Electronic Journal of Statistics
Vol. 9 (2015) 419–443
ISSN: 1935-7524
DOI: 10.1214/15-EJS1003

A nonparametric analysis of waiting times from a multistate model using a novel linear hazards model approach

Douglas J. Lorenz and Somnath Datta

Department of Bioinformatics and Biostatistics
School of Public Health and Information Science
University of Louisville
485 E. Gray St.
Louisville, KY 40202,
e-mail: djlore01@louisville.edu; somnath.datta@louisville.edu

Abstract: Traditional methods for the analysis of failure time data are often employed in the analysis of waiting times of transient states from multistate models. However, such methods can exhibit bias when waiting times among model states are dependent, even when censoring is random. Furthermore, right-censoring can occur prior to entry into the transient state of interest, preventing the observation of transitions from the state and providing another potential source of bias. We introduce a nonparametric linear hazards model for waiting times from multistate models, analogous to Aalen's linear hazards model for failure time data, where proper estimation can be carried out via reweighting, a method flexible enough to incorporate general forms of induced and other dependent censoring. We illustrate the approximate unbiasedness of the proposed regression coefficient estimators through a simulation study, while also demonstrating the bias arising from traditional Aalen's linear hazards model estimators obtained from correlated waiting time data. Theoretical results for the parameter estimators are provided. The reweighted estimators are used in the analysis of two data sets, to identify predictors of ambulatory recovery in a data set of spinal cord injury patients receiving activity-based rehabilitation and to identify prognostic indicators for patients receiving bone marrow transplant.

MSC 2010 subject classifications: Primary 62N02, 62N01; secondary 62G05.
Keywords and phrases: Inverse probability weighting, Aalen's linear model, multivariate survival data.

Received January 2014.

Contents

1. Introduction

The motivating example for this article is a data set of 273 patients with incomplete spinal cord injury (SCI) participating in a national activity-based rehabilitation program [1, 2]. Participating patients undergo a comprehensive assessment at enrollment, receive sessions of standardized therapy (called locomotor training), and are periodically evaluated for functional progress until program discharge. Walking speed is among the measures of function collected on these patients, and there are several clinical benchmarks – 0.44 m/s represents the minimum walking speed associated with the ability to walk in the community, 0.7 m/s separates those who require assistive walking devices from those who do not, and 1.2 m/s approximately defines the speed required to cross a street at a stoplight [3]. The achievement of these benchmarks provides an example of a multistate model, specifically, a five state model with states (1) patient unable to walk, (2) patient able to walk no faster than 0.44 m/s, (3) patient able to walk no faster than 0.7 m/s, (4) patient able to walk no faster than 1.2 m/s, (5) patient able to walk faster than 1.2 m/s (Figure 1).

Community ambulation is a frequently-cited goal of rehabilitation therapy for SCI patients with limited ambulatory capacity. Therefore, clinicians in the rehabilitation program have been interested in identifying prognostic indicators of the amount of time it takes ambulatory patients to achieve this goal.

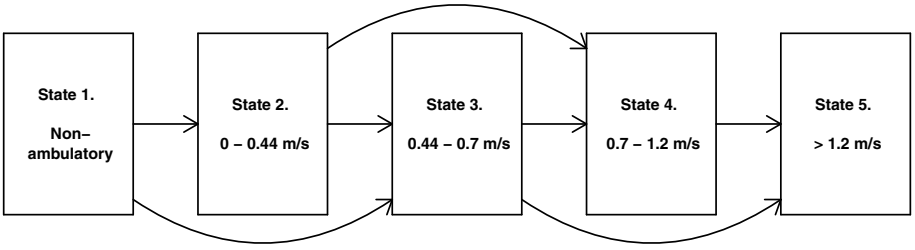

FIG 1. *Multistate model of SCI data example, with states defined by thresholds for walking speed attained by SCI patients. Only transitions observed in the data are included in the depiction of the model. Note that patients could enter the system at any state other than the absorbing state (> 1.2 m/s). In Section 4.1, we consider waiting times in the second state (0–0.44 m/s).*

In particular, it is of interest whether the intensity of therapy – which varies from patient to patient – has an impact on the time it takes a non-community ambulator to become a community ambulator. In the context of our multistate model, this goal is represented as a transition out of state 2 to any later state. The amount of time spent in state 2 prior to exit is the waiting time in state 2, also referred to as the sojourn time. A simple approach to modeling waiting times in transient model states as a function of one or more covariates may involve the use of a regression model for survival data, such as the popular Cox relative risk model or accelerated failure time model. This approach may in part be motivated by the conceptual similarity between survival times and waiting times. Both measure the time to an event – failure or death for survival times and state exit for waiting times. This approach can be flawed, particularly when observation of the multistate system is right censored. We describe these flaws in the context of the SCI example.

Rehabilitation patients with severe functional deficits – those who are non-ambulatory or non-community ambulators at enrollment – very rarely regain the ability to walk at street-crossing speed. Thus, most of these patients are right-censored prior to reaching the absorbing state of the multistate system (able to walk faster than 1.2 m/s). In particular, patients that enter the rehabilitation program unable to walk may not regain the ability to walk before discharge from the rehabilitation program. In the context of the multistate model describing the SCI data, such patients are censored while in state 1 of the model. As such, these patients have contributed observation time prior to the commencement of the event process of interest (the exit process from state 2), and it is impossible for the observer to know whether they would have entered state 2 and how much time they may have spent there. Survival data methods applied to multistate waiting times necessarily ignore these patients, which can potentially bias any inferences made, particularly in more complex multistate networks with several paths. Another drawback to the use of survival data methods for multistate waiting times is the need to assume the independence of waiting times among states, the semi-Markov property, in addition to the independence of censoring times traditionally required of survival data methods. Even in the presence of random censoring, dependence among waiting times can induce patterns of dependent censoring; a heuristic demonstration of this can be found in [4]. In practice, the semi-Markov property can be implausible for many multistate models, obviating the need for alternative methods. We return to the spinal cord injury data set in Section 4 and show how to resolve these issues following our approach.

The nonparametric linear hazard model introduced by Aalen [5, 6, 7] provides a flexible regression model for the analysis of failure time data. Like the Cox relative risk and AFT models, Aalen's linear model permits fixed and time-varying covariates both internal and external in scope. The flexibility of Aalen's linear model comes from the model coefficients being defined as functions of time rather than static quantities as in the Cox and AFT models. To our knowledge, Aalen's linear model has yet to be extended to waiting time hazards from multistate models.

In this paper, we develop a novel linear hazards model for multistate waiting times and related estimation procedure. Some regression models for sojourn times have been developed for the special case of a progressive multistate system (time ordered serial events) where the sojourn times are the gap times between events [8, 9]. In contrast, our approach is applicable to a very general multistate system under induced and other forms of dependent censoring. Further, the reweighting procedure we utilize provides a proper adjustment to the bias that can be introduced from right censoring prior to entry into the state under analysis. In Section 2, we introduce notation for multistate models, develop an estimator for the integrated coefficient functions, and suggest a martingale representation for the estimator, useful for the development of asymptotic properties. Section 3 details the results of a simulation study evaluating the present estimators against the estimators from Aalen's linear model for failure time data applied to waiting times in a simple multistate system with correlated waiting times. We illustrate the utility of our methodology by analyzing two data sets, the multistate SCI data described above and a follow-up study of individuals receiving bone marrow transplantation [10].

2. Methods

2.1. Preliminaries and the linear Hazards model for waiting times

We consider an acyclic network of J states through which n individuals progress. For individual i $(1 \leq i \leq n)$, the uncensored data consist of the entry and exit times for each state, T_{ij}^* and U_{ij}^*, respectively, for states $j \in \{1, \ldots, J\}$. Let $T_{ij}^* = \infty$ if state j is never entered and $U_{ij}^* = \infty$ if state j is never entered or never exited, for example, if state j is absorbing. The right censoring time for individual i is denoted by C_i, and we note that censoring applies to the full multistate system, (i.e.) after censoring no further transitions between states are observed. The right-censored entry and exit times for finite T_{ij}^* and U_{ij}^* are $T_{ij} = \min(T_{ij}^*, C_i)$ and $U_{ij} = \min(U_{ij}^*, C_i)$. Let $\gamma_{ij} = I[C_i \geq T_{ij}^*]$ and $\delta_{ij} = I[C_i \geq U_{ij}^*]$ be the indicators of individual i having been observed to enter and exit state j, respectively. The time of last transition is defined as $T_i^* = \max_j\{T_{ij}^* \mid T_{ij}^* < \infty\}$ in the uncensored data and $T_i = \min\{T_i^*, C_i\}$ in the censored data, and let $\delta_i = I[T_i \leq C_i^*]$ be the indicator of the last transition being observed. Let $X_{ij} = I[T_{ij}^* < \infty]$ be the indicator of eventual state j entry. Lastly, in addition to the entry and exit times and entry and exit indicators, we observe a vector of possibly time-varying covariates $\mathbf{Z}_i(t)$ for each individual.

In what follows, we will have need to refer to both calendar time, the time from which observation of the multistate system began, and waiting time, the time from entry into a given model state. Additionally, we will need to define processes and filtrations evolving in calendar time as well as those evolving in waiting time. In order to distinguish the two, we will denote calendar time as t and waiting time as w. Processes and filtrations evolving in calendar time will be represented as functions of t and those evolving in waiting time as functions of w.

The full event history for the multistate system consists of transitions observed up to time t along with any information contributed by censoring. Let $\mathcal{H}_{t,i} = \sigma\{T_{ij}^* I[T_{ij}^* \leq t]; j = 1, \ldots, J\}$ represent the event history for individual i, defined as the σ-algebra generated by transitions observed up to time t. Define the event history for all individuals to be $\mathcal{H}_t = \vee_i \mathcal{H}_{t,i}$. Let $\bar{\mathbf{Z}}_i(t)$ represent the covariate history for individual i up to time t, and let $\bar{\mathbf{Z}}(t) = \vee_i \bar{\mathbf{Z}}_i(t)$.

