Published in final edited form as:

Lifetime Data Anal. 2018 July ; 24(3): 464–491. doi:10.1007/s10985-017-9403-6.

Flexible Semi-parametric Regression of State Occupational Probabilities in a Multistate Model with Right-censored Data

Chathura Siriwardhana,

Department of Complementary & Integrative Medicine, University of Hawaii John A. Burns School of Medicine, Honolulu, HI, USA

K.B. Kulasekera, and

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, USA

Somnath Datta

Department of Biostatistics, University of Florida, Gainesville, FL, USA

Abstract

Inference for the state occupation probabilities, given a set of baseline covariates, is an important problem in survival analysis and time to event multistate data. We introduce an inverse censoring probability re-weighted semi-parametric single index model based approach to estimate conditional state occupation probabilities of a given individual in a multistate model under right-censoring. Besides obtaining a temporal regression function, we also test the potential time varying effect of a baseline covariate on future state occupation. We show that the proposed technique has desirable finite sample performances and its performance is competitive when compared with three other existing approaches. We illustrate the proposed methodology using two different data sets. First, we re-examine a well-known data set dealing with leukemia patients undergoing bone marrow transplant with various state transitions. Our second illustration is based on data from a study involving functional status of a set of spinal cord injured patients undergoing a rehabilitation program.

Keywords

Binary choice single index model; Bone marrow transplant; Multistate models; Right-censoring; Spinal code injury; State occupational probabilities

1 Introduction

Multistate models represent subjects' movement along time in terms of state occupation starting from an initial state to a final (absorbing) state. It can be a simple survival model that describes the transition between two states or a more complex model which contains several intermediate and final states. The well studied bone marrow transplant data described

Correspondence to: Somnath Datta.

Conflicts of interest

The authors declare that they have no competing interests.

by Copelan et al. (1991) is an example of such a system, which illustrates the transition of acute leukemia subjects in numerous different clinical states in time, after a bone marrow transplant. An important investigation for such a system centers around future state occupancy of an individual at a specific time since enrollment into the system given subject specific information. In reality, the complete movement of a subject in a multistate model may not be fully observable due to censoring, which restricts using typical regression concepts for this problem. In this work, we propose a new method to estimate the conditional state occupation probability given a subject's covariates in the presence of right-censoring. In our model, transitions between states are allowed to follow a dynamically varying nonlinear relationship with the individuals' covariates. Furthermore, the functional form of this non-linear relationship is semi-parametrically estimated at every time point using two single index models and thereby offering a greater flexibility in practice.

The literature of multistate models has been fairly dominated by parametric approaches over a long period; see Andersen and Keiding (2002) for examples. However several works based on fully nonparametric concepts have added a great momentum to this area. In the past, Aalen (1976, 1978) and Aalen and Johansen (1978) introduced nonparametric estimators of state occupational probabilities of a multistate model based on Nelson-Aalen type transition hazards. Datta and Satten (2001, 2002) showed the validity of these estimators under non-Markovian settings and extended their work further for subject dependent censoring. Andersen and Klein (2007) introduced a pseudo-values based regression approach starting with a marginal estimator, which could be both parametric or nonparametric, in order to take advantage of the flexible GEE approach of inference. Scheike and Zhang (2007) introduced direct modeling of covariate effects on state occupation probabilities using a generalized linear model incorporating inverse probability censoring weighting (IPCW), which is a straightforward approach that is easy to implement. Mostajabi and Datta (2013) developed a fully nonparametric approach to estimate conditional 'state-to-state transition counts' and 'the number at-risk' processes of a progressive multistate model under right-censoring given a covariate value, integrating the IPCW concept. However, their method handles only a single continuous covariate limiting the applicability. For competing risks multistate settings, the use of cause-specific cumulative incidence function has been popular (Fine and Gray, 1999).

We propose a novel method to estimate underlying temporal processes of a multistate model, conditionally given a covariate vector, by introducing the IPCW re-weighted binary choice Single Index Models (SIM) for estimating transition counts and number at-risk processes. This approach allows one to estimate the conditional transition matrix for a given covariate vector at a specific time point even when the transition mechanism has a highly nonlinear and rapidly varying dependency on multivariate baseline covariates. This is followed by a product limit calculation as in Datta and Satten (2002) to produce estimated conditional state occupying probabilities given baseline covariates. A detailed simulation study shows that the proposed method has desirable finite sample properties and it is robust under departures from the SIM assumption. Numerical comparisons show that our method is fairly competitive against existing suitable methods for estimating the regression function and testing the effect of a baseline covariate for future state occupation at a given time. We

demonstrate the utility of the proposed methodology using two real data sets; one pertaining to bone marrow transplantation and the other related to spinal cord injuries.

The rest of the paper is organized as follows. In Section 2, we describe the proposed semi-parametric estimation and resulting inferences for state occupying probabilities. Section 3 provides results of our simulation studies. In Section 4, we illustrate the proposed methods using two datasets: the bone marrow transplant data and the spinal code injury data. Finally, Section 5 contains a discussion.

2 The Proposed Methodology

In this section, we briefly describe the outline of a multistate model and notations used throughout the paper. Suppose that a set of n individuals move independently along an interconnected networked system of J states, $0, 1, \dots, (J-1)$, in a time-continuous multistate model. Such a model can be graphically represented with nodes to represent the states and arrows to represent allowable transitions from one state to another. We also assume that the model is progressive so that each state is entered at most once. For example, Figure 1 presents the progressive illness-death model that contains three states.

Consider the i th, $i = 1, \dots, n$, individual in the multistate system. Let $S_i(t)$ be the state occupied by the individual i at time t . Suppose, the time taken to reach an absorbing state is T_i^* , which can be subjected to right-censoring. Let C_i be the right-censoring time and $T_i = \min\{T_i^*, C_i\}$ is the observed final transition time. Define $\delta_i = I(C_i \geq T_i^*)$ which is an indicator of observing the final transition. Let $\delta_i(t) = I(C_i \geq t)$ and $K_i(t)$ be the “survival” function of the censoring distribution for the i th individual, given its generalized covariates; see Section 2.3 for a more detailed and precise description. Suppose $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})'$ is a p dimensional baseline covariate vector that we observe for all individuals in the study. Thus, our data consists of $\{S_i(s), 0 \leq s \leq T_i, X_i, \delta_i\}$, $1 \leq i \leq n$, corresponding to n individuals. Our goal is to estimate conditional state occupational probabilities given baseline covariate vector $X = x$: $p_j(t|x) = \Pr\{S(t) = j|X = x\}$, $j = 0, 1, \dots, J-1$, based on the observed data.

2.1 Binary Choice SIM Model for the Right-censored Data

Use of data weighting schemes for the purpose of bias reduction is well known in the statistical literature. One way to handle censored observations in the context of regression is to introduce a re-weighting scheme to the original estimator developed for complete data, in a way that the bias caused due to censoring asymptotically fades away. This idea was first introduced by Koul et al. (1981) for the randomly right censored data in linear regression.

The main body of this work is based on the IPCW re-weighted Klein and Spady (1993) binary choice SIM model, which is used to estimate conditional “transition counts” and “number at-risk processes” to formulate a conditional transition hazard matrix given a covariate value at a specific time point. Generally speaking, the normalized conditional number of at-risk processes estimated for a multistate model provides natural estimators for conditional state occupational probabilities. However, such estimators may not have a

smooth form with respect to time. Thus, we consider obtaining fairly stable estimates using a product limit function of transition hazards, which we will discuss in the sequel.

