recipients or the US Government.

References

- Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. *Econometrica*. 2008; 76:1537–1557.
- Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*. 1982; 10:1100–1120.
- Chen Q, Zeng D, Ibrahim JG, Akacha M, Schmidli H. Estimating time-varying effects for overdispersed recurrent events data with treatment switching. *Biometrika*. 2013; 100:339–354. [PubMed: 24465031]
- Cook RJ, Lawless JF, Lakhal-Chaieba L, Lee KA. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: Application to skeletal complications in cancer metastatic to bone. *Journal of the American Statistical Association*. 2009; 104:60–75.
- Cook, RJ.; Lawless, JF. *The Statistical Analysis of Recurrent Event Data*. New York: Wiley; 2007.
- Cox DR. Partial likelihood. *Biometrika*. 1975; 62:269–276.
- Davidson R, MacKinnon JG. Bootstrap tests: how many bootstraps? *Econometric Reviews*. 2000; 19:55–68.
- Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*. 1979; 7:1–26.
- Ghosh D, Lin DY. Marginal regression models for recurrent and terminal event. *Statistica Sinica*. 2002; 12:663–688.
- Gong Q, Schaubel DE. Partly conditional estimation of the effect of a time-dependent factor in the presence of dependent censoring. *Biometrics*. 2013; 69:338–347. [PubMed: 23635094]
- Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008; 95:481–488.
- Hansen BB. The essential role of balance tests in propensity-matched observational studies: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*. *Statistics in Medicine*. 2008; 27:2037–2049. [PubMed: 18038446]
- Huang CY, Qin J, Wang MC. Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics*. 2010; 66:39–49. [PubMed: 19459839]
- Huang CY, Wang MC. Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*. 2004; 99:1153–1165. [PubMed: 24068850]

- Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd. New York: Wiley; 2002.
- Kalbfleisch JD, Schaubel DE, Ye Y, Gong Q. An estimating function approach to the analysis of recurrent and terminal events. *Biometrics*. 2013; 69:366–374. [PubMed: 23651362]
- Kim S, Zeng D, Chambless L, Li Y. Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in the Biosciences*. 2012; 4:262–281.
- Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics*. 1995; 37:158–168.
- Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine*. 2014; 33:3488–3508. [PubMed: 24151187]
- Li Y, Schaubel DE, He K. Matching methods for obtaining survival functions to estimate the effect of a time-dependent treatment. *Statistics in Biosciences*. 2014; 6:105–126. [PubMed: 25309633]
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B*. 2000; 62:711–730.
- Liu L, Wolfe RA, Huang X. Shared frailty model for recurrent events and a terminal event. *Biometrics*. 2004; 60:747–756. [PubMed: 15339298]
- Lu B. Propensity score matching with time dependent covariates. *Biometrics*. 2005; 61:721–728. [PubMed: 16135023]
- Merion RM, Shearon TH, Berg CL, Everhart JE, et al. Hospitalization rates before and after adult-to-adult living donor or deceased donor liver transplantation. *Annals of Surgery*. 2010; 251:542–549. [PubMed: 20130466]
- Miloslavsky M, Keles S, Van Der Laan MJ, Butler S. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society, Series B*. 2004; 66:239–257.
- Pena EA, Strawderman RL, Hollander M. Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association*. 2001; 96:1299–1315.
- Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time-dependent covariates. *Journal of the American Statistical Association*. 1993; 88:811–820.
- Pepe MS, Couper D. Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association*. 1997; 92:991–998.
- Robins, JM.; Rotnitzky, A. *AIDS Epidemiology - Methodological Issues*. Birkhäuser; Boston: 1992. Recovery of information and adjustment for dependent censoring using surrogate markers; p. 297–331.
- Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*. 2000; 95:573–585.
- Schaubel DE, Wolfe RA, Sima CS, Merion RM. Estimating the effect of a time-dependent treatment by levels of an internal time-dependent covariate. *Journal of the American Statistical Association*. 2009; 104:49–59.
- Schaubel DE, Zhang M. Estimating treatment effects on the marginal recurrent event mean in the presence of a terminating event. *Lifetime Data Analysis*. 2010; 16:451–477. [PubMed: 20063183]
- Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*. 2013; 66:S84–S90. [PubMed: 23849158]
- van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*. 2007; 34:70–85.
- van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*. 2008; 14:447–463. [PubMed: 18836831]
- van Houwelingen, HC.; Putter, H. *Dynamic prediction in clinical survival analysis*. Boca Raton, FL: CRC Press Inc; 2011.

- Ye Y, Kalbfleisch JD, Schaubel DE. Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*. 2007; 63:78–87. [PubMed: 17447932]
- Zhao X, Liu L, Liu Y, Xu W. Analysis of multivariate recurrent event data with time-dependent covariates and informative censoring. *Biometrical Journal*. 2012; 54:585–599. [PubMed: 22886587]
- Zheng YY, Heagerty PJ. Partly conditional survival models for longitudinal data. *Biometrics*. 2005; 61:379–391. [PubMed: 16011684]

Figure 1.

Recurrent event process for subject i under two scenarios. In the top time line, subject i receives the experimental treatment at time $T_i^E = s$. For the bottom time line, the experimental treatment is unavailable, such that subject i receives conventional therapy; this second time line is latent beyond time s . For subject i , follow-up time $> s$ is pertinent to assessing the effect of the experimental treatment on the recurrent event rate.

Figure 2. Histogram of Parameter Estimate with Normal Density

Figure 3. Time-dependent rate ratio (RR) for LDLT and 95% confidence intervals

Table 1**Results of simulation study**

Scenario	β_*	Matching	Bias	ESE	ASE	CP
1	-0.901	Prognostic	0.030	0.130	0.158	0.980
2	-0.457		0.029	0.127	0.151	0.972
3	0.082		0.024	0.121	0.142	0.972
4	0.261		0.015	0.119	0.140	0.976
1	-0.901	Prognostic	0.017	0.129	0.158	0.976
2	-0.457	w/adjustment	0.017	0.125	0.153	0.980
3	0.082		0.012	0.120	0.144	0.984
4	0.261		0.004	0.118	0.141	0.976
1	-0.901	# of Events	0.022	0.142	0.188	0.988
2	-0.457	with Covariate	0.027	0.140	0.181	0.976
3	0.082	Adjustment	0.029	0.135	0.172	0.976
4	0.261		0.024	0.133	0.170	0.980

Table 2

A2ALL Results

		Proposed Model				Time-dependent Model			
Model		Rate Ratio	95% CI	p-value		Rate Ratio	95% CI	p-value	
1	LDLT	0.986	0.857-1.134	0.8425		1.217		<0.0001	
2	LDLT MELD 6-11	1.152	0.849-1.562	0.3628		1.606		<0.0001	
	LDLT MELD 12-15	0.962	0.752-1.230	0.7589		1.351		<0.0001	
	LDLT MELD 16-19	0.915	0.696-1.202	0.5226		1.009		0.8906	
	LDLT MELD 20-29	0.918	0.702-1.202	0.5344		1.040		0.5483	
	LDLT MELD 30-40	1.150	0.136-9.702	0.8978		1.143		0.3148	