Processes evolving in state j waiting time will be adapted to a filtration generated by observation of the multistate system after state j entry. This will consist of exits from state j as well as covariate histories after state j entry. We will require notation for these waiting time process and filtrations. For individuals that enter state j, the waiting time in state j can be written as $w = t - T_{ij}^*$. For individual i, the history associated with the state j waiting time process is $\mathcal{H}_{w,i}^{(j)} = \sigma((U_{ij}^* - T_{ij}^*) I[U_{ij}^* - T_{ij}^* \leq w], X_{ij})$. The state j waiting time history for all individuals is $\mathcal{H}_w^{(j)} = \vee_i \mathcal{H}_{w,i}^{(j)}$. Define the value of the vector of covariates at waiting time w as

$$\mathbf{Z}_i^{(j)}(w) = \begin{cases} \mathbf{Z}_i(w + T_{ij}^*) & \text{if } X_{ij} = 1 \\ 0 & \text{if } X_{ij} = 0. \end{cases}$$

The value of $\mathbf{Z}_i^{(j)}$ at waiting time w is simply the value of $\mathbf{Z}_i$ at the calendar time t corresponding to waiting time w, which is $w + T_{ij}^*$ for individuals having entered state j. Let $\bar{\mathbf{Z}}_i^{(j)}(w)$ represent the covariate history for individual i during state j occupation, and let $\bar{\mathbf{Z}}^{(j)}(w) = \vee_i \bar{\mathbf{Z}}_i^{(j)}(w)$ be the state j covariate history for all individuals. These histories of transitions and covariate values, both for the full multistate system and state j waiting times, will be employed later to define filtrations for martingales associated with our proposed estimator.

We define a linear model for the waiting time hazard in transient state j, $U_{ij}^* - T_{ij}^*$ denoting the state j waiting time in the uncensored data and $U_{ij} - T_{ij}$ in the censored data, which we assume to be continuous. Let $\lambda_j(w|\cdot) = \lim_{\mathrm{d}w \to 0} P[U_{ij}^* - T_{ij}^* \in [w, w + \mathrm{d}w) \mid U_{ij}^* - T_{ij}^* \geq w, X_{ij} = 1, \cdot]/\mathrm{d}w$ be the hazard rate function for exits from state j of the multistate model, and let $\Lambda_j(w|\cdot) = \int_0^w \lambda_j(v|\cdot)\mathrm{d}v$ be the cumulative hazard function. Our linear hazards model for waiting times defines $\lambda_j(w)$ as a linear function of the covariates $\mathbf{Z}_i^{(j)}(w)$:

$$\lambda_j\left(w \mid \bar{\mathbf{Z}}_i^{(j)}(w), X_{ij} = 1\right) = \beta_{0j}(w) + \sum_{m=1}^{p} \beta_{mj}(w) Z_{im}^{(j)}(w), \tag{1}$$

where the $\beta_{mj}(w)$ are the coefficient functions. We note that values of the time-varying covariates Z_{im} impact the state j hazard only at times during which an individual is in state j, as implied by the notation $Z_{im}^{(j)}(w)$. Values taken by time-varying covariates prior to state j entry have no effect on the state j exit hazard.

The set of covariates in (1) can include so-called internal covariates, functions of the observed transitions through the multi-state system, and we note one important consequence of this. Since the waiting time linear model (1) is marginal

with respect to the waiting time in state j but conditional on the covariates $Z_{im}(w)$, a linear model that includes the time of entry into state j (T_{ij}) or functions thereof as one of the covariates can be analyzed with the standard Aalen model. In the introduction, we noted that dependencies among waiting times in successive states can induce dependent censoring in the analysis of waiting times, potentially biasing traditional estimators for survival data like the covariate estimators for Aalen's linear model. However, if the state j entry time is included as a covariate in waiting time model (1), then the dependencies among waiting times are effectively "conditioned out" of the model. In other words, the inclusion of the state j waiting time as a model covariate effectively fixes it, thereby removing any waiting time dependencies that may induce dependent censoring and bias survival data methods. In this special case, the traditional coefficient estimators for Aalen's linear model can be employed. In general, the traditional estimators will fail while our proposed estimators remain valid, as will be shown in our simulation study in Section 3.

We estimate the integrated coefficient vector $\mathbf{B}_j(w) = (B_{0j}(w), \ldots, B_{pj}(w))$ for model (1), where $B_{mj}(w) = \int_0^w \beta_{mj}(v)\mathrm{d}v$. The integrated coefficients $\mathbf{B}_j(w)$ represent a cumulative sum of the regression coefficients over time and are a cumulative measure of the impact of a covariate on the hazard of state j exit. We begin with uncensored data, for which Aalen's estimator for failure time data [5, 6, 7] can be directly applied. To define the uncensored data estimator, let $N^*_{ij}(w) = I[U^*_{ij} - T^*_{ij} \leq w, X_{ij} = 1]$ indicate whether individual i has exited state j by time w after state j entry (given that state j has been entered), and define the vector $\mathbf{N}^*_j(w) = (N^*_{1j}(w), \ldots, N^*_{nj}(w))$. Let $Y^*_{ij}(w) = I[U^*_{ij} - T^*_{ij} \geq w, X_{ij} = 1]$ be the indicator that individual i has yet to leave state j just prior to time w after state j entry, given that state j has been entered. Define the $n \times (p+1)$ "at-risk" covariate matrix $\mathbf{Y}^*_j(w)$ to have i^{th} row $Y^*_{ij}(w) \cdot (1, Z^{(j)}_{i1}(w), \ldots, Z^{(j)}_{ip}(w))$. If individual i has yet to leave state j by time w after state j entry, then the i^{th} row of $\mathbf{Y}^*_j(w)$ contains the covariate vector for individual i at time w after state j entry. Otherwise, the i^{th} row of $\mathbf{Y}^*_j(w)$ is a vector of zeros. Aalen's estimator of $\mathbf{B}_j(w)$, derived as the solution to a martingale estimating equation (cf. [11], Chapter VII for a detailed derivation) is

$$\mathbf{B}^*_j(w) = \int_0^w J^*_j(v)\mathbf{Y}^{*-}_j(v)\mathrm{d}\mathbf{N}^*_j(v), \tag{2}$$

where $\mathbf{Y}^{*-}_j(v)$ is a generalized inverse of $\mathbf{Y}^*_j(v)$ and $J^*_j(v) = I[\mathrm{rank}(\mathbf{Y}^*_j(v)) = p+1]$. From the martingale estimating equation follows a martingale representation for estimator (2), $(\mathbf{B}^*_j - \mathbf{B}_j)(w) = \int_0^w J^*_j(v)\mathbf{Y}^{*-}_j(v)\mathrm{d}\mathbf{M}^*_j(v)$, where $\mathbf{M}^*_j(v)$ is the vector with components $M^*_{ij}(v) = N^*_{ij}(v) - \int_0^w \lambda_j(v)Y^*_{ij}(v)\mathrm{d}v$. For uncensored data, standard results [11] for Aalen's linear model give that $\mathbf{M}^*_j(w)$ is a martingale with respect to the filtration $\mathcal{F}^{(j)}_w = \sigma\{\bar{\mathbf{Z}}^{(j)}(w), \mathcal{H}^{(j)}_w\}$, the σ-algebra generated by state j exits and the state j covariate process. The predictable

variation process associated with this martingale is

$$\left\langle \mathbf{B}_j^* - \mathbf{B}_j \right\rangle (w) = \int_0^t J_j^*(v)\mathbf{Y}_j^{*-}(v)\mathbf{diag}\left\{\lambda_j(v)\mathrm{d}v\right\}\mathbf{Y}_j^{*-}(v)^{\mathrm{T}}. \tag{3}$$

Asymptotic results and practical applications for this estimator require the choice of a specific generalized inverse for $\mathbf{Y}_j^*(w)$. A popular choice is based on a least squares principle by letting $\mathbf{Y}_j^{*-}(w) = (\mathbf{Y}_j^*(w)^{\mathrm{T}}\mathbf{Y}_j^*(w))^{-1}\mathbf{Y}_j^*(w)^{\mathrm{T}}$, although other choices are possible. Weak convergence to a Gaussian martingale with variance equal to the expectation of (3) follows under regularity conditions on the matrix $\mathbf{Y}_j^*(w)$ guaranteeing applicability of the martingale central limit theorem (cf. [11], Thm. VIII.4.1).

2.2. Inverse probability of censoring weighted estimator for integrated regression coefficient functions

The estimator (2) can be adapted to censored data by replacing the individual level counting processes $N_{ij}^*(w)$ and $Y_{ij}^*(w)$ with censored data equivalents $N_{ij}(w) = I[U_{ij} - T_{ij} \leq w, \delta_{ij} = 1]$ and $Y_{ij}(w) = I[U_{ij} - T_{ij} \geq w, \gamma_{ij} = 1]$, producing the estimator

$$\widetilde{\mathbf{B}}_j(w) = \int_0^w J_j(v)\mathbf{Y}_j^-(v)\mathrm{d}\mathbf{N}_j(v), \tag{4}$$

where $J_j(v) = I[\mathrm{rank}(\mathbf{Y}_j(v)) = p + 1]$. As noted in Section 1, a semi-Markov assumption and an independent censoring assumption [11] are required for this estimator to be valid. To relax this requirement, our proposed estimator operates by weighting the basic counting processes composing uncensored data estimators like (2) with the inverse probability of censoring. Thus, in order to develop weighted estimators for the integrated regression coefficient functions, we first require a model for the censoring hazard.

Let $\lambda^c(t|\cdot)$ denote the censoring hazard. Following previous work on weighted estimation for failure time and multistate data [12, 13, 14], we assume that knowledge of future transition times does not impact the hazard of censoring, a condition formally stated as

$$\lambda^c\left(t \mid \bar{\mathbf{Z}}_i(t), \mathcal{H}_{\infty,i}\right) = \lambda^c\left(t \mid \bar{\mathbf{Z}}_i(t), \mathcal{H}_{t-,i}\right). \tag{5}$$

For notational convenience we write $\lambda_i^c(t)$ for $\lambda^c(t \mid \bar{\mathbf{Z}}_i(t), \mathcal{H}_{t-,i})$. Let $\Lambda_i^c(t) = \int_0^t \lambda_i^c(s)\mathrm{d}s$ be the cumulative hazard of censoring and define the associated product integral $K_i(t) = \prod_{s\leq t}[1 - \mathrm{d}\Lambda_i^c(s)]$, which provides the probability of censoring for individual at time t.