Let $N_{i,jj'}(t) = I(S_{i,jj'}(s-) = j, S_{i,jj'}(s) = j')$, for some $s \leq t$ be the individual specific counting process for transitions from state j to j' and let $Y_{i,j}(t) = I(S_{i,j}(t-) = j)$ be the individual specific at-risk process of occupying state j just prior to t , for the i th individual. These two processes may not be completely observable due to right-censoring. We define

$$N_{jj'}(t) = \sum_{i=1}^n N_{i,jj'}(t) \text{ and } Y_j(t) = \sum_{i=1}^n Y_{i,j}(t).$$

We assume conditional means of individual specific transition counting process $\mu_{i,jj'}^N(t | x_i)$ between states j to j' and individual specific at-risk process $\mu_{i,j}^Y(t | x_i)$ in the j th state, for the i th individual at time t are in following SIM forms:

$$\mu_{i,jj'}^N(t | x_i) = E(N_{i,jj'}(t) | X_i = x_i) = f_{t,jj'}(\beta'_{t,jj'} x_i), \quad (1)$$

$$\mu_{i,j}^Y(t | x_i) = E(Y_{i,j}(t) | X_i = x_i) = f_{t,j}(\beta'_{t,j} x_i), \quad (2)$$

where $f_{t,jj'}$, $f_{t,j}$ are unknown smooth functions and $\beta_{t,jj'}$, $\beta_{t,j}$ are p dimensional unknown parameter vectors. Here we develop the IPCW modified version of the Klein and Spady (1993) binary choice SIM model to estimate $\mu_{u,jj'}^N(t | x)$ and $\mu_{u,j}^Y(t | x)$, for a hypothetical individual (u) with a specified covariate vector $X = x$. Since both these functions are estimated in the same fashion, we describe the proposed algorithm in terms of a generic subject specific indicator process denoted by $I_{i,j}(t)$, without index notations j or j' . For our purpose, we let $I_{i,j}(t) = I(S_{i,jj'}(s-) = j, S_{i,jj'}(s) = j')$, for some $s \leq t$ or $I_{i,j}(t) = I(S_{i,j}(t-) = j)$ as appropriate.

We proceed to estimate $E(I_{i,j}(t) | X = x)$ in the following manner. Suppose, for a given t , the true model is

$$P(I_{i,j}(t) = 1 | X = x_i) = P_i(t) = F_t(\beta'_t x_i); i = 1, \dots, n,$$

where $F_t(\cdot)$ is a smooth function and β_t is a p dimensional parameter vector. Consider the IPCW re-weighted log-likelihood function L given by

$$L = \sum_{i=1}^n \frac{\delta_i(t')}{K_i(t')} \{I_{i,j}(t) \ln [F_t(\beta'_t x_i)] + [1 - I_{i,j}(t)] \ln [1 - F_t(\beta'_t x_i)]\},$$

where t' is the smaller of t and the exit time from state j , $K_i(t) = \prod_{s \leq t} [1 - \lambda_C(s | \bar{Z}_i(s)) ds]$ and $\bar{Z}_i(t)$ is a generalized covariate defined for the i th individual, which we will explain in detail

in the sequel when estimating $K_{\lambda}(t)$; here λ_C is the hazard rate function of the censoring distribution. Generally speaking $K_{\lambda}(t)$ does not have a survival function interpretation unless $\bar{Z}_{\lambda}(t)$ is formed with non-time-varying covariates. We will later describe a flexible model to estimate $K_{\lambda}(t)$, for any time t .

For the case of completely observed data, Klein and Spady (1993) introduced a semi-parametric likelihood, by approximating $F_{\lambda}(\beta' x_i)$ using a nonparametric estimator, which is similar to the leave-one-out estimator described by Ichimura et al. (1993). Thus, for right-censored data, we propose to use a similar estimator using the IPCW technique as

$$\hat{F}_{t-i}(\beta'_i x_i) = \frac{\sum_{l \neq i}^n \phi\left(\frac{\beta'_i x_i - \beta'_l x_l}{h}\right) \frac{I_l(t) \delta_l(t')}{K_{\lambda}(t')}}{\sum_{l \neq i}^n \phi\left(\frac{\beta'_i x_i - \beta'_l x_l}{h}\right)},$$

where $\phi(\cdot) \geq 0$ is a user specified kernel function with $\int \phi(u) du = 1$, and h is a smoothing parameter. This leads to a quasi log-likelihood function given by

$$L_q = \sum_{i=1}^n \frac{\delta_i(t')}{K_{\lambda}(t')} \left\{ I_i(t) \ln \left[\hat{F}_{t-i}(\beta'_i x_i) \right] + \left[1 - I_i(t) \right] \ln \left[1 - \hat{F}_{t-i}(\beta'_i x_i) \right] \right\} \hat{\tau}_i,$$

where $\hat{\tau}_i$ is a trimming sequence that is introduced by Klein and Spady (1993) to obtain a feasible likelihood criterion, which down-weights observations for which the corresponding densities are small. $\hat{\tau}_i$ is a weighting factor based on the density estimate of $\beta'_i x_i$. However, Klein and Spady (1993) show that the trimming procedure has a minor effect on the numerical performance of the estimator.

We maximize the quasi log-likelihood function L_q with respect to both β_t and h simultaneously to obtain the estimator $\hat{\beta}_t$ and the optimal smoothing parameter h_0 . Finally, an estimator of $F_t(\beta'_i x)$ for a given $X = x$ is given by

$$\hat{F}_t(\hat{\beta}'_i x) = \frac{\sum_{i=1}^n \phi\left(\frac{\hat{\beta}'_i x - \hat{\beta}'_l x_l}{h_0}\right) \frac{I_l(t) \delta_l(t')}{K_{\lambda}(t')}}{\sum_{i=1}^n \phi\left(\frac{\hat{\beta}'_i x - \hat{\beta}'_l x_l}{h_0}\right)}.$$

We estimate conditional means of the at-risk process and the event process between all possible pairs of nodes in the multistate model at time t for a given individual u with a covariate $X = x$ using the above procedure. Thus, we find $\hat{\mu}_{u,j}^Y(t | X = x)$ and $\hat{\mu}_{u,jj'}^N(t | X = x)$, estimators of $\mu_{u,j}^Y(t | X = x)$ and $\mu_{u,jj'}^N(t | X = x)$, respectively, by letting $I_{\lambda}(t)$ to be respective indicators corresponding to $Y_{i,j}(t)$ and $N_{i,jj'}(t)$. Also, we replace $K_{\lambda}(t)$ by its estimator $\hat{K}_{\lambda}(t)$.

Finally, we estimate $E(Y_j(t)|X=x)$ and $E(N_{jj'}(t)|X=x)$ by $n\hat{\mu}_{u,j}^Y(t|X=x)$ and $n\hat{\mu}_{u,jj'}^N(t|X=x)$, respectively.

Note that, we use the binary choice single index process to regress transition count processes. This requires only a single model to regress the transition from state j to j' at a given time when the multistate model is progressive. Otherwise, we will have to use similar models for every repeated transition (the first, second, etc.), which requires multiple models to handle transitions between those two states, making the methodology too complicated and then, the estimation becomes impossible due to sparsity in transition data.

In the estimation process, one can conveniently choose elements of t as the union of all event times of the multistate model. Since state-to-state conditional transition counts are non-decreasing functions of time, estimated means of the conditional transition counts are monotonized with respect to time via isotonic regression using the generalized pooled adjacent violators algorithm (Barlow et al., 1972; Leeuw et al., 2009).

2.2 Conditional Transition Hazard Rates and State Occupation Probabilities

In an uncensored experiment, the conditional hazard of transitions from state j to j' ($j \neq j'$) given a specific value of the covariate vector $X=x$ is given by

$$\alpha_{jj'}(t|X=x) = \lim_{dt \rightarrow 0} Pr\{S(s) = j', \text{ for some } s \in [t, t+dt) | S(t-) = j, X=x\}/dt,$$

where $S(t-)$ is the state occupied just before time t . Note that this conditional transition hazard only conditions on the current state for a given $X=x$. Hence this is also known as the partially conditional transition hazards (Pepe and Cai, 1993).

