

Landmark estimation of survival and treatment effects in observational studies

Layla Parast¹ · Beth Ann Griffin¹

Received: 18 May 2015 / Accepted: 12 January 2016 / Published online: 15 February 2016
© RAND Corporation 2016

Abstract Clinical studies aimed at identifying effective treatments to reduce the risk of disease or death often require long term follow-up of participants in order to observe a sufficient number of events to precisely estimate the treatment effect. In such studies, observing the outcome of interest during follow-up may be difficult and high rates of censoring may be observed which often leads to reduced power when applying straightforward statistical methods developed for time-to-event data. Alternative methods have been proposed to take advantage of auxiliary information that may potentially improve efficiency when estimating marginal survival and improve power when testing for a treatment effect. Recently, Parast et al. (J Am Stat Assoc 109(505):384–394, 2014) proposed a landmark estimation procedure for the estimation of survival and treatment effects in a randomized clinical trial setting and demonstrated that significant gains in efficiency and power could be obtained by incorporating intermediate event information as well as baseline covariates. However, the procedure requires the assumption that the potential outcomes for each individual under treatment and control are independent of treatment group assignment which is unlikely to hold in an observational study setting. In this paper we develop the landmark estimation procedure for use in an observational setting. In particular, we incorporate inverse probability of treatment weights (IPTW) in the landmark estimation procedure to account for selection bias on observed baseline (pretreatment) covariates. We demonstrate that consistent estimates of survival and treatment effects can be obtained by using IPTW and that there is improved efficiency by using auxiliary intermediate event and baseline informa-

Electronic supplementary material The online version of this article (doi:[10.1007/s10985-016-9358-z](https://doi.org/10.1007/s10985-016-9358-z)) contains supplementary material, which is available to authorized users.

✉ Layla Parast
parast@rand.org

¹ RAND Corporation, 1776 Main Street, Santa Monica, CA 90403, USA

tion. We compare our proposed estimates to those obtained using the Kaplan–Meier estimator, the original landmark estimation procedure, and the IPTW Kaplan–Meier estimator. We illustrate our resulting reduction in bias and gains in efficiency through a simulation study and apply our procedure to an AIDS dataset to examine the effect of previous antiretroviral therapy on survival.

Keywords Treatment effect · Survival analysis · Intermediate event · Nonparametric · Robust

1 Introduction

While randomized studies are the gold standard for estimating treatment effectiveness, there are numerous occasions when they are not feasible. Moreover, there are numerous times when meaningful information is available from observational studies regarding the potential effectiveness of a particular treatment on an outcome. Unfortunately, with rare diseases or outcomes, observational and clinical studies aimed at identifying effective treatments to reduce the risk of disease or death often require long term follow-up of participants in order to observe a sufficient number of events to precisely estimate the treatment effect. In such studies, observing the outcome of interest during follow-up may be difficult and high rates of censoring may be observed which often leads to reduced power when applying straightforward statistical methods developed for time-to-event data.

In light of these challenges, alternative methods have been proposed to take advantage of auxiliary information that may potentially improve efficiency when estimating marginal survival and improve power when testing for a treatment effect in a randomized study (Cook and Lawless 2001). For example, when the available auxiliary information consists of a single discrete variable, fully nonparametric approaches (Rotnitzky and Robins 2005; Murray and Tsiatis 1996) that incorporate this variable when estimating marginal survival have been shown to produce more efficient estimates when compared to the Kaplan–Meier estimator (Kaplan and Meier 1958). When the auxiliary information includes continuous variables and/or multiple variables, semi parametric and parametric approaches such as regression adjustment are often considered. However, while these methods can be used to improve efficiency, they often rely on correct model specification. For example, the Cox proportional hazards model (Cox 1972) incorporating baseline covariates is often used to obtain an estimate of marginal survival and test for a treatment effect but the validity and performance of this approach also depends on the correct specification of the Cox model (Lagakos 1988; Lagakos and Schoenfeld 1984; Lin and Wei 1989).

A promising alternative to regression adjustment methods that has gained much recent attention is augmentation approaches which generally involve an augmentation term that is a function of the auxiliary information (Lu and Tsiatis 2008; Garcia et al. 2011; Tian et al. 2012; Zhang 2015; Zhang et al. 2008; Parast et al. 2014). For example, Lu and Tsiatis (2008) proposed an augmentation procedure to improve the efficiency of estimating the log hazard ratio from the Cox model, and demonstrated substantial gains in power when compared to the standard log-rank test. Garcia et al. (2011) used

a similar covariate augmentation approach to improve efficiency when using a more general class of survival models and Zhang (2015) developed augmented versions of the Nelson–Aalen and Kaplan–Meier estimators.