Returning to the analysis of waiting times, define the weighted counting processes $\overline{N}_{ij}(w) = I[U_{ij} - T_{ij} \leq w, \delta_{ij} = 1]/K_i(U_{ij}-)$ and $\overline{Y}_{ij}(w) = I[U_{ij} - T_{ij} \geq w, \gamma_{ij} = 1]/K_i(T_{ij} + w-)$ and let the vector $\overline{\mathbf{N}}_j(w)$ and the matrix $\overline{\mathbf{Y}}_j(w)$ be defined by replacing $N_{ij}^*(w)$ and $Y_{ij}^*(w)$ with $\overline{N}_{ij}(w)$ and $\overline{Y}_{ij}(w)$ throughout. Inference based on these reweighted counting processes is asymptotically

equivalent to that based on the uncensored data counting processes based on the previously proven [15] expectation equalities $E[\overline{N}_{ij}(w)] = E[N^*_{ij}(w)]$ and $E[\overline{Y}_{ij}(w)] = E[Y^*_{ij}(w)]$. The weighted processes $\overline{N}_{ij}(w)$ and $\overline{Y}_{ij}(w)$ in essence estimate the uncensored data processes, which are unobservable when right censoring is present.

The processes $\overline{N}_{ij}(w)$ and $\overline{Y}_{ij}(w)$ are not of practical use, being based on the generally unknown functions $K_i(t)$. We obtain an estimate of $K_i(t)$ by again employing Aalen's linear hazard model, and define the model for the censoring hazard as $\lambda^c_i(t) = \beta^c_0(t) + \sum_{m=1}^q \beta^c_m(t) Z_{im}(t)$. In contrast to the state j waiting times linear hazards model (1), the covariates $Z_{im}(t)$ are used directly since the censoring process evolves in calendar time. Further, we note that the covariates used to model the censoring hazard are subsets of the full covariate history $\overline{\mathbf{Z}}_i(t)$, but may be distinct from the covariates used to model the waiting time hazard in (1). To define the estimator of the vector of integrated regression coefficients $\mathbf{B}^c(t) = (B^c_0(t), \ldots, B^c_q(t))$ where $B^c_m(t) = \int_0^t \beta^c_m(s)\mathrm{d}s$, we introduce the required counting processes for censoring. Let $N^c_i(t) = I[C_i \le t, C_i \le T^*_i]$ be the indicator of individual i having been censored by time t and let $Y^c_i(t) = I[T_i \ge t]$ be the indicator of individual being at risk of censoring just before time t. Define the vector $\mathbf{N}^c(t) = (N^c_1(t), \ldots, N^c_n(t))$ and the matrix $\mathbf{Y}^c(t)$ with i^{th} row $Y^c_i(t) \cdot (1, Z_{i1}(t), \ldots, Z_{iq}(t))$. The estimator of $\mathbf{B}^c(t)$ is defined in similar fashion as the estimator $\mathbf{B}^*_j(t)$ for uncensored data:

$$\widehat{\mathbf{B}}^c(t) = \int_0^t J^c(s)\mathbf{Y}^{c-}(s)\mathrm{d}\mathbf{N}^c(s), \tag{6}$$

where $J^c(s) = I[\mathrm{rank}(\mathbf{Y}^c(s)) = q + 1]$. We again select the generalized inverse $\mathbf{Y}^{c-}(s) = (\mathbf{Y}^c(s)^{\mathrm{T}}\mathbf{Y}^c(s))^{-1}\mathbf{Y}^c(s)^{\mathrm{T}}$ for practical application of (6). Let $\widehat{\Lambda}^c_i(t) = \int_0^t \mathbf{Z}^{\mathrm{T}}_i(s)\mathrm{d}\widehat{\mathbf{B}}^c(s)$ and $\widehat{K}_i(t) = \prod_{s \le t}(1 - \mathrm{d}\widehat{\Lambda}^c_i(s))$.

Using the estimated hazard of censoring, we can now define a weighted estimator for the integrated regression coefficients for the waiting time model. Define the data-based weighted counting processes $\widehat{N}_{ij}(w) = I[U_{ij} - T_{ij} \le w, \delta_{ij} = 1]/\widehat{K}_i(U_{ij}-)$ and $\widehat{Y}_{ij}(w) = I[U_{ij} - T_{ij} \ge w, \gamma_{ij} = 1]/\widehat{K}_i(T_{ij} + w-)$, the vector $\widehat{\mathbf{N}}_j(w) = (\widehat{N}_{1j}(w), \ldots \widehat{N}_{nj}(w))$, and the matrix $\widehat{\mathbf{Y}}_j(w)$ with i^{th} row $\widehat{Y}_{ij}(w) \cdot (1, Z^{(j)}_{i1}(w), \ldots, Z^{(j)}_{ip}(w))$. The weighted estimator of $\mathbf{B}_j(t)$ is given by replacing the uncensored data counting processes with the weighted equivalents:

$$\widehat{\mathbf{B}}_j(t) = \int_0^t \widehat{J}_j(s)\widehat{\mathbf{Y}}^-_j(s)\mathrm{d}\widehat{\mathbf{N}}_j(s), \tag{7}$$

where $\widehat{J}_j(s) = I[\mathrm{rank}(\widehat{\mathbf{Y}}_j(s)) = p + 1]$, and we again use the least squares generalized inverse $\widehat{\mathbf{Y}}^-_j(t) = (\widehat{\mathbf{Y}}^{\mathrm{T}}_j(t)\widehat{\mathbf{Y}}_j(t))^{-1}\widehat{\mathbf{Y}}^{\mathrm{T}}_j(t)$. In Appendix A, we derive a martingale representation for the coefficient estimator (7). This representation results from the decomposition of $(\widehat{\mathbf{B}}_j - \mathbf{B}_j)(t)$ into two asymptotically independent martingales, defined with respect to different filtrations; the first a stochastic integral of a predictable process with respect to the martingale associated with state j exits (a state j waiting time martingale), and the second

a stochastic integral of a predictable process with respect to the martingale for censoring events (a calendar time martingale). This decomposition has previously been used to establish martingale representations for weighted estimators for survival data [15], waiting time distributions [14], and log rank tests for waiting times [4].

Given the martingale representation and suitable regularity conditions, weak convergence to a Gaussian limit can be expected via the martingale central limit theorem. However, the martingale representation and its associated predictable variation process and variance-covariance matrix are complex, making the technical conditions for the martingale central limit theorem difficult to characterize and complicating variance computations made on real data. We thus recommend use of the bootstrap to generate variance estimates, which is in part validated by the martingale representation for (7).

3. Simulation study

We conducted a simulation study on a simple multistate model to evaluate the validity of our estimator $\widehat{\mathbf{B}}_j(t)$ and examine the performance of the unweighted estimator $\widetilde{\mathbf{B}}_j(t)$ defined in Section 2.2. We considered a multistate model with a single root node (state 0), from which individuals could progress to a transient state (state 1) or absorbing state (state 2). Individuals progressing to state 1 from state 0 could then transition to a second absorbing state (state 3). We selected this structure as it represents the simplest acyclic network exhibiting a transient state (state 1) as well as an alternative absorbing state (state 2) precluding entry into the transient state. Data were simulated for four models of the waiting time hazard in the transient state (state 1), where parameters governing the bivariate distribution of waiting times in the root and transient states (states 0 and 1) were varied:

1. *One-Factor Model*. The logarithm of waiting times in states 0 and 1 were simulated from the bivariate normal distribution with means $k/4$ for group k, $k = 1, \ldots, 4$, marginal variances of 1, and covariance equal to ρ, where ρ was set equal to -0.5, 0, and 0.5 for negatively correlated, uncorrelated, and positively correlated waiting times. The state of entry from state 0 was simulated via a Bernoulli random variable with $p = 0.75$. If this random variable took value 1, the individual entered transient state 1 from state 0; otherwise the individual entered absorbing state 2. The per group sample size was 500. Censoring times were generated from the Weibull distribution with scale parameter 5 and shape parameter $2 * (5 - k)/k$ for group k, so that censoring depended on the covariate in question and varied by sample. The linear hazard model for state 1 waiting times under this design is

$$\lambda_{i1}(w) = \beta_{11}(w)Z_{i1} + \beta_{21}(w)Z_{i2} + \beta_{31}(w)Z_{i3} + \beta_{41}(w)Z_{i4},$$

where the Z_{ik} are indicator functions for group k membership and β_{k1} the coefficient functions for group k corresponding to the lognormal hazard associated with group k.

2. *One-Factor Model with Time-Dependent Covariates.* A random "switching" time was added to the first model, a time at which individuals could switch groups, $1 \leftrightarrow 4$ and $2 \leftrightarrow 3$, and subsequently experience a new hazard of state 1 exit. Switching times were generated from the uniform distribution on $(0, 5)$ and represented the calendar time of a group switch, (i.e.) an individual could switch groups before or after state 1 entry, or not at all. The linear hazard model for this design is

$$\lambda_{i1}(w) = \beta_{11}(w)Z_{i1}(w) + \beta_{21}(w)Z_{i2}(w) + \beta_{31}(w)Z_{i3}(w) + \beta_{41}(w)Z_{i4}(w)$$

   Note that the covariates $Z_{ik}(w)$ are now time-varying to account for the group switch. The definition and interpretation of all model terms are the same as in the first model, with the covariates $Z_{ik}(w)$ now reflecting group membership at waiting time w.
3. *Simple Regression.* State 0 and 1 waiting times were generated for $n = 2000$ individuals as correlated exponential variates with rate parameters 0.25 and $0.25 + 0.05Z$, where Z was a continuous covariate generated from the uniform distribution on $(-4, 4)$. The correlation between state 0 and 1 waiting times was -0.5, 0, and 0.5. Censoring times were generated from the Weibull distribution with shape parameter 2 and scale parameter $2Z$. The state of entry from the root node was simulated via a Bernoulli random variable, identical to simulation model 1. The linear hazard model for state 1 waiting times is

$$\lambda_{i1}(w) = \beta_{01}(w) + \beta_{11}(w)Z_{i1}.$$

For each simulation model, the linear model for the censoring hazard included the "correct" covariate, indicators denoting group status for simulation models 1 and 2 and the continuous covariate Z in simulation model 3. Additionally, a vector of time-varying indicators denoting state occupation at calendar time t for each model state was included in the censoring hazard model, defined as $(S_{i0}(t)\ S_{i1}(t)\ S_{i2}(t))^{\mathrm{T}}$ where $S_{ij}(t)$ takes value 1 if individual i is in state j and value 0 otherwise. To evaluate the effect of misspecification of the censoring hazard model on $\widehat{\mathbf{B}}_1(w)$, we calculated $\widehat{\mathbf{B}}_1(w)$ for simulation model 3 using two alternative censoring hazard estimators – (1) a no-covariate model for which the weights $\hat{K}_i(t)$ were simply the Kaplan-Meier estimator for the censoring hazard and (2) a model including the vector of state occupation indicators but excluding the continuous covariate Z.