The cumulative (integrated) conditional state-to-state transition hazard matrix ($A(t|X=x)$) for the multistate model can be obtained as

$$A_{jj'}(t|X=x) = \begin{cases} \int_0^t \alpha_{jj'}(u|X=x)du, & \text{if } j \neq j' \\ -\sum_{j \neq j'} A_{jj'}(t|X=x), & \text{otherwise.} \end{cases}$$

Now we obtain an estimator for $A_{jj'}(t|X=x)$ as described in Datta and Satten (2002) as

$$\hat{A}_{jj'}(t|X=x) = \begin{cases} \int_0^t d\hat{E}(N_{jj'}(u|X=x))/\hat{E}(Y_j(u|X=x)), & \text{if } j \neq j' \\ -\sum_{j \neq j'} \hat{A}_{jj'}(t|X=x), & \text{otherwise.} \end{cases}$$

Note that components of the $\hat{A}(t|X=x)$ have the Nelson-Aalen form. Thus, $\hat{A}(t|X=x)$ is referred as the generalized Nelson-Aalen estimator for a multistate model. The process given by $d\hat{E}(N_{jj'}(t|X=x))$ can be obtained from the corresponding jumps of estimated

conditional state-to-state transitions between $[t-, t)$. In this calculation, we define the division by zero to be zero. Next we consider the product integral,

$$\hat{P}(s, t | X = x) = \prod_{(s, t]} (I + d\hat{A}(u | X = x)).$$

Estimators of state conditional occupation probabilities $p_j(t | X = x) = Pr(S(t) = j | X = x)$ on a given value of $X = x$ is given by

$$\hat{p}_j(t | X = x) = \sum_{k=0}^{J-1} \frac{\hat{Y}_k(0+ | X = x)}{n} \hat{p}_{kj}(0, t | X = x),$$

where $\hat{p}_{kj}(0, t | X = x)$ is the (k, j) -th element of the matrix $\hat{P}(0, t | X = x)$. Note that the validity of such state occupation formulas even for non-Markov models has been established by Datta and Satten (2001).

2.3 Censoring Hazards and Estimation of the Weights $K_i(t)$

Let $Z_i(t)$ be a generalized covariate defined for individual i , $1 \leq i \leq n$, at time t , which may contain both baseline and additional covariates (could be time varying) than covariates of primary interest (X) that affect the censoring hazards. For example, the current state occupation indicator at a given time may affect the censoring hazard besides baseline covariates. Suppose $\bar{Z}_i(t) = \sigma\{Z_i(s): 0 \leq s < t\}$ is the observed covariate and transition history prior to t . We assume that the censoring mechanism satisfies, $\lambda_C^i(\bar{Z}_i(t), S_i(\cdot)) = \lambda_C^i(t | \bar{Z}_i(t))$,

where $\lambda_C^i(t | \cdot) = \lim_{dt \rightarrow 0} Pr(C_i \in [t, t + dt) | T_i \geq t, \cdot) / dt$, which means the censoring hazard, given the covariate and transition history, does not depend on individuals' future state occupation (Datta and Satten, 2002). We use the Aalen's nonparametric additive model (Aalen, 1980 and 1989) to calculate IPCW weights, which provides a flexible structure to estimate the censoring hazards by allowing covariates to be varying over time. The additive censoring hazard of i th individual at time t is given by the linear form

$\lambda_C^i(t | \bar{Z}_i(t)) = \sum_{k=1}^m \eta_k(t) W_{ik}(t)$, where $W_{i0}(t) \equiv 1$ and $W_{ik}(t) = f_k(\bar{Z}_i(t))$, $k = 1, \dots, m$, are possibly time-dependent functions of the past history of the covariate process for subject i and $\eta_k(t)$ are unknown regression functions that measure the effect of corresponding covariate functions on the censoring hazard. Define $W_i(t) = (W_{i1}(t), \dots, W_{ip}(t))$. Then the Aalen's estimator of cumulative censoring hazard for the i th individual is given by

$$\Lambda_C^i(t | \bar{Z}_i(t)) = \int_0^t \hat{\lambda}_C^i(u | \bar{Z}_i(t)) du = \sum_{j=1}^n I(T_j \leq t) (1 - \delta_j) W_i(T_j) R^{-1}(T_j) W_j(T_j),$$

where $R(t) = \sum_{i=1}^n I(T_i \geq t) W_i(t) W_i'(t)$. The estimated IPCW weight for i th individual can be expressed as $\hat{K}_i(t) = \exp(-\hat{\Lambda}_C^i(t | \bar{Z}_i(t)))$, where $\Lambda_C^i(t | \bar{Z}_i(t)) = \int_0^t \lambda_C^i(u | \bar{Z}_i(t)) du$.

2.4 Bootstrap Confidence Interval for a State Occupation Probability

We describe a smooth bootstrap procedure to calculate a confidence interval for a state occupation probability. Generally speaking, since the estimated state occupation probabilities involve smoothing, the standard bootstrap procedure will not achieve the desired confidence level for the resulting confidence interval and an oversmoothed estimate is needed for centering in the “bootstrap world” (Marron, 1992; Li and Datta, 2001).

Let $\hat{\theta}_h$ be the estimator of θ , which is calculated using a sample of size n with a bandwidth h that results from a cross-validation process. Suppose $\hat{\theta}^*$ is the respective bootstrap estimate with the same bandwidth h computed from a bootstrap sample of size n that is obtained by re-sampling the original data with replacement. In the context of nonparametric regression of survival times, Li and Datta (2001) described the approximate distributional relationship given by $D(\hat{\theta}_h - \theta) \approx D(\hat{\theta}^* - \hat{\theta}_g)$, where $\hat{\theta}_g$ is the corresponding over-smoothed estimate that involves a larger smoothing parameter g ($g > h$), satisfying $\lim_{n \rightarrow \infty} \frac{h}{g} = 0$. For $0 < \varphi < 1$, one can select g as

$$g = \begin{cases} h^\varphi, & \text{if } h < 1 \\ h^{1/\varphi}, & \text{otherwise.} \end{cases}$$

There does not appear to be a usable formula for an optimal φ in practice. We made an ad hoc choice of $\varphi = 0.8$ in our calculations which seems to produce confidence intervals with reasonable coverage for settings investigated by us. We use the same principle as in Li and Datta (2001) to compute our bootstrap confidence interval. To this end let $v^{*b} = \hat{\theta}^{*b} - \hat{\theta}_g$, $b = 1, \dots, B$, where B is the number of bootstrap samples drawn from the given data. Suppose $(Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}})$ are the respective $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles $v^{*1}, \dots, v^{*B}$. Thus, $(\hat{\theta}_h - \theta)$ lies in the interval of $[Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}}]$ with approximate probability $(1-\alpha)$, which yields an approximate $(1-\alpha)100\%$ confidence interval for θ as $[\hat{\theta}_h - Q_{1-\frac{\alpha}{2}}, \hat{\theta}_h - Q_{\frac{\alpha}{2}}]$. This interval will be called a bootstrap (based) confidence interval for a state occupation probability, when $\theta = p_j(t|X = x)$.

2.5 Testing For Covariate Effects

In this section, we describe testing for covariate effects on state occupation probabilities using the IPCW re-weighted binary choice SIM model. For this purpose, we focused on the parameters estimated for risk processes, which are natural approximations for state occupation probabilities in a multistate model. Note that for any given state j , $j = 1, \dots, J$, and individual i , $i = 1, \dots, n$, $E(Y_{ij}(t)|X_i) = Pr(S_i(t) = j|X_i)$ is the conditional state occupation probability which, according to the SIM model (1) is given by $Pr(S_i(t) = j|X_i) = f_{t,j}(\beta'_{t,j}X_i)$. Therefore the effect of a covariate can be determined by Wald type tests based on the corresponding estimators $\hat{\beta}_{t,j}$. We have conducted empirical studies on testing of covariate effects using the asymptotic variance that comes from the sandwich variance estimator.

Although outcomes of this study is not presented in this manuscript, those indicated that the size is severely inflated when following this approach. Thus, we suggest a bootstrap method based approach for testing the covariate effect in finite sample problems. In this section, we focus on testing the effects of covariates on state occupation probabilities at a given time. However, it is also possible to test the overall effect of the covariate using a method based on a re-sampling scheme. In fact, we provide an illustration of this test in Section 4.1.