When auxiliary information consists of information collected over time such as repeated measurements after baseline or the occurrence of an intermediate event, incorporating this information to improve efficiency becomes more difficult due to the semi-competing risks nature of the data (Fine et al. 2001). That is, when the primary outcome is a terminal event such as death and the intermediate event is a non-terminal event such as hospitalization or cancer recurrence, the occurrence of the terminal event would censor the non-terminal event but not vice versa. Therefore, if an individual dies before the intermediate event occurs or before the repeated measurements are obtained, this auxiliary information is not available for that individual. Recently, Parast et al. (2014) proposed a landmark estimation procedure that uses a landmarking approach to overcome these semi-competing risk issues. Specifically, this procedure incorporates intermediate event information observed up to a landmark time, t_0 , for those who have survived and are still under observation at t_0 , in the estimation of marginal survival and a treatment effect. In addition, a smoothing component of the landmark estimation procedure ensures the consistency of survival estimates and thus these estimates do not require one to correctly specify a model relating the intermediate event to the primary outcome. Parast et al. (2014) demonstrated that significant gains in efficiency can be obtained. Other previously proposed methods to improve efficiency by using intermediate information include a kernel estimation approach (Gray 1994), a three-state model approach (Finkelstein and Schoenfeld 1994), an augmented score and augmented likelihood approach (Fleming et al. 1994), a multiple imputation approach (Faucett et al. 2002), a nonparametric approach (Murray and Tsiatis 1996, 2001), and a targeted shrinkage regression approach (Li et al. 2011).

While the methods described above allow for increased efficiency and power through the use of auxiliary information, they are generally not valid in observational study settings. That is, these methods require the assumption that the potential outcomes for each individual under treatment and control are independent of treatment group assignment, an assumption that holds in a randomized clinical trial setting but is very unlikely to hold in an observational setting. When this assumption is violated, methods that do not account for this “selection” bias can result in biased estimates of survival and treatment effectiveness.

There are a number of statistical methods available that attempt to account for potential selection bias including regression adjustment, matching methods, and inverse probability of treatment (IPT) weighting (or propensity score weighting). The goal of such methods is to estimate survival and treatment effects appropriately adjusting for the fact that individuals in one treatment group may differ from those in another group on factors other than treatment alone. In the case of IPT weighting, an average treatment effect in the population can be estimated by re-weighting individuals based on their probability of treatment such that the treatment groups are, in essence, balanced on all observed factors other than treatment (Hernán et al. 2000; Rosenbaum and Rubin 1983b, 1984). Xie and Liu (2005) proposed an IPT weighted Kaplan–Meier estimate of survival and a corresponding test statistic to test for a difference in survival distributions and showed that consistent estimates that account for selection bias can

be obtained. Other methods based on weighting and stratification include Nieto and Coresh (1996) and Amato (1988) where the general approach is to stratify individuals by the observed confounders, estimate survival in each strata, and appropriately combine the resulting survival estimates. Alternatively, survival estimates can be adjusted for observed confounders and compared using a specified regression model such as the Cox model, but as in the case where one aims to gain efficiency by using a Cox model, when the model is not correctly specified the resulting estimates may not be valid (Thomsen et al. 1991; Therneau 2000; Chen and Tsiatis 2001). A number of doubly robust estimators that combine IPT weights (IPTW) and a model for survival, often a Cox regression model, have been proposed and lead to consistent estimates when either the model used to obtain the IPTW or the regression model is correct (Zhang and Schaubel 2012a, b; Bai et al. 2013).

While these previously developed time-to-event methods provide valuable tools for inference in an observational setting, methods that can improve efficiency through the use of auxiliary information that includes intermediate event information and are valid in an observational setting are still lacking. In this paper we develop the landmark estimation procedure of Parast et al. (2014) for use in an observational setting such that one can obtain consistent estimates of survival and a treatment effect with improved efficiency by taking advantage of baseline and intermediate event auxiliary information. We compare our proposed estimates to those obtained using the Kaplan–Meier estimator, the original landmark estimation procedure (which one would expect to be biased as selection bias is not accounted for), and the IPT weighted Kaplan–Meier estimator (which we expect to be unbiased but less efficient since auxiliary information is not incorporated). We illustrate the resulting reduction in bias and gains in efficiency through a simulation study and apply our procedure to an AIDS dataset to examine the effect of previous antiretroviral therapy on survival.

2 Estimation of survival in an observational study

2.1 Notation and potential outcomes framework

For the i th subject, let T_{Li} denote the time of the primary event of interest, \mathbf{T}_{Si} denote the vector of intermediate event times, \mathbf{Z}_i denote the vector of baseline (pretreatment) covariates, and C_i denote the censoring time assumed independent of $(T_{Li}, \mathbf{T}_{Si}, \mathbf{Z}_i)$. Due to censoring, T_{Li} and \mathbf{T}_{Si} are only potentially observed. Instead, we observe $X_{Li} = \min(T_{Li}, C_i)$, $\mathbf{X}_{Si} = \min(\mathbf{T}_{Si}, C_i)$ and $\delta_{Li} = I(T_{Li} \leq C_i)$, $\delta_{Si} = I(\mathbf{T}_{Si} \leq C_i)$. When T_{Li} is a terminal event, such as death, this would represent a semi-competing risks setting where \mathbf{T}_{Si} is additionally subject to *informative* censoring by T_{Li} , while T_{Li} is only subject to administrative censoring and cannot be censored by \mathbf{T}_{Si} . Let t_0 denote some landmark time prior to t , such as a 1-year check up time following disease diagnosis. Our goal is to estimate $S(t) = P(T_{Li} > t)$ appropriately using baseline covariate information