The estimator $\widehat{\mathbf{B}}_1(w)$ and its variance were calculated for each of 10,000 Monte Carlo replications. The estimated coverage of the asymptotic confidence interval at several confidence levels was calculated for each coefficient as the proportion of the 10,000 replications in which the true value of the coefficient resided in the confidence interval at waiting time $w = 2.5$ (arbitrarily selected). Identical calculations were made for the unweighted estimator $\widetilde{\mathbf{B}}_1(w)$. Simulation code was written and executed in the open source R software environment [16].

The empirical coverage of the asymptotic confidence interval for the weighted estimator corresponded well with the nominal coverage rate for confidence levels

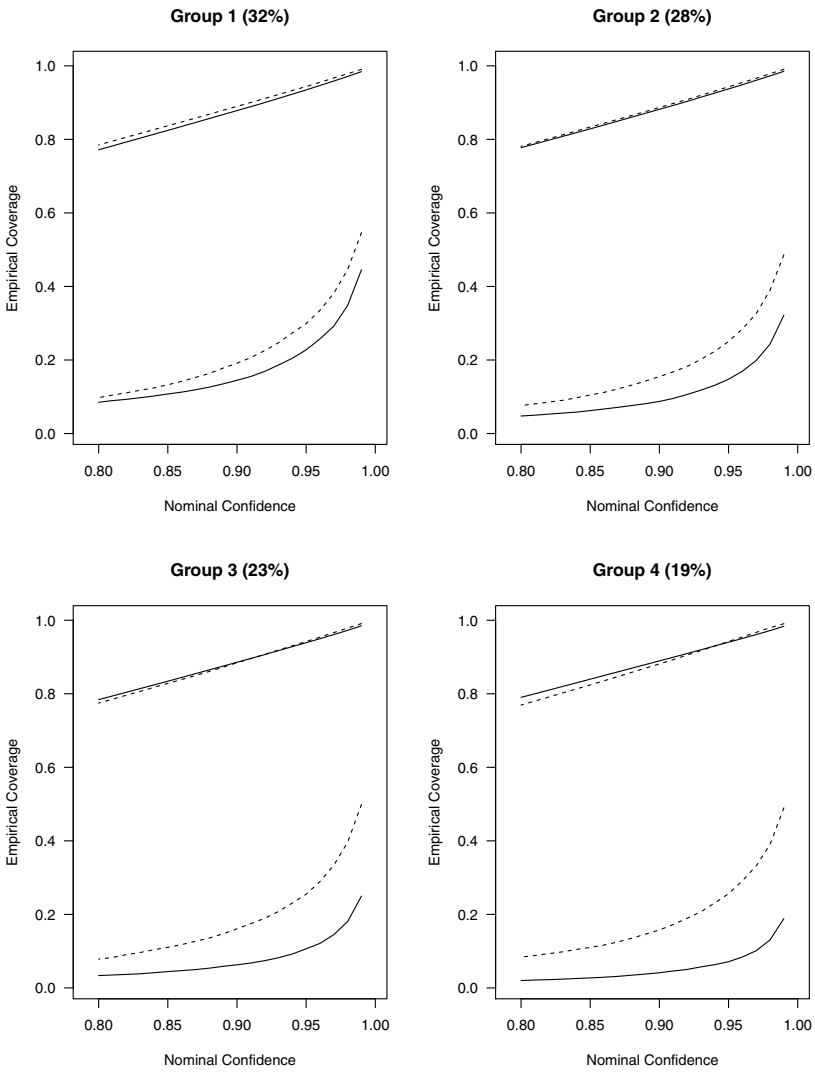

FIG 2. *P-P plot of confidence interval coverage of the weighted and unweighted integrated coefficient estimators from simulation model 2 for negatively (solid line) and positively (dashed line) correlated state 0 and 1 waiting times (calculated at time* $w = 2.5$*). Lines for the weighted estimator are in the top of each panel, exhibiting approximately correct coverage. Lines for the unweighted estimator are in the bottom of each panel, exhibiting undercoverage. Values in parentheses represent the proportion the sample that were observed to transition out of state 1, (i.e.) were uncensored and did not transition to state 2 from state 0.*

ranging from 0.80 to 0.99 in simulation model 2 (Figure 2). Notably, empirical coverage rates were on target for each level of correlation between state 0 and 1 waiting times. Heavier censoring resulted in some disparity from nominal confidence levels, as coverage rates for groups 3 and 4 were farther from the nominal

TABLE 1

Results of simulation model 3. For each waiting time correlation (Corr.), the empirical biases and standard errors (in parentheses) for the unweighted estimator ($\widetilde{\mathbf{B}}_1$) and the weighted estimator ($\hat{\mathbf{B}}_1$) are provided for waiting time $w = 2.5$. Weights were generated using two misspecified models (K-M and State Occ.) and a correctly specified model (Correct)

Parameter	Corr.	$\widetilde{\mathbf{B}}_1(2.5)$	$\hat{\mathbf{B}}_1(2.5)$		
			K-M	State Occ.	Correct
Intercept (B_{01})	−0.5	−0.47 (0.14)	−0.04 (0.27)	−0.02 (0.25)	−0.01 (0.26)
	0.0	0.00 (0.19)	−0.01 (0.21)	0.00 (0.25)	0.00 (0.26)
	0.5	0.74 (0.29)	0.11 (0.26)	0.06 (0.27)	0.02 (0.24)
Slope (B_{11})	−0.5	0.08 (0.05)	0.05 (0.11)	0.02 (0.13)	0.00 (0.12)
	0.0	0.00 (0.07)	0.01 (0.15)	0.00 (0.12)	0.00 (0.13)
	0.5	−0.13 (0.10)	−0.11 (0.13)	−0.04 (0.11)	−0.01 (0.11)

levels. The unweighted estimator $\widetilde{\mathbf{B}}_1(w)$ was substantially biased when state 0 and 1 waiting times were negatively and positively correlated. In simulation model 2, the unweighted estimator underestimated the true integrated coefficient values at $w = 2.5$ by 15% to 22% when waiting times were negatively correlated, and overestimated by 22% to 24% when positively correlated. Due to this bias, the empirical coverage of the asymptotic confidence intervals for the unweighted estimator did not exceed 68% for any group at any of the nominal confidence levels tested. When waiting times were uncorrelated, $\widetilde{\mathbf{B}}_1(w)$ was unbiased and the empirical coverage of the asymptotic confidence interval corresponded well with nominal rates. Further, the estimated variance of the unweighted estimator was substantially lower than that of the weighted estimator. Similar phenomena were observed under simulation models 1 and 3, the results of which can be found in the supplemental materials associated with this manuscript [17].

Table 1 reports the effect of model misspecification on the bias of the weighted estimators. The weighted estimator performed worst in terms of bias under the no-covariate censoring hazard model, although the amount of bias was substantially smaller than that exhibited by the unweighted estimator, particularly for the intercept term. When time-varying state occupation indicators were added to the censoring hazard model, the weighted estimator continued to exhibit bias but to a much lesser magnitude than the no-covariate censoring hazard model. The model with correct censoring hazard specification exhibited almost no bias.

4. Analysis of real data

We provide analyses of two multistate data sets to demonstrate the practical application of our linear hazards model for waiting times: the novel spinal cord injury (SCI) data set [2] noted in the Introduction and a data set tracking outcomes for individuals that received bone marrow transplant (BMT) [10]. The SCI multistate network was simple in structure – a five-state tracking model – but lacked a true root node, as patients could enter the system in any non-absorbing state. Further, it was possible for patients to skip a state in a tran-

sition, since functional outcomes were measured only at pre-defined intervals (approximately every 20 treatment sessions) and the states are defined according to discretization of a continuous outcome. For example, a patient may walk at 0.43 m/s at a given session and reside in state 2 and then walk at 0.71 m/s at the next session – a transition to state 4. The BMT network was more complex, but all patients entered the system at a root node at calendar time $t = 0$.

4.1. Spinal cord injury data

We examined waiting times in state 2, the time until the next speed benchmark for patients able to walk no faster than 0.44 m/s. We modeled the hazard of exit from state 2 as a function of 8 covariates: (1) patient age at enrollment, (2) sex, (3) time from spinal cord injury to enrollment, (4) lower motor score from the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI) exam, (5) American Spinal Injury Association Impairment Scale (categorized as C or D), (6) neurological level of injury (cervical or thoracic), (7) state of entry into the multistate system (state 1 or 2), and (8) treatment intensity, measured as the cumulative number of training session received divided by the number of days enrolled in the program. These covariates were selected based on known or suspected influence on functional recovery; patients advanced in age, female patients, farther removed from their SCI at enrollment, with more severe injuries (AIS C), with injuries higher up the spinal cord (cervical), with reduced lower motor function (lower motor score), and that entered the program non-ambulatory (in state 1) were known or suspected to recover more slowly, if at all. The covariate of primary interest was treatment intensity, defined as the ratio of treatment session received to days elapsed between functional evaluations. The investigators were particularly interested in determining whether higher intensity therapy resulted in a more rapid progression beyond non-community ambulation, (i.e.) a more rapid exit from state 2. The linear hazard model for state 2 waiting times was $\lambda_{i2}(w) = \beta_{02}(w)+\beta_{12}(w)Z_{i1}+\beta_{22}(w)Z_{i2}+\beta_{32}(w)Z_{i3}+\beta_{42}(w)Z_{i4}+\beta_{52}(w)Z_{i5}+\beta_{62}(w)Z_{i6}+\beta_{72}(w)Z_{i7}+\beta_{82}(w)Z_{i8}(w)$, where the first subscript in each $\beta_{\cdot 2}(w)$ corresponds to the above list of covariates.