Since this testing does not require the smooth bootstrap technique that was described in Section 2.5, the proposed testing procedure is conducted with the standard bootstrap technique. Let $\hat{\beta}_{t,j}$ be the estimator of $\beta_{t,j}$ which is calculated using a sample of size n . Suppose $\hat{\beta}_{t,j}^{*b}$ is the respective estimate of $\beta_{t,j}$ calculated from $b = 1, \dots, B$, bootstrap sample of size n that is obtained from the original data with replacement. One may test the

hypotheses $H_0: \beta_{t,jp} = 0$ vs $H_1: \beta_{t,jp} \neq 0$, $p = 1, \dots, P$, using the test statistic $T = \frac{\hat{\beta}_{t,jp}}{\sigma_{\hat{\beta}_{t,jp}}^*}$,

where $\sigma_{\hat{\beta}_{t,jp}}^*$ is the standard error based on bootstrap samples of $\hat{\beta}_{t,jp}$. For large values of n ,

the test statistic T can be assumed to be approximately distributed as $N(0, 1)$, under the null hypothesis.

3 Simulation Study

3.1 Study Design

In this section we present a detailed simulation study that investigates the properties of the proposed procedure in finite samples.

We conducted the simulation study using the progressive illness-death model given in Figure 1. In this model, an individual starting from the initial healthy state (state-0) at time 0 moves to the absorbing state (state-2), denoted as death, by following one of two possible paths, where the path choice is controlled by an independent Bernoulli random variable. An individual at state-0 has a 60% chance of moving through the intermediate disease state (state-1) to the final state and 40% chance of reaching the final state directly. We simulated n independent individuals starting from state-0 at time 0, where n is chosen from the set $\{100, 500, 1000\}$. For each individual, we then generated a three dimensional covariate vector $X = (X_1, X_2, X_3)'$ that contains subject specific information, where its first component X_1 is generated from the Bernoulli distribution with probability 0.5 and the other two components (X_2, X_3) are obtained from a multivariate normal distribution with a zero mean vector and a dispersion matrix D given by

$$D = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}.$$

We assumed that individuals' transition times follow Log-Normal distributions, in such a way that the log-mean parameters depend on their baseline covariate vectors. Specifically,

depending on which path an individual takes, transition times T_{01} , T_{12} , and T_{02} are independently drawn using Log-Normal distribution $LN(\mu_{ij}, \sigma^2)$. We applied the proposed procedure to estimate state occupation probabilities for a specific value of an arbitrarily selected covariate vector $x_0 = (1.0, 0.2, -0.1)'$. In this work, we used uniform kernels for the IPCW re-weighted binary choice SIM models.

The proposed method is developed to incorporate cases where the state occupation probabilities contain nonlinear effects of covariates. Therefore, we primarily evaluate the performance of the proposed method using a simulation setting that uses such nonlinear effects (Setting-1). Because we also compare the new method with several existing approaches that are primarily taking into account linear effects of covariates, another simulation setting with linear effects (Setting-2) was used for that comparison.

Setting-1—Transition times T_{01} , T_{12} , and T_{02} are independently drawn from Log-Normal distributions $LN(\mu_{ij}, 0.5^2)$ with parameters $\mu_{01} = (\beta'X)^2 + 0.1$, $\mu_{12} = (\beta'X)^2$, and $\mu_{02} = (\beta'X)^2 + 0.5$, respectively. We chose $\beta = (0.40, 0.79, 0.46)'$ where $\|\beta\| = 1$.

We considered both random and covariate dependent right-censoring cases with 25% and 50% censoring rates. The random censoring times are generated from Exponential distributions with scale parameters 3×10^{-2} and 1.5×10^{-1} to achieve 25% and 50% censoring, respectively. For the covariate dependent censoring setting, we obtained censoring time C using an Exponential distribution that is specified by an indicator function for a threshold value of a linear combination of covariates given by

$$C \sim I(\rho'X > w) \exp(\zeta_1) + I(\rho'X \leq w) \exp(\zeta_2).$$

We took $w = 0.1$, $\rho = (0.3, 0.2, 0.5)'$, and pairs (ζ_1, ζ_2) equaling $(6 \times 10^{-2}, 1 \times 10^{-2})$ and $(1.8 \times 10^{-1}, 9 \times 10^{-2})$, to achieve 25% and 50% censoring rates, respectively.

Setting-2—Transition times T_{01} , T_{12} , and T_{02} are independently generated using Log-Normal distributions $LN(\mu_{ij}, 0.25^2)$ with parameters $\mu_{01} = \frac{1}{2}(\beta'X)$, $\mu_{12} = 0.4$, and $\mu_{02} = \frac{1}{2}(\beta'X) + 0.5$, respectively. Here β is fixed to be $\beta = (0.40, 0.79, 0.46)'$ with $\|\beta\| = 1$.

For Setting-2, we only consider random censoring cases with 25% and 50% censoring rates. These censoring times are generated from Exponential distributions with scale parameters 1.2×10^{-1} and 3×10^{-1} to achieve 25% and 50% censoring, respectively.

3.2 Absolute Error of Estimated State Occupation Probability

We examined the expected absolute error $L_1(t|x_0) = E|\hat{p}(t|x_0) - p(t|x_0)|$, measured between estimated and “true” state occupying probabilities, at 25th, 50th, and 75th percentile points of the temporal range for each state given $X = x_0$. Here $p(t|x_0)$ is the true conditional state occupation probability for $X = x_0$ for a given state and $\hat{p}(t|x_0)$ is the corresponding estimator obtained by the proposed method.

We approximated $p(t|x_0)$ by the proportion of state occupation using an uncensored experiment with large number (10,000) of data points that are generated using the covariate vector $X = x_0$. We calculated $L_1(t|x_0)$ via averaging absolute errors over 500 Monte-Carlo replicates.

In this study, we compared two different estimators with our proposed estimator. The first estimator is based on the Cox's model (Cox, 1972) and the second is the method given by Scheike and Zhang (Scheike and Zhang, 2007). In the Cox's approach, the state-to-state transition hazards in the multistate models are estimated using separate Cox's regression model fitted for each transition, where the corresponding baseline hazard function is estimated by Breslow's method (Breslow, 1972). Thus, the conditionally estimated transition hazards of transitions from states j to j' ($j, j' = 0, 1, 2$) at a given time t can be obtained as, $\hat{\alpha}_{jj'}^{Cox}(t | X) = \hat{\lambda}_0^{jj'}(t) \exp(\hat{\gamma}'X)$, where $\lambda_0^{jj'}(t)$ is the baseline hazard of transition from state j to j' at time t , $\gamma = (\gamma_1, \gamma_2, \gamma_3)'$ is the three dimensional vector that contains respective regression coefficients of baseline covariates $X = (X_1, X_2, X_3)'$. Next, we used the conditional transition hazards matrices based on $\hat{\alpha}_{jj'}^{Cox}(t | X)$ to obtain estimates for the state occupation probabilities $\hat{p}_j^{Cox}(t | X)$ at a specific covariate value $X = x_0$, by following the Aalen-Johansen (Aalen and Johansen, 1978) formulation as discussed in Section 2.3. Scheike and Zhang's method uses a direct modeling of individuals' state occupation indicators using IPCW integrated generalized linear model approach, which allows one to estimate the state occupation probabilities given a covariate and test the effect of covariates using censored data. In this work, we applied Scheike and Zhang's method using regression models with a logit link function. L_1 error rates for these two methods are computed in the same manner as above.

Table 1 presents the resulting $L_1(t|x_0)$ for the simulation Setting-1. Define $\Delta_{L_1}^{New}(t | x_0)$ to be $L_1(t|x_0)$ values for the proposed method. Demonstrating reasonable performances, results for the proposed method show a clear decline in the $\Delta_{L_1}^{New}(t | x_0)$, as the sample size is increased for all types of censoring and rates. As expected, $\Delta_{L_1}^{New}(t | x_0)$ is higher for scenarios with a larger censoring rate. $\Delta_{L_1}^{New}(t | x_0)$ at 50th and 75th percentile times are relatively larger than corresponding values at 25th percentiles.

Suppose $\log\{\Delta_{L_1}^{New}(t | x_0)\} \sim -\kappa \log(n)$. We observed a linear trend between $\log(n)$ and $\log\{\Delta_{L_1}^{New}(t | x_0)\}$ in every case, seemingly complying with this relationship. For example,

Figure 2 shows approximate linear relationships in plots of $\log(n)$ vs $\log\{\Delta_{L_1}^{New}(t | x_0)\}$ that are developed for a 50% random censoring rate using Setting-1. This infers that the proposed estimator is reaching the true conditional state occupying probability, in an asymptomatic fashion. Clearly, throughout the whole experiment $L_1(t|x_0)$ values for Cox's approach and Scheike and Zhang's method remain almost unchanged with sample size. Also, both these

competing methods appear to have considerably larger L_1 errors, especially at 50% and 75% percentile points compared with the proposed concept suggesting that these two methods are incapable of providing reliable state occupation probability estimates for the simulation Setting-1, which uses the nonlinear covariate effects.