In the SCI data, 122 patients entered state 2, of which 68 were censored in state 2, 41 entered state 3, and 13 entered state 4 directly. Seventy-six of the 122 patients entered the multistate system in state 2 and the remaining 46 came from state 1, (i.e.) walked for the first time after enrollment in the NRN. A full description of observed transitions for this model can be found in the supplemental materials associated with this manuscript [17]. Inverse probability of censoring weights were calculated from a linear hazard model for censoring that included the above eight covariates as well as patient race, Berg Balance Scale score at time t (a measure of postural balance assessed at every patient evaluation), fastest walking speed from enrollment to time t, and patient classification of functional status (1, 2, or 3) by the Neuromuscular Recovery Scale [18].

Four of the covariates – time since spinal cord injury, lower motor score, state of model entry, and treatment intensity – were associated with the hazard of

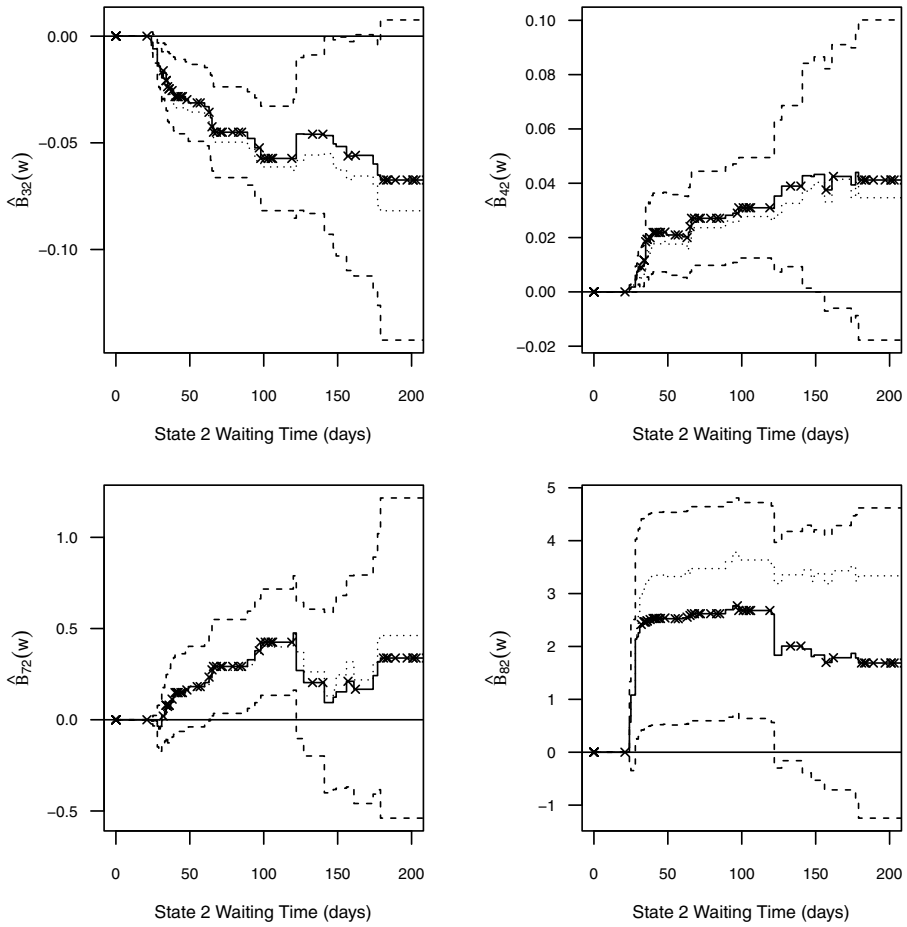

FIG 3. *Integrated coefficient functions from linear model fit to spinal cord injury data, representing regression effect of time since spinal cord injury ($\widehat{B}_{32}$), lower motor score ($\widehat{B}_{42}$), state 2 as point of entry (relative to state 1, $\widehat{B}_{72}$), and treatment intensity ($\widehat{B}_{82}$). Solid lines plot the weighted estimators, dashed lines the asymptotic 95% CI, and dotted lines the unweighted estimator. Points plotted on the weighted estimates mark censored observations.*

exit from state 2 (Figure 3). We briefly note that no patient was evaluated for progress before 20 days of enrollment had passed, thus the integrated coefficient function estimates were zero for at least the first twenty days. Bootstrapped pointwise 95% confidence intervals for each of these covariates excluded zero for some interval of state 2 waiting time, although zero was contained in all of the confidence intervals for waiting times exceeding 150 days, times for which the at-risk set was small. We also noted that the weighted estimator for the treatment intensity covariate ($\hat{B}_{82}$) differed substantially from the unweighted estimator ($\tilde{B}_{82}$).

The observed associations were clinically reasonable. Patients farther removed from their injury at enrollment tend to recover ambulatory function less rapidly, and the negative coefficient $\widehat{B}_{32}$ indicates that the hazard of progressing to the next speed benchmark was lower for patients farther removed from injury. Lower motor scores loosely describe motor function in the lower extremities – the higher the motor score, the more control a patient has over lower motor function. The positive coefficient $\widehat{B}_{42}$ indicated that patients with higher motor scores progressed more rapidly to the next speed benchmark. Patients that entered the treatment program already able to ambulate (entered into state 2) tended to advance to the next speed benchmark more quickly than those who were unable to ambulate at enrollment, as $\widehat{B}_{72}$ was positive. Higher intensity treatment was associated with more rapid progression to the next speed benchmark ($\widehat{B}_{82}$). The effect of treatment intensity appeared to be among the most important factors determining the rate of exit from state 2, as the integrated coefficient function rapidly attained a significant value prior to day 50 of waiting time in state 2. The impact of high intensity treatment stabilized thereafter, indicating that high intensity treatment after a long wait in state 2 provides little benefit to patients with regard to achieving the next speed benchmark. Plots of the integrated coefficient estimators for all covariates can be found in the supplemental materials associated with this manuscript [17].

4.2. Bone marrow transplant data

The bone marrow transplant (BMT) data set consists of 137 patients who received an experimental preparatory medication prior to transplant. After the transplant procedure, patients were followed and the time to several clinically important events recorded – platelet recovery, acute graft-versus-host-disease (aGVHD), chronic GVHD (cGVHD), relapse of leukemia, and death. The multistate model representing these events and their occurrence has the following states: (1) bone marrow transplant, the root state, (2) aGVHD as first event, (3) platelet recovery as first event, (4) platelet recovery secondary to aGVHD, (5) aGVHD secondary to platelet recovery, (6) cGVHD, (7) relapse or death, the absorbing state (Figure 4).

The network structure depicted in Figure 4 is not unique [19], and was chosen in this analysis in part to provide an acyclic network. Further, we have treated relapse as an absorbing state (with death) when in fact events subsequent to relapse (platelet recovery, aGVHD, cGVHD) did occur. The original investigators measured eleven prognostic indicators: patient and donor age, sex, and cytelomegalovirus (CMV) status, waiting time to transplant, French-American-British (FAB) classification of morphological status, treating hospital, administration of prophylactic treatment, and type of leukemia – acute lymphoblastic leukemia (ALL), and low- and high-risk acute myelogenous leukemia (AML). These data are available online [20] and have been described in greater detail elsewhere [10, 19]. While these data have received considerable attention as a multistate data set, to our knowledge an analysis of waiting times has yet to be conducted and our findings below are novel.

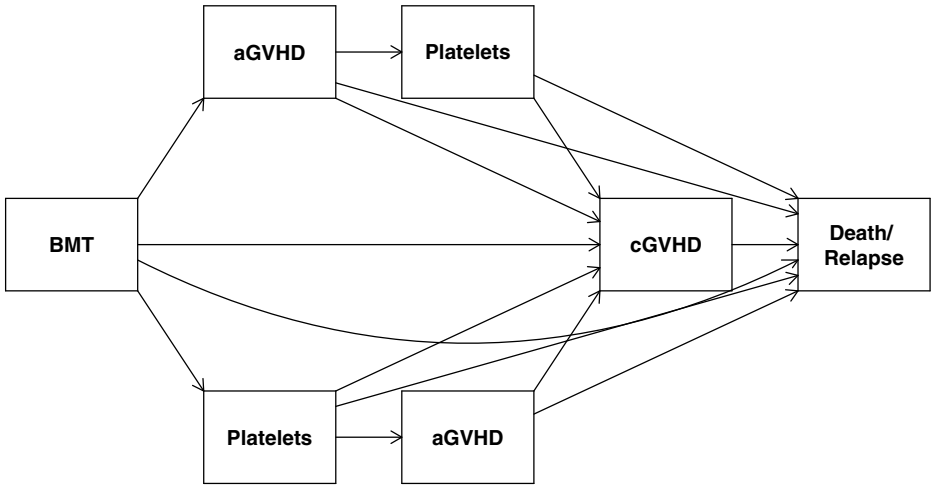

FIG 4. *Multistate network for bone marrow transplant data. In the analysis of these data, we consider waiting times in the transient cGVHD state (state 6).*

We modeled time to death or relapse after the acquisition of cGVHD (the waiting time in state 6) as a function of patient age, patient disease group, and time to onset of cGVHD after transplantation. These covariates were selected to represent a continuous external covariate, and categorical external covariate, and an internal covariate, respectively. Fifty-nine of the 137 patients experienced cGVHD, of which 27 relapsed or died and 32 were censored. Fifty-five of these 59 experienced platelet recovery as a first event prior to cGVHD or experienced platelet recovery followed by aGVHD, (i.e.) proceeded through states 3 and/or 5 of the model. Twenty-two patients were censored and 56 died or relapsed without acquiring cGVHD. A full description of observed transitions through the multistate network for these data is available in the supplemental materials associated with this manuscript [17]. The linear model for the state 6 hazard was $\lambda_{i6}(w) = \beta_{06}(w) + \beta_{16}(w)Z_{i1} + \beta_{26}(w)Z_{i2} + \beta_{36}(w)Z_{i3} + \beta_{46}(w)Z_{i4}$, where $Z_{i1} = T_{i6} - T_{i1}$ represented the time to onset of cGVHD after transplantation, Z_{i2} patient age at transplantation, Z_{i3} the indicator for low risk AML, and Z_{i4} the indicator for high risk AML (ALL was the reference group). In fitting this model, we calculated the inverse probability of censoring weights via a linear model for the censoring hazard containing the aforementioned eleven external covariates as well as six internal covariates denoting state occupation at calendar time t. The model specification exhausted all external covariates included in the data set.

The most significant predictor of death/relapse hazard after cGVHD onset was the time to onset of cGVHD (Z_{i1}, Figure 5). The negative estimate $\widehat{B}_{16}$ indicated that patients who more rapidly developed cGVHD also more rapidly relapsed or died after cGVHD onset. Tests based on asymptotic normality and bootstrapped standard errors showed that the estimated integrated coefficient

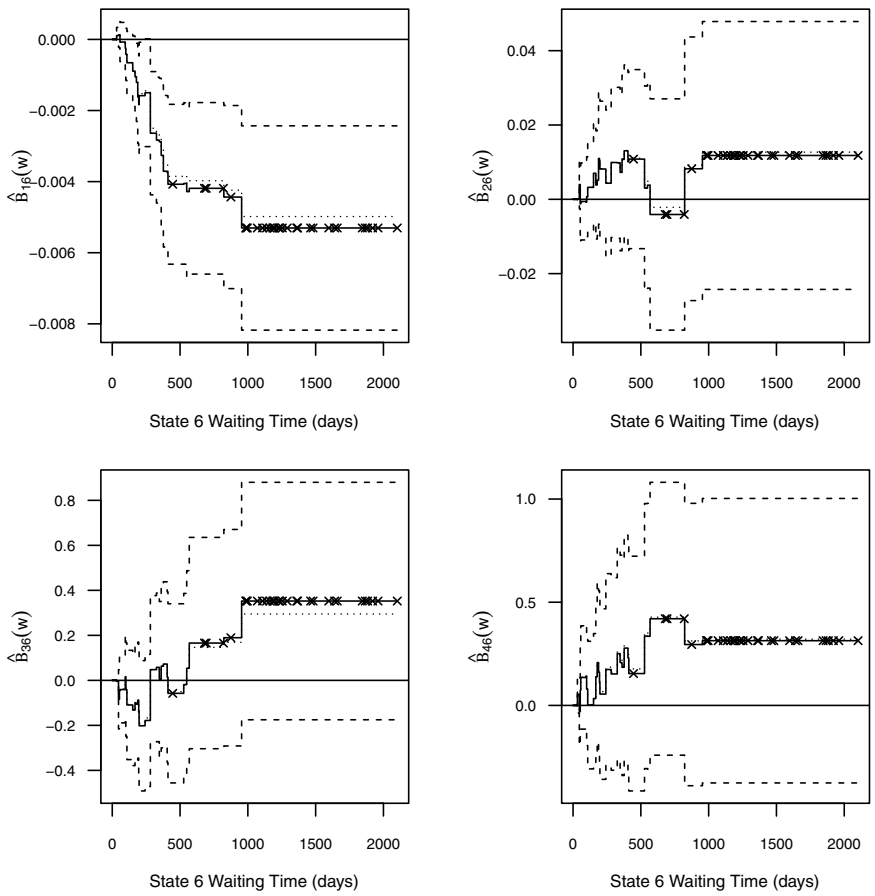

FIG 5. *Integrated coefficient functions from linear model fit to BMT data, representing regression effect of time to cGVHD onset ($\widehat{B}_{16}$), patient age ($\widehat{B}_{26}$), low risk AML ($\widehat{B}_{36}$), and high risk AML relative to ALL ($\widehat{B}_{46}$). Model intercept not shown. Solid lines plot the weighted estimators, dashed lines the asymptotic 95% CI, and dotted lines the unweighted estimator. Points plotted on the weighted estimate mark censored observations.*

$\widehat{B}_{16}$ was significantly different from zero 200 days after entry in state 6, and remained so until the end of observation at day 2102 ($z = 3.59, p = .0003$, Figure 5). The disease groups were not significantly different from each other as indicated by the bootstrapped 95% confidence intervals for $\widehat{B}_{36}$ and $\widehat{B}_{46}$, and patient age did not play a significant role in the hazard of death/relapse following cGVHD ($\widehat{B}_{26}, p = .53$). We additionally noted that the weighted estimators $\hat{\mathbf{B}}$ did not differ substantially from the unweighted estimators $\tilde{\mathbf{B}}$. This may have been a side effect of the inclusion of an internal covariate that was a function of the state 6 entry time – time to entry into state 6 (Z_{i1}) – a phenomenon noted in Section 2.

5. Discussion

The Cox relative risk and accelerated failure time models enjoy considerable popularity as regression models for survival data. Fully parametric models also receive frequent use, particularly for data adhering to the distributional assumptions made by parametric models. Aalen's linear model is less frequently used in applications, and it is possible that technical issues with the model are the reason for this. Two common issues with the model are the possibility of rank deficiency of the covariate matrix at a given time and generating negative cumulative hazard estimates. These issues appear to be a trade-off for the unique flexibility afforded Aalen's linear model through the definition of regression coefficients as functions of time rather than as static values. Ad hoc solutions to these problems have been suggested, including using alternative generalized inverses for the at-risk matrix $\mathbf{Y}_j(w)$ [15], discontinuing estimation of $\mathbf{B}_j(w)$ when $\mathbf{Y}_j(w)$ become rank deficient, and bounding cumulative hazard estimates below by zero [21], and appear to behave reasonably. It has been previously noted [15] that when using Aalen's linear model to generate inverse probability of censoring weights, these issues do not adversely impact weighted estimates of hazard and survival functions for failure time data.

As noted in the introduction, two issues plague the application of failure time methods in the analysis of right-censored multistate waiting times – (1) a semi-Markov property is required, even when censoring can be assumed to be independent, and (2) the observer does not know where individuals censored prior to entry into the transient state of interest would progress, and specifically whether or not such individuals would progress to the state of interest. Our simulation study was designed to highlight each of these issues, and showed the necessity of the semi-Markov property in the application of failure time methods in marginal analyses of waiting times from multistate models. When waiting times were correlated and many individuals were censored prior to exit from state 0, the survival data version of Aalen's linear model was substantially biased, a problem corrected by weighting by the inverse probability of censoring. Under simulation scenarios under which both the weighted and unweighted methods were unbiased, (i.e.) when waiting times were uncorrelated, the unweighted estimator exhibited lower variance and appeared optimal. McKeague [22] and Huffer and McKeague [23] suggested utilizing weighted least squares in defining the generalized inverse of the at risk matrix, thereby optimizing the variance of the integrated coefficient estimator. A comparison of variance estimates between the weighted and unweighted estimators under weighted least squares may be of interest.

The impact of mispecification of the censoring hazard has been a relatively unexplored research topic. The results of simulation model 3 showed that the weighted estimators can be biased when the censoring hazard model is misspecified, although the magnitude of this bias was far smaller than the bias exhibited by the unweighted estimator when waiting times were correlated. This bias was further reduced (although not completely eliminated) when time-varying state occupation indicators were used to model the censoring hazard. This point is of

particular importance, since information on state occupation is always available in the observation of a multistate system. We recommend the inclusion of state occupation indicators in modeling the censoring hazard. Further research into the impact of censoring hazard model misspecification is ongoing.

A potential shortcoming of our treatment of the spinal cord injury data was in a tacit assumption about censoring, that individuals in the rehabilitation program were right censored if discharged before reaching the absorbing state. Since individuals were evaluated for walking speed approximately every 20 treatment sessions rather than continuously monitored, an interval censoring approach may be more appropriate. We are currently working on extensions to the results presented here for interval censored data.

In our analysis of the bone marrow transplant and spinal cord injury data, we have used the asymptotic normality of the weighted estimator to judge the significance of model coefficients. Aalen [5] suggested that hypothesis test of individual and sets of model covariates be conducted by examining weighted integrated coefficients of the form $\int_0^t L_m(s)\mathrm{d}\widehat{B}_m(s)$, where $L_m(s)$ is a predictable weight process for covariate m. The weight process provides flexibility for optimizing power against certain classes of alternative hypotheses. These weighted tests can easily be applied to the weighted estimators for waiting times we have introduced, with appropriate definition of the weight process $L_m(s)$ for waiting times. The marginal analysis of clustered survival data with informative cluster size has received recent attention [24, 25]. The results presented here can conceivably be extended to apply to clustered waiting time data with informative cluster size through the simultaneous reweighting by the inverse probability of censoring and inverse cluster size [25], a potential next step in the marginal analysis of multistate waiting times.

Acknowledgements

This work was in part supported by the Centers for Disease Control and Prevention and the Christopher and Dana Reeve Foundation (grant/cooperative agreement U10/CCU220379). The authors thank the Christopher and Dana Reeve Foundation and all current and past members of the NeuroRecovery Network. Datta's research was supported by grants from the United States National Science Foundation (DMS-0706965) and National Security Agency (H98230-11-1-0168).

Appendix A: Martingale representation of the reweighted coefficient estimator $\hat{\mathbf{B}}_j(w)$

We first restate our notational convention for time. Calendar time will be denoted by t with variable of integration s. Processes evolving in calendar time, like the censoring process, will be functions of t (or s in stochastic integrals). Waiting time will be denoted by w with variable of integration v, and waiting time processes will be functions of w (or v in stochastic integrals).