Table 2 provides $L_1(\hat{\lambda}|x_0)$ for the simulation Setting-2. Clearly, $L_1(\hat{\lambda}|x_0)$ values decrease for all three methods when n increases. Even though Cox's and Scheike and Zhang's methods under performed in the previous setting, both these methods exhibit better performances in Setting-2. In fact, $L_1(\hat{\lambda}|x_0)$ values by the Cox's approach appear to be the lowest in most cases. Overall, we observe reasonable $L_1(\hat{\lambda}|x_0)$ values for all three methods for different censoring rates for the Setting-2, which only considers linear covariate effects.

3.3 Robustness of the Proposed Method

The proposed method for estimating conditional state-to-state transition count and at-risk processes is established on a strict assumption that the underlying functions have the SIM form. However, in practice this may not necessarily be the case. Thus, we performed a simulation study to investigate the robustness of the proposed method under perturbed SIM functions. Similar to Section 3.2, we examined the absolute error denoted by $L_1(\hat{\lambda}|x_0)$ at percentiles of transition times, using the same three-state progressive illness-death model described in our previous simulations, but using a new set of model functions that have a perturbed SIM structure. In particular, in this experiment, transition times T_{01} , T_{12} , and T_{02} are generated from Log-Normal distributions $LN(\mu_{ij}', 0.5^2)$ with parameters $\mu_{01} = (\beta^T X)^2 + (\vartheta X_2)^2 + 0.1$, $\mu_{12} = (\beta^T X)^2 + (\vartheta X_2)^2$, and $\mu_{02} = (\beta^T X)^2 + (\vartheta X_2)^2 + 0.5$ respectively, where X_2 is the second component in the covariate vector $X = (X_1, X_2, X_3)'$. Consequently, this new set of models violate the SIM assumption in temporal processes of the multistate model.

In this investigation, we chose β to be $(0.40, 0.79, 0.46)'$ and varied the value of ϑ from 0.5 to 1.25. The study was conducted under a 25% random censoring rate, using a sample of size $n=500$. Table 3 summarizes the $\Delta_{L_1}^{New}(t | x_0)$ values that are obtained by averaging 500 Monte-Carlo replicates. As to be expected, $\Delta_{L_1}^{New}(t | x_0)$ is increased with ϑ ; a further assessment of this result showed that the percentage change in $\Delta_{L_1}^{New}(t | x_0)$ for $\vartheta=0.5$, $\vartheta=0.75$, $\vartheta=1.00$, and $\vartheta=1.25$, compared to $\vartheta=0$ were in 5–47%, 6–94%, 16–141%, and 23–162%, respectively for the three different time points under consideration. Overall, we observed fairly reasonable $\Delta_{L_1}^{New}(t | x_0)$ values under above perturbed functions, indicating the robustness of the proposed procedure.

4 Applications

In this section, we illustrate two applications of the proposed method on real data. The first illustration uses a well known dataset on bone marrow transplant patients (Copelan et al., 1991) and the second illustration utilizes a dataset collected on patients suffering from severe spinal cord injury (Harkema et al., 2012).

4.1 Bone Marrow Transplant study

The bone marrow transplant study was conducted during the years 1984 to 1989 for acute leukemia patients in four worldwide centers. Bone marrow transplant is considered one of the standard therapies for acute leukemia. Usually, a subject experiences various clinical conditions during the recovery process after receiving the transplant from a donor, which can be represented as states of a multistate model. In this study, 137 acute leukemia subjects underwent bone marrow transplantation. After that, these individuals were followed up to a maximum of 7 years. In a multistate representation, starting from the primary state of receiving the bone marrow transplant, there are five intermediate states that an individual may reach before he/she reaches the final (absorbing) state, which is relapse or death by leukemia. Five intermediate states are represented by conditions of developing acute-graft-versus-host-disease (GVHD); returning of platelet levels to normal levels; returning of platelet levels to normal levels after developing acute GVHD; developing acute GVHD after platelet recovery; developing chronic GVHD. In this dataset, the censoring rate is rate is approximately 39% in terms of reaching to the final state. A schematic representation of the multistate model is given in Figure 3. We summarize the transition counts for the study in Table 4. For additional details of this study and the dataset, we refer readers to Klein and Moeschberger (1997).

The bone marrow transplant data described here has been analyzed by many authors from various angles. In this work, our primary goal is to estimate the state occupation probability for a hypothetical individual represented by a specific covariate vector. Among several available covariates, we use patient age and donor age for our analysis. We apply the proposed procedure by choosing the covariate vector $x = (28, 28)'$ corresponding to a hypothetical patient who receives a transplant from a healthy donor of the same age to estimate his/her state occupation probabilities as functions of time, for $\hat{p}_j(t|x), j = 0, \dots, 6$. We then construct pointwise bootstrap method based 95% confidence intervals following the same procedure described in the simulations section. In a previous analysis, Datta and Satten (2001) showed that the censoring hazard depends on individual's currently occupied state. Hence, we consider patient age and also the current state which is a time varying covariate, for calculating the subject specific IPCW weights.

Figure 4 provides a set of plots that show the estimated conditional state occupation probabilities for a hypothetical bone marrow transplant recipient of age 28 who receives a bone marrow transplant from a donor of the same age. The plot of state-0 shows the probability of staying at this state dramatically declines within a short period, which means that the individual possibly moves to another clinical state soon after he/she undergoes the transplant. As indicated by the plot for state-2, the most probable state after state-0 for such an individual appears to be the platelet recovery state (state-2). The probability of staying in this state at the very beginning appears to be as high as 0.7. However, the state occupation probability for this state quickly decreases below 0.3 and reaches an almost steady level within first 500 days of the surgery. Plots for states 1 and 3 clearly indicate that the chances of staying in acute GVHD state and platelet recovery after acute GVHD state are very low for an individual with these covariate values. Occupying an acute GVHD after platelet recovery state near the beginning is close to 0.1, which then declines gradually. The

occupation probability of the chronic GVHD state quickly increases to approximately 0.25 and remains almost unchanged for a long span. As expected, the chance of moving to the absorbing state (Relapse/Death) increases with time and reaches a plateau after nearly 1000 days after the transplant. Pointwise confidence bands based on 95% bootstrap confidence intervals (with $B = 500$) for state occupation probabilities indicate a reasonable precision of our estimates.

We provide an additional illustration (see Figure 5) of estimated conditional state occupation probabilities for two covariate vectors: $x = (20, 20)'$ and $x = (40, 40)'$. These values were chosen to represent the two scenarios: (i) young patient, young donor and (ii) older patient, older donor. Certainly, the analysis could be repeated to study other possible combinations as well.

Visual inspection of these plots shown some noticeable differences in the estimated state occupation probabilities. For example, the overall state occupation probability for platelet recovery after acute GVHD state for an individual with $x = (20, 20)'$ is larger compared to an individual with $x = (40, 40)'$. On the other hand, this trend reverses for occupying the chronic GVHD state. To determine if these differences are statistically significant, we developed a test statistic and computed its corresponding p -value using a re-sampling scheme. First, we calculated the mean absolute difference $D = \int |\hat{p}(t|X = x_1) - \hat{p}(t|X = x_2)| dE_{\hat{p}}(t)$, to quantify the overall difference for the above covariate vectors, where $\hat{p}(t|X = x)$ and $dE_{\hat{p}}(t)$ are the proposed estimator of state occupation probability given $X = x$ and the empirical cumulative distribution function of event times, respectively. We then drew a bootstrap sample of size n from the original data by re-sampling its labels with replacement. From this step, we obtained re-sampled event times. Next, we re-sampled X values from the original data with replacement. This was performed independently from the previous step. After that, we composed the final re-sampled dataset by attaching re-sampled X values with re-sampled event times. We calculated D using the above re-sampled data, denoted by D^* . This procedure was repeated for B times to obtain $D_1^*, \dots, D_B^*$. Then the p -value was computed by $p = \frac{1}{B} \sum_{b=1}^B I(D_b^* \geq D)$. In Table 5, we present the p -values based on $B = 500$ bootstrap samples. As shown in the table, the difference of state occupation probabilities for the platelet recovery after acute GVHD state is significant at a 5% level. Other differences such as in acute GVHD after platelet recovery appear to be borderline significant.