Let the counting process martingale for the censoring event for individual i be $M_i^c(t) = N_i^c(t) - \int_0^t I[T_i \geq s]\mathrm{d}\Lambda_i^c(s)$, where $N_i^c(t) = I[T_i \leq t, \delta_i = 0]$, $\delta_i = I[T_i \leq C_i]$. The process $M_i^c(t)$ is a martingale with respect to the filtration $\mathcal{F}_{t,i}^c = \sigma\{I[T_i \leq t], \delta_i, \bar{\mathbf{Z}}_i(t), \mathcal{H}_{\infty,i}\}$, where δ_i, $\bar{\mathbf{Z}}_i(t)$, and $\mathcal{H}_{t,i}$ are as defined in Section 2. Note that $\mathcal{H}_{\infty,i}$ is included in the definition $\mathcal{F}_{t,i}$ for all $t \geq 0$ by assumption (5), which asserts that knowledge of future transition times does not impact the hazard of censoring. Let $\mathcal{F}_t^c = \vee_i \mathcal{F}_{t,i}^c$ and note that the vector $\mathbf{M}^c(t) = (M_1^c(t), \ldots, M_n^c(t))$ is a martingale with respect to $\mathcal{F}_t^c$.

A process closely related to $M_i^c(t)$ is $\widehat{M}_i^c(t) = N_i^c(t) - \int_0^t I[T_i \geq s]\mathrm{d}\widehat{\Lambda}_i^c(s)$, which we note is not a martingale, as it based on the estimated censoring hazard $\widehat{\Lambda}_i^c(s)$ rather than the true hazard $\Lambda_i^c(s)$. After taking differentials, these equalities can be combined to produce $\mathrm{d}\widehat{M}_i^c(t) = \mathrm{d}M_i^c(t) - I[T_i \geq t]\mathrm{d}(\widehat{\Lambda}_i^c(t) - \Lambda_i^c(t))$. By Aalen's linear model (cf. [11], Ch. VII), we note that $\mathrm{d}(\widehat{\Lambda}_i^c(t) - \Lambda_i^c(t)) = \mathbf{Y}_i^c(t)^{\mathrm{T}}\mathbf{Y}^c(t)^{-}\mathrm{d}\mathbf{M}^c(t)$, where $\mathbf{Y}^c(t)$ is the at-risk covariate matrix used in fitting Aalen's linear model to the censoring hazard, $\mathbf{Y}_i^c(t)$ is the i^{th} column of $\mathbf{Y}^c(t)$ corresponding to individual i, and $\mathbf{M}^c(t) = (M_1^c(t), \ldots, M_n^c(t))^{\mathrm{T}}$. We thus arrive at an equality relating $M_i^c(t)$ and $\widehat{M}_i^c(t)$,

$$\begin{aligned}
\mathrm{d}\widehat{M}_i^c(t) &= \mathrm{d}M_i^c(t) - I[T_i \geq t]\mathrm{d}(\widehat{\Lambda}_i^c - \Lambda_i^c)(t) \\
&= \mathrm{d}M_i^c(t) - \mathbf{Y}_i^c(t)^{\mathrm{T}}\mathbf{Y}^c(t)^{-}\mathrm{d}\mathbf{M}^c(t) \\
&= \left(\mathbf{E}_i^{\mathrm{T}} - \mathbf{Y}_i^c(t)^{\mathrm{T}}\mathbf{Y}^c(t)^{-}\right)\mathrm{d}\mathbf{M}^c(t) \\
&= \mathbf{P}_i(t)\mathrm{d}\mathbf{M}^c(t), \qquad (8)
\end{aligned}$$

where $\mathbf{E}_i$ is the i^{th} standard basis vector and $\mathbf{P}_i(t) = \mathbf{E}_i^{\mathrm{T}} - \mathbf{Y}_i^c(t)^{\mathrm{T}}\mathbf{Y}^c(t)^{-}$.

In proving variants of the equalities $E[\overline{N}_{ij}(w)] = E[N_{ij}^*(w)]$ and $E[\overline{Y}_{ij}(w)] = E[Y_{ij}^*(w)]$, Satten, Datta, and Robins [15] related the weighted counting processes to the uncensored data counting process by showing

$$\begin{aligned}
\widehat{N}_{ij}(w) &= \left(1 - \int_0^{U_{ij}^*-} \widehat{K}_i(s)^{-1}\mathrm{d}\widehat{M}_i^c(s)\right) N_{ij}^*(w), \\
\widehat{Y}_{ij}(w) &= \left(1 - \int_0^{T_{ij}^*+w-} \widehat{K}_i(s)^{-1}\mathrm{d}\widehat{M}_i^c(s)\right) Y_{ij}^*(w).
\end{aligned}$$

After substituting equality (8), these results take vectorized form

$$\begin{aligned}
\widehat{\mathbf{N}}_j(w) &= \mathbf{diag}\left\{1 - \int_0^{U_{ij}^*-} \widehat{K}_i(s)^{-1}\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)\right\}\mathbf{N}_j^*(w), \\
\widehat{\mathbf{Y}}_j(w) &= \mathbf{diag}\left\{1 - \int_0^{T_{ij}^*+w-} \widehat{K}_i(s)^{-1}\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)\right\}\mathbf{Y}_j^*(w). \qquad (9)
\end{aligned}$$

For simplicity, denote the diagonal matrices in (9) as $\mathbf{D}_N$ and $\mathbf{D}_Y(w)$, respectively, so that $\widehat{\mathbf{N}}_j(w) = \mathbf{D}_N\mathbf{N}_j^*(w)$ and $\widehat{\mathbf{Y}}_j(w) = \mathbf{D}_Y(w)\mathbf{Y}_j^*(w)$. For notational convenience in what follows, we suppress the term $\widehat{J}_j(w) = I[\mathrm{rank}(\widehat{\mathbf{Y}}_j(w)) =$

$p+1]$ in the estimator $\widehat{\mathbf{B}}_j(w)$, and assume throughout that the matrix $\widehat{\mathbf{Y}}_j(w)$ is of full column rank. We then have

$$\begin{aligned}
&\left(\widehat{\mathbf{B}}_j - \mathbf{B}_j\right)(w) = \\
&= \int_0^w \widehat{\mathbf{Y}}_j^-(v)\mathrm{d}\widehat{\mathbf{N}}_j(v) - \mathbf{B}_j(w) \\
&= \int_0^w \widehat{\mathbf{Y}}_j^-(v)\mathrm{d}\widehat{\mathbf{N}}_j(v) - \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{N}_j^*(v) + \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{N}_j^*(v) - \mathbf{B}_j(w) \\
&= \int_0^w \widehat{\mathbf{Y}}_j^-(v)(\mathbf{I}-\mathbf{D}_N)\mathrm{d}\mathbf{N}_j^*(v) - \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{N}_j^*(v) + \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{M}_j^*(v) \\
&= \int_0^w \left(\widehat{\mathbf{Y}}_j^-(v) - \mathbf{Y}_j^{*-}(v)\right)\mathrm{d}\mathbf{N}_j^*(v) - \int_0^w \widehat{\mathbf{Y}}_j^-(v)\mathbf{D}_N\mathrm{d}\mathbf{N}_j^*(v) + \\
&\quad \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{M}_j^*(v), \qquad (10)
\end{aligned}$$

where $\mathbf{M}_j^*(w) = (M_1^*(w), \ldots, M_n^*(w))^{\mathrm{T}}$ is the vector of counting process martingales for state j exits. We note that the $M_i^*(w)$ are martingales with respect to the filtration $\mathcal{F}_{w,i}^{(j)} = \sigma\{\bar{\mathbf{Z}}_i^{(j)}(w), \mathcal{H}_{w,i}^{(j)}\}$, the σ-algebra generated by observation of the covariates after state j entry and exits from state j. Further, $\mathbf{M}_j^*(w)$ is a martingale with respect to $\mathcal{F}_w^{(j)} = \vee_i \mathcal{F}_{w,i}^{(j)}$. Note that the above has established $\int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{N}_j^*(v) - \mathbf{B}_j(w) = \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{M}_j^*(v)$ in the uncensored data, a standard result for Aalen's linear model (cf. [11]).

Let $\mathbf{Y}_j^{*-}(w)$ denote the least squares left inverse of $\mathbf{Y}_j^*(w)$. A generalized inverse of $\widehat{\mathbf{Y}}_j(w)$ is then $\mathbf{Y}_j^{*-}(w)\mathbf{diag}\{1/(1-\int_0^{T_{ij}^*+w-}\widehat{K}_i(s)^{-1}\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s))\}$, which we write in shorthand as $\mathbf{Y}_j^{*-}(w)\mathbf{D}_{Y-}(w)$. The first integral in (10) is

$$\begin{aligned}
&\int_0^w \left(\widehat{\mathbf{Y}}_j^-(v) - \mathbf{Y}_j^{*-}(v)\right)\mathrm{d}\mathbf{N}_j^*(v) = \\
&\qquad = \int_0^w \left(\mathbf{Y}_j^*(v)\mathbf{D}_{Y-}(v) - \mathbf{Y}_j^{*-}(v)\right)\mathrm{d}\mathbf{N}_j^*(v) \\
&\qquad = \int_0^w \mathbf{Y}_j^*(v)\left(\mathbf{D}_{Y-}(v) - \mathbf{I}\right)\mathrm{d}\mathbf{N}_j^*(v) \\
&\qquad = \int_0^w \mathbf{Y}_j^*(v)\mathbf{diag}\left\{\frac{\int_0^{T_{ij}^*+v-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}{1-\int_0^{T_{ij}^*+v-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}\right\}\mathrm{d}\mathbf{N}_j^*(v).
\end{aligned}$$

The second integral in (10) is

$$\begin{aligned}
&\int_0^w \widehat{\mathbf{Y}}_j^-(v)\mathbf{D}_N\mathrm{d}\mathbf{N}_j^*(v) = \\
&\qquad = \int_0^w \mathbf{Y}_j^{*-}(v)\mathbf{D}_{Y-}(v)\mathbf{D}_N\mathrm{d}\mathbf{N}_j^*(v) \\
&\qquad = \int_0^w \mathbf{Y}_j^{*-}(v)\mathbf{diag}\left\{\frac{\int_0^{U_{ij}^*-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}{1-\int_0^{T_{ij}^*+v-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}\right\}\mathrm{d}\mathbf{N}_j^*(v),
\end{aligned}$$

whence the difference of the first two integrals in (10) is

$$\int_0^w \mathbf{Y}_j^{*-}(v)\mathbf{diag}\left\{\frac{\int_{T_{ij}^*+v}^{U_{ij}^*-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}{1-\int_0^{T_{ij}^*+v-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}\right\}\mathrm{d}\mathbf{N}_j^*(v), \tag{11}$$

and

$$\begin{aligned}\left(\widehat{\mathbf{B}}_j-\mathbf{B}_j\right)(w) &= \\ &= \int_0^w \mathbf{Y}_j^{*-}(v)\mathbf{diag}\left\{\frac{\int_{T_{ij}^*+v}^{U_{ij}^*-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}{1-\int_0^{T_{ij}^*+v-}\widehat{K}_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)}\right\}\mathrm{d}\mathbf{N}_j^*(v)+ \\ &\quad \int_0^w \mathbf{Y}_j^{*-}(v)\mathrm{d}\mathbf{M}_j^*(v).\end{aligned} \tag{12}$$