4.2 Spinal Cord Injury Study

The spinal cord injury (SCI) data set contains information on 296 subjects with clinically incomplete spinal cord injury from a national activity-based rehabilitation program. The program initially focused on individuals with clinically incomplete SCI after discharged from inpatient rehabilitation (Harkema et al., 2012). During the program, patients underwent sessions of locomotor training, based on a screening assessment at enrollment. The locomotor training program is an activity-based therapy for functions relating to standing and walking. The functional progress of these individual has been thoroughly evaluated periodically in terms of mobility, standing, and stepping after receiving therapeutic sessions. For instance, this study assessed 6-minute walk and 10-meter walk tests. There are several

clinical benchmarks defined based on the walking speed of an individual. For example, 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 (van Hedel and Dietz, 2010). Thus, we represent the progress of a SCI subject through the rehabilitation program as states of a multistate model as follows: (1) nonambulatory, (2) able to walk but slower than 0.44 m/s, (3) able to walk but in between 0.44 to 0.7 m/s, (4) able to walk but in between 0.7 m/s to 1.2 m/s, (5) able to walk faster than 1.2 m/s. A graphical representation of the multistate model with 5 states is given in Figure 6. We also provide a summary of transition counts in Table 6. It is important to note that individuals in this study have been entered into the multistate model from various states, which is one of the primary differences between SCI and bone marrow transplant studies. In a previous analysis of this dataset, Lorenz and Datta (2015) estimated an individuals' average waiting time to reach state-1 in the model, based on a linear hazards model approach.

Among several available subject specific measures, we used (1) initial speed at enrollment, (2) patient age at enrollment, (3) time from spinal cord injury to enrollment, and (4) lower motor score from the International Standards for Neurological Classification to estimate the state occupation probability for a given individual with a specified baseline covariate vector. We noticed that SCI subjects with severe conditions, such as nonambulatory or patients with lower walking speed than 0.44 m/s, very rarely regain a better walking ability than 1.2 m/s. Thus, observations for most of these patients have been subjected to right-censoring before they were discharged from the program upon recovery, resulting in approximately 82% censoring rate in terms of reaching the final state. We believe the censoring hazards may have an effect from the cumulative number of training sessions received by the individual prior to censoring, which can be considered as a time varying covariate. Thus, we use this information in addition to baseline covariates to calculate the IPCW weights. Figure 7 presents conditional state occupation probabilities estimated using the proposed method along with 95% bootstrap confidence intervals ($B = 150$), for a test individual with a covariate vector $x = (0.08, 38.0, 0.92, 33.0)'$, which corresponds to medians of baseline covariates for study participants. For an individual with this covariate vector, the chance of occupying state-0 at the early period is close to 0.5, which then sharply declines. There is an approximately 25% chance of occupying state-1 at the beginning, which relatively increases over time. The chance of occupying state-2 declines quickly, while the chance of occupying states 3, and 4 considerably elevates when reaching 500 days after enrollment.

5 Discussion

We proposed a novel and flexible method to estimate the conditional state occupational probabilities of a multistate model, given a set of covariates, in the presence of right-censoring. The primary advantage of the proposed method is its ability to model a complex multistate system, where the transition mechanism follows time varying nonlinearity with multiple baseline covariates. Fully nonparametric approaches are available in the literature for one covariate at a time due to the curse of dimensionality. Thus this paper offers a practical solution to study the effects of multiple covariates in a robust fashion. Furthermore, hypotheses testing have been conducted with this setting for studying the effect of a

covariate on the occupation probability of a given state. Likewise, a bootstrap based significance test for comparing the state occupation probabilities based on different baseline covariate profiles has also been developed.

We presented a series of simulations that investigate properties of the proposed method in finite samples, under nonlinear and linear covariate effects. We primarily focused on the L_1 distance between the estimated and true conditional state occupational probabilities given a vector of covariates at a selected set of time points. Clearly, L_1 distance seems to be decreasing with the sample size for both random and covariate dependent censoring mechanisms, under both low and high censoring rates, which seemingly suggests that the proposed estimator converges to the true conditional state occupation probability in an asymptotic sense. The robustness against the departure from the SIM structure signifies the potential capability of the proposed method for handling many real applications. A study that investigates the coverage of bootstrap confidence intervals demonstrated reasonable agreement between the empirical coverage and the nominal level. We also demonstrated the performance of the proposed method in examining covariate effects using a power study, reflecting the advantage of the new concept compared to two other existing approaches under complex transition mechanisms. We refer readers to the Electronic Supplement for additional details of these simulation studies. Two applications of the proposed method on real data were illustrated using bone marrow transplantation and SCI data, which contain several possible clinical states.

One of the drawbacks of the proposed concept is the heavy computational cost. In this work, we conducted all simulation studies and analyses on a high performance research computing cluster. It approximately takes 200 hours to generate results for Table 1 by the proposed method, using 2000 parallel cluster nodes.

A possible extension for the current methodology could be a generalization for high dimensional covariates such as genomic biomarkers for greater applicability in a broad class of medical data. This could be achieved by imposing strong dimension reduction criteria.

We are primarily concerned with baseline covariates in this paper. It is possible to incorporate time varying covariates to the models for estimating transition counts and number at-risk processes. However, the product limit formulas leading to state occupation probabilities may not hold in general with time varying covariates and the interpretation of such conditional probabilities is not possible.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Christopher and Dana Reeve Foundation and all current and past members of the NeuroRecovery Network for the provision of the spinal cord injury data. Also, we thank the associate editor and two anonymous reviewers for many constructive suggestions.

References

- Aalen OO. Non-parametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics*. 1976; 3:15–27.
- Aalen OO. Non-parametric inference for a family of counting processes. *The Annals of Statistics*. 1978; 6:701–726.
- Aalen, OO. A model for non-parametric regression analysis of counting processes. In: Klonecki, W.Kozek, A., Rosiski, J., editors. *Lecture Notes on Mathematical Statistics and Probability 2*. Springer; New York: 1980. p. 1-25.
- Aalen OO. A linear regression model for the analysis of lifetimes. *Statistics in Medicine*. 1989; 8:907–925. [PubMed: 2678347]
- Aalen OO, Johansen S. An empirical transition matrix for nonhomogeneous Markov chains based censored observations. *Scandinavian Journal of Statistics*. 1978; 5:141–150.
- Andersen PK, Keiding N. Multi-state models for event history analysis. *Statistical Methods in Medical Research*. 2002; 11:91–115. [PubMed: 12040698]
- Andersen PK, Klein JP. Regression analysis for multistate models based on a pseudo-value approach, with application to bone-marrow transplant studies. *Scandinavian Journal of Statistics*. 2007; 34:3–16.
- Barlow, RE., Bartholomew, DJ., Bremner, JM., Brunk, HD. *Statistical Inference under Order Restrictions*. New York: Wiley; 1972.
- Breslow NE. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B*. 1972; 34:216–217.
- Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, et al. Treatment for Acute Myelocytic Leukemia With Allogeneic Bone Marrow Transplantation Following Preparation With BuCy2. *Blood*. 1991; 78:838–843. [PubMed: 1859895]
- Cox DR. Regression model and life-tables. *Journal of Royal Statistical, Society Series B*. 1972; 34:187–220.
- Datta S, Satten GA. Validity of the Aalen-Johansen estimators of state occupation probabilities and integrated transition hazards for non-Markov models. *Statistics and Probability Letters*. 2001; 55:403–411.
- Datta S, Satten GA. Estimation of integrated transition hazards and stage occupation probabilities for non-Markov system under dependent censoring. *Biometrics*. 2002; 58:792–802. [PubMed: 12495133]
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 1999; 94:496–509.
- Harkema SJ, Schmidt-Read M, Behrman AL, Bratta A, Sisto SA, et al. Establishing the NeuroRecovery Network: Multisite rehabilitation centers that provide activity-based therapies and assessments for neurologic disorders. *Archives of Physical Medicine and Rehabilitation*. 2012; 93:1498–1507. [PubMed: 21777906]
- Ichimura H, Hall P, Hardle W. Optimal smoothing in single index models. *Annals of Statistics*. 1993; 21:157–178.
- Klein, JP., Moeschberger, ML. *Survival Analysis: Techniques for Censored and Truncated data*. New York: Springer-Verlag; 1997.
- Klein RW, Spady RH. An efficient estimator for binary response models. *Econometrica*. 1993; 66:387–421.
- Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right censored data. *Annals of Statistics*. 1981; 9:1276–1288.
- Leeuw J, Hornik K, Mair P. Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. *Journal of Statistical Software*. 2009; 32:1–24.
- Li G, Datta S. A Bootstrap Approach to Nonparametric Regression for Right Censored Data. *Annals of the Institute of Statistical Mathematics*. 2001; 53:708–729.
- Lorenz DJ, Datta S. A nonparametric analysis of waiting times from a multistate model using a novel linear hazards model approach. *Electronic Journal of Statistics*. 2015; 9:419–443.

- Marron, J. Bootstrap bandwidth selection. In: LePage, R., Billard, L., editors. Exploring the Limits of Bootstrap. Wiley; New York: 1992. p. 249-262.
- Mostajabi F, Datta S. Nonparametric regression of state occupation, entry, exit, and waiting times with multistate right-censored data. *Statistics in Medicine*. 2013; 32:3006–3019. [PubMed: 23225570]
- Pepe MS, Cai J. Some graphical displays and marginal regression analysis for recurrent failure times and time dependent covariates. *Journal of American Statistical Association*. 1993; 88:811–820.
- Satten GA, Datta S. Marginal estimation for multi-state models: waiting time distribution and competing risks analyses. *Statistics in Medicine*. 2002; 21:3–19. [PubMed: 11782047]
- Scheike TH, Zhang M. Direct Modelling of Regression Effects for Transition Probabilities in Multistate Models. *Scandinavian Journal of Statistics*. 2007; 34:17–32.
- van Hedel HJ, Dietz V. Rehabilitation of locomotion after spinal cord injury. *Restorative Neurology and Neuroscience*. 2010; 28:123–134. [PubMed: 20086289]

Fig. 1.
Graphical representation of a three state progressive illness-death model

Fig. 2.

Plots of $\log(n)$ vs $\log(\Delta_{L_1}^{New}(t | x_0))$ generated at 25th, 50th, and 75th percentiles of state entry (states - 1, 2) and exit (state - 0) times, under random censoring with a rate of 50% for Setting-1. Dotted lines represent a least squares line fitted to three points.

Fig. 3.
Graphical representation of a multistate model for the bone marrow transplant data.

Fig. 4.

Plots of estimated conditional state occupational probabilities of seven clinical states in the bone marrow transplant data obtained using the proposed method for a hypothetical individual of age 28 who receives a bone marrow transplant from a donor of the same age, along with 95% bootstrap based pointwise confidence intervals (represented by the dotted lines).

Fig. 5.

Plots of estimated conditional state occupational probabilities of seven clinical states in the bone marrow transplant data obtained using the proposed method for two hypothetical individuals with covariate vectors: $(x_1, x_2) = (20, 20)$ and $(x_1, x_2) = (40, 40)$. x_1 and x_2 represent patient age and donor age, respectively.

Fig. 6.
Graphical representation of a multistate model for the spinal code injury data.

Fig. 7.

Plots of estimated conditional state occupational probabilities of five clinical states in the spinal code injury data obtained using the proposed method for a hypothetical individual with a covariate vector given by $x = (0.08, 38.0, 0.92, 33.0)'$, along with 95% bootstrap based pointwise confidence intervals (represented by the dotted lines).

L_1 Distances for estimated state conditional occupying probabilities ($L_1(d|x_0)$), at 25th, 50th, and 75th percentiles of state entry (states -1, 2) and exit (state - 0) times for Setting-1 with two different censoring mechanisms for the proposed method, Cox-Regression approach, and Scheike and Zhang’s method. The true conditional state occupation probability given the covariate value x_0 , is represented by $P(d|x_0)$.

L ₁ (d x ₀) (10 ⁻²)																	
State	Percentile Time Point & p(d x ₀)	Sample Size (n)	No Censoring			Random Censoring			Covariate Dependent Censoring								
			Proposed Method	Cox's Method	Scheike & Zhang's Method	25%	50%	Proposed Method	Cox's Method	Scheike & Zhang's Method	25%	50%	Proposed Method	Cox's Method	Scheike & Zhang's Method		
State-0	25% 0.7500	100	6.67	8.46	9.47	6.61	8.34	8.98	6.63	8.28	9.24	6.92	8.05	9.12	7.47	8.21	9.01
		500	3.61	8.50	9.10	3.38	8.61	9.10	3.55	8.49	9.02	3.71	8.25	8.77	3.88	7.95	8.58
		1000	2.71	8.45	9.14	2.58	8.53	9.22	2.58	8.45	9.12	2.84	8.22	8.95	3.65	8.22	8.89
	100	7.72	18.36	20.03	7.40	18.36	19.33	8.12	18.36	18.78	7.93	18.21	19.26	8.16	17.92	18.76	
	500	4.57	18.21	19.85	4.32	18.68	19.83	4.40	18.46	19.16	4.54	17.96	19.39	4.90	18.11	18.87	
	1000	3.35	18.14	19.90	3.27	18.43	19.94	3.36	18.29	19.40	3.43	17.80	19.58	3.71	17.80	19.11	
State-1	75% 0.2500	100	7.19	26.35	29.67	8.03	26.28	28.65	8.15	26.24	26.55	7.63	26.00	28.95	7.68	25.45	26.68
		500	4.31	25.96	29.35	4.44	26.75	29.30	4.72	26.40	27.92	4.16	25.57	28.63	4.49	25.90	27.13
		1000	3.22	25.87	29.65	3.23	26.39	29.46	3.50	26.21	28.03	2.95	25.39	28.86	3.71	25.39	27.48
	100	5.24	5.15	6.25	5.52	5.44	6.08	5.87	5.38	6.02	5.46	5.15	6.03	5.90	4.90	5.73	
	500	2.97	4.97	5.61	2.87	5.11	5.70	2.87	5.06	5.35	2.80	4.94	5.33	3.09	5.14	5.06	
	1000	2.33	4.92	5.64	2.25	5.01	5.60	2.28	4.98	5.34	2.32	4.84	5.40	3.68	4.84	5.20	
State-1	50% 0.2757	100	6.56	11.15	12.19	6.25	11.45	11.62	6.33	11.51	11.19	6.87	11.27	11.73	7.48	11.03	11.19
		500	4.19	11.05	11.91	3.78	11.24	12.13	3.60	11.17	11.58	3.58	11.07	11.71	4.18	11.26	11.41