The second term of (12) is a martingale with respect to $\mathcal{F}_w^{(j)}$, since $\mathbf{Y}_j^{*-}(w)$ is $\mathcal{F}_w^{(j)}$-predictable by definition. The first term of (12) has k^{th} component

$$\begin{aligned}&\sum_{i=1}^n\int_0^w Y_{j,ki}^{*-}(v)\int_{T_{ij}^*+v}^{U_{ij}-}K_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)\times \\ &\qquad\left(1-\int_0^{T_{ij}^*+v-}K_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s)\right)^{-1}\mathrm{d}N_i^*(v)+o_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned} \tag{13}$$

where $Y_{j,ki}^{*-}(v)$ is the $(k,i)^{th}$ element of $\mathbf{Y}_j^{*-}(v)$ and the $o_p(\frac{1}{\sqrt{n}})$ is added with replacement of $\widehat{K}_i$ with K_i via the consistency of Aalen's linear model for the censoring hazard $(\Lambda_i^c(t))$ and the functional delta method for product integrals $(K_i(t))$. From (8), we have that

$$\begin{aligned}&1-\int_0^{T_{ij}^*+v-}K_i^{-1}(s)\mathbf{P}_i(s)\mathrm{d}\mathbf{M}^c(s) \\ &\quad=1-\int_0^{T_{ij}^*+v-}K_i^{-1}(s)\mathrm{d}M_i^c(s)+\int_0^{T_{ij}^*+v-}K_i^{-1}(s)\mathbf{Y}_i^c(s)\mathbf{Y}^{c-}(s)\mathrm{d}\mathbf{M}^c(s)\end{aligned} \tag{14}$$

The last term of (14) is $o_p(1)$ by laws of large numbers for martingales. By an argument similar to (A.5) of Satten, Datta, and Robins (2001) [15], we have that

$$1-\int_0^{T_{ij}^*+v-}K_i^{-1}(s)\mathrm{d}M_i^c(s)=\frac{I[C_i\geq T_{ij}^*+v]}{K_i(T_{ij}^*+v)}$$

Combining this with (12) and (13) and interchanging the order of integration gives

$$\left(\widehat{B}_{jk}-B_{jk}\right)(w)=\int_0^\infty \mathbf{H}_{n,k}(s,w)\mathrm{d}\mathbf{M}^c(s)+\int_0^w \mathbf{Y}_{j,k}^{*-}(v)\mathrm{d}\mathbf{M}_j^*(v)+o_p\left(\frac{1}{\sqrt{n}}\right), \tag{15}$$

where $\mathbf{H}_{n,k}(s,w) = \frac{1}{n}\sum_{i=1}^{n} I[T_{ij}^* \leq s < U_{ij}^*]K_i^{-1}(s)\zeta_i(s,w)\mathbf{P}_i(s)$, $\mathbf{Y}_{j,k}^{*-}(v)$ is row k of $\mathbf{Y}_j^{*-}(v)$, and $\zeta_i(s,w) = \int_0^{s-T_{ij}^* \wedge w} Y_{j,ki}^{*-}(v)K_i(v)\mathrm{d}N_i^*(v)$. Noting that $\mathbf{H}_{n,k}(s,w)$ is $\mathcal{F}_t^c$-predictable, we then have that $(\widehat{B}_{jk} - B_{jk})(w)$ and hence $(\widehat{\mathbf{B}}_j - \mathbf{B}_j)(w)$ are sums of two martingales – one with respect to $\mathcal{F}_t^c$ and one with respect to the waiting time filtration $\mathcal{F}_w^{(j)}$ – plus an asymptotically negligible term. Both martingales are asymptotically normal under suitable regularity conditions via the martingale central limit theorem [11]. The asymptotic independence of the two terms can be shown through a standard conditioning argument using characteristic functions. The conditioning argument exploits the fact that the σ-algebra $\mathcal{F}_0^c$ corresponding to the censoring process martingale contains the entire multistate transition history, $\mathcal{H}_\infty$ which in turn contains $\mathcal{F}_w^{(j)}$ for each w. Let M_1 and M_2 be the first and second terms of (15), respectively. Then, under appropriate conditions,

$$\begin{aligned} E\left[e^{it\sqrt{n}(M_1+M_2)}\right] &= E\left[e^{it\sqrt{n}M_1}E\left[e^{it\sqrt{n}M_2} \mid \mathcal{F}_0^c\right]\right] \\ &= E\left[e^{it\sqrt{n}M_2}\ e^{t^2\sigma_1^2/2}\right] + o(1) \\ &\to e^{t^2\sigma_1^2/2}\ e^{t^2\sigma_2^2/2} \end{aligned}$$

Through assumption (5) about the censoring hazard, $\mathcal{F}_0^c$ contains $\mathcal{H}_\infty$. Since $\mathcal{H}_\infty$ is the full history of transitions through the multistate system, M_2 is fixed by conditioning on $\mathcal{F}_0^c$ and pulled out of the inner conditional expectation. Under suitable regularity conditions and the martingale CLT, both M_1 and M_2 will each converge to a Gaussian martingale.

Supplementary Material

Supplement to "A nonparametric analysis of waiting times from a multistate model using a novel linear hazards model approach"
(doi: 10.1214/15-EJS1003SUPP; .pdf).

References

[1] Harkema, S.J., Schmidt-Read, M., Behrman, A.L., Bratta, A., Sisto, S.A., Edgerton, V.R., Establishing the NeuroRecovery Network: Multisite rehabilitation centers that provide activity-based therapies and assessments for neurologic disorders. *Archives of Physical Medicine and Rehabilitation* 2012; **93(9)**: 1498–1507.

[2] Harkema, S.J., Schmidt-Read, M., Lorenz, D., Edgerton, V.R., Behrman, A.L., Balance and ambulation improvements in individuals with chronic incomplete spinal cord injury using locomotor training-based rehabilitation. *Archives of Physical Medicine and Rehabilitation* 2012; **93(9)**: 1508–1517.

[3] van Hedel, H.J., Dietz, V., Rehabilitation of locomotion after spinal cord injury. *Restorative Neurology and Neuroscience* 2010; **28**: 123–134.
[4] Lorenz, D.J., Datta, S., Comparing waiting times in a multi-stage model: A log-rank approach. *Journal of Statistical Planning and Inference* 2012; **142**: 2832–2843. MR2925969
[5] Aalen, O.O., A model for nonparametric regression analysis of counting processes. In *Lecture Notes on Mathematical Statistics and Probability*, 2, Klonecki, W., Kozek, A., Rosiski, J. (eds.), New York: Springer-Verlag, 1980: 1–25. MR0577267
[6] Aalen, O.O., A linear regression model for the analysis of lifetimes. *Statistics in Medicine* 1989; **8**: 907–925.
[7] Aalen, O.O., Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine* 1993; **12**: 1569–1588.
[8] Huang, Y., Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society Series B* 2002; **64**: 17–29. MR1881842
[9] Schaubel, D.E., Cai, J., Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. *Biometrika* 2004; **91**: 291–303. MR2081302
[10] Copelan, E.A., Biggs, J.C., Thompson, J.M., Crilley, P., Szer, J., Klein, J.P. et al., Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with BuCy2. *Blood* 1991; **78**: 838–843.
[11] Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., *Statistical Models Based on Counting Processes*. New York: Springer-Verlag, 1993. MR1198884
[12] Robins, J.M., Rotnitsky, A., Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology – Methodological Issues*, Jewell, N., Dietz, K., Farewell, V. (eds.), 1993; 297–331. Boston: Birkhauser.
[13] Robins, J.M., Rotnitsky, A., Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 1995; **82**: 805–820. MR1380816
[14] Satten, G.A., Datta, S., Marginal estimation for multi-stage models: Waiting time distributions and competing risks analyses. *Statistics in Medicine* 2002; **21**: 3–19.
[15] Satten, G.A., Datta, S., Robins, J.M., Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters* 2001; **54**: 397–403. MR1861385
[16] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2012. Available at http://www.R-project.org/.
[17] Lorenz D.J., Datta, S., Supplement to "A nonparametric analysis of waiting times from a multistate model using a novel linear hazards model approach". 2015; DOI: 10.1214/15-EJS1003SUPP.

[18] Behrman, A.L., Ardolino, E.A., VanHiel, L., Kern, M., Atkinson, D., Lorenz D., et al., Assessment of functional improvement without compensation reduces variability of outcome measures after human spinal cord injury. *Archives of Physical Medicine and Rehabilitation* 2012; **93(9)**: 1518–1529.

[19] Klein, J.P., Moeschberger, M.L., *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer Verlag, 1997.

[20] Klein, J.P., Moeschberger, M.L., Modified by Yan J., Data sets from Klein and Moeschberger (1997), Survival Analysis 2003; R package version 0.1-3.

[21] Hosmer, D.W., Royston, P., Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model. *The Stata Journal* 2002; **2(4)**: 331–350.

[22] McKeague, I.W., Asymptotic theory for weighted least squares estimation in Aalen's additive risk model. *Contemporary Mathematics* 1988; **80**: 139–152. MR0999011

[23] Huffer, F.W., McKeague, I.W., Weighted least squares estimation for Aalen's additive risk model. *Journal of the American Statistical Association* 1991; **86**: 114–129.

[24] Cong, X.J., Yin, G., Shen, Y., Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics* 2007; **63**: 663–672. MR2395702

[25] Williamson, J.M., Kim, H.Y., Manatunga, A., Addiss, D.G., Modeling survival data with informative cluster size. *Statistics in Medicine* 2008; **27**: 543–555. MR2418464