L ₁ (d x ₀) (10 ⁻²)																	
State	Percentile Time Point & p(d x ₀)	Sample Size (n)	No Censoring			Random Censoring			Covariate Dependent Censoring								
			Proposed Method	Cox's Method	Scheike & Zhang's Method	Proposed Method	Cox's Method	Scheike & Zhang's Method	Proposed Method	Cox's Method	Scheike & Zhang's Method						
State-1	75% 0.3447	1000	2.86	10.98	12.18	2.96	11.07	12.12	3.08	11.06	11.86	3.04	10.81	11.92	3.89	10.81	11.57
		100	7.48	14.09	15.65	7.51	14.64	15.27	8.12	14.66	14.43	7.93	14.46	15.42	8.28	13.92	14.78
		500	4.37	14.06	15.56	4.45	14.27	15.87	5.04	14.11	14.73	4.42	13.92	15.44	5.11	14.28	14.63
		1000	3.06	13.90	15.92	3.08	14.05	15.67	3.36	14.02	14.91	2.99	13.87	15.54	4.21	13.87	14.88
	100	7.21	7.58	9.10	6.40	7.57	8.66	6.79	7.31	8.94	7.16	7.13	8.79	8.48	6.85	8.51	
	500	3.40	7.19	8.57	3.51	7.52	8.22	3.72	7.39	8.36	3.88	6.95	8.20	4.21	7.32	8.03	
	1000	2.74	7.23	8.36	2.92	7.46	8.50	3.17	7.37	8.53	2.95	6.94	8.30	4.38	6.94	8.07	
	100	7.87	15.72	19.62	9.15	16.07	19.33	8.93	15.76	18.89	8.91	15.07	19.28	10.10	15.08	18.17	
	500	4.58	15.64	19.62	4.25	16.22	19.68	4.74	16.06	19.35	4.70	15.31	19.00	5.51	15.17	18.64	
	1000	3.56	15.61	19.58	3.82	16.11	19.70	3.83	15.93	19.51	3.67	14.89	19.17	4.86	14.89	18.61	
State-2	75% 0.7500	100	7.23	22.66	30.71	7.32	23.76	30.35	8.29	23.33	29.47	7.61	22.09	29.82	8.09	22.22	28.54
		500	4.59	23.10	30.92	4.33	24.03	31.09	4.90	23.72	30.88	4.62	22.57	30.23	6.08	22.02	29.21
		1000	3.10	23.06	31.01	3.57	23.83	31.05	4.04	23.54	31.06	3.36	22.11	30.29	5.28	22.11	29.61

Table 2

L_1 Distances of estimated state conditional occupying probabilities ($L_1(d|x_0)$), at 25th, 50th, and 75th percentiles of state entry (states- 1, 2) and exit (state - 0) times for Setting-2 with random censoring for the proposed method, Cox-Regression approach, and Scheike and Zhang's method. The true conditional state occupation probability, given the covariate value x_0 , is represented by $p(d|x_0)$.

State	Percentile Time Point & $p(d x_0)$	Sample Size (n)	No Censoring			25% Censoring Rate			50% Censoring Rate		
			Proposed Method	Cox's Method	Scheike & Zhang's Method	Proposed Method	Cox's Method	Scheike & Zhang's Method	Proposed Method	Cox's Method	Scheike & Zhang's Method
State-0	25% 0.2500	100	8.04	5.94	6.88	8.39	5.89	7.13	8.67	5.63	8.47
		500	4.57	3.11	3.54	4.70	2.80	3.80	5.29	2.38	4.11
		1000	3.84	2.68	3.22	3.93	2.35	3.06	4.32	1.75	3.49
	50% 0.5000	100	6.87	7.40	8.35	7.62	7.11	9.24	7.96	7.71	11.34
		500	4.18	3.91	3.32	4.10	4.03	4.06	4.65	4.54	5.95
		1000	3.22	3.56	2.46	3.50	3.56	2.79	3.74	3.63	3.25
State-1	75% 0.2500	100	5.43	4.96	7.18	6.40	11.06	8.62	7.08	8.63	10.41
		500	3.34	2.12	4.43	3.06	4.84	5.07	3.55	4.48	5.54
		1000	2.59	1.65	3.28	2.67	3.82	4.35	2.91	4.09	4.03
	25% 0.1480	100	7.41	4.40	5.77	7.61	7.02	6.29	7.75	4.02	6.85
		500	4.50	2.50	3.06	4.72	2.50	3.23	4.99	2.04	3.88
		1000	4.01	2.25	2.55	4.20	2.59	2.83	4.36	1.84	3.23
State-1	50% 0.2958	100	6.05	5.65	8.28	5.84	5.42	8.85	6.33	5.96	11.04
		500	3.44	2.28	6.38	3.49	2.52	6.52	3.75	3.52	6.54

State	Percentile Time Point & $p(d x_0)$	Sample Size (n)	$L_1(t x_0) (10^{-2})$					
			No Censoring		25% Censoring Rate		50% Censoring Rate	
			Proposed Method	Cox's Method	Scheike & Zhang's Method	Proposed Method	Cox's Method	Scheike & Zhang's Method
State-2	75% 0.4434	1000	2.85	1.59	4.36	2.90	1.76	5.20
		100	5.25	5.51	8.51	6.21	5.58	9.28
		500	3.10	2.43	4.99	3.21	2.98	5.25
		1000	2.43	1.68	3.25	2.66	2.45	4.74
	25% 0.2500	100	7.82	5.78	5.53	8.87	6.94	6.07
		500	4.42	2.70	2.65	4.90	2.94	2.88
		1000	3.57	2.42	2.12	3.88	2.49	2.29
		100	6.48	5.18	8.02	7.71	6.51	8.93
	50% 0.5000	500	3.48	2.41	4.85	3.58	2.75	5.07
		1000	2.63	1.68	3.20	2.76	2.02	4.28
		100	5.72	5.22	6.98	6.34	5.85	9.01
		500	3.24	2.45	3.56	3.36	2.67	3.93
		1000	2.53	1.98	2.79	2.66	2.04	3.21
State-1	75% 0.7500	1000	2.53	1.98	2.79	2.66	2.04	3.21
		100	5.72	5.22	6.98	6.34	5.85	9.01
		500	3.24	2.45	3.56	3.36	2.67	3.93
		1000	2.53	1.98	2.79	2.66	2.04	3.21
	50% 0.5000	1000	2.63	1.68	3.20	2.76	2.02	4.28
		100	5.72	5.22	6.98	6.34	5.85	9.01
		500	3.48	2.41	4.85	3.58	2.75	5.07
		1000	2.63	1.68	3.20	2.76	2.02	4.28
	25% 0.2500	1000	3.57	2.42	2.12	3.88	2.49	2.29
		100	7.82	5.78	5.53	8.87	6.94	6.07
		500	4.42	2.70	2.65	4.90	2.94	2.88
		1000	2.43	1.68	3.25	2.66	2.45	4.74

L_1 distances of estimated state conditional occupying probabilities using data generated for perturbed SIM models, at 25th, 50th, and 75th percentiles of state entry (states - 1, 2) and exit (state - 0) times, with 500 Monte-Carlo replicates, with 25% random censoring. The value of the ϑ is ranging from 0 to 1.25.

Table 3

ϑ	(10 ⁻²)											
	State-0			State-1			State-2					
	25%	50%	75%	25%	50%	75%	25%	50%	75%	25%	50%	75%
0	3.378	4.317	4.441	2.867	3.776	4.448	3.506	4.245	4.326			
0.50	3.969	5.353	5.437	3.073	4.387	4.656	4.112	6.185	6.354			
0.75	5.089	7.037	6.575	3.225	5.221	4.733	4.767	8.221	8.225			
1.00	5.558	8.234	8.298	3.889	5.882	5.140	5.904	10.190	10.417			
1.25	6.860	9.576	9.497	4.534	6.500	5.471	6.735	10.469	11.325			

Table 4
Matrix showing the state-to-state transition counts for the Bone Marrow Transplant data

To								
From	0	1	2	3	4	5	6	
0	0	7	117	0	0	1	12	
1		0	0	3	0	2	2	
2			20	0	19	44	34	
3				0	0	1	2	
4					2	11	6	
5						32	27	
6							83	

Table 5

Mean absolute difference between conditional state occupation probabilities of two hypothetical individuals with covariate vectors: $(x_1, x_2) = (20, 20)$ and $(x_1, x_2) = (40, 40)$, using the proposed method. x_1 and x_2 represent patient's age and donor's age, respectively. The corresponding p -values are in parenthesis.

State	$D (10^{-2})$
0: Bone Marrow Transplant	0.533 (0.65)
1: Acute GVHD	0.854 (0.15)
2: Platelet Recovery	3.457 (0.13)
3: Platelet Recovery After Acute GVHD	1.710 (0.03)
4: Acute GVHD After Platelet Recovery	2.152 (0.08)
5: Chronic GVHD	3.506 (0.15)
6: Relapse/Death	1.843 (0.42)

Table 6

Matrix showing the state-to-state transition counts for the Spinal Code Injury data

From	To				
	0	1	2	3	4
0	79	47	10	0	0
1		68	38	16	0
2			39	33	7
3				55	27
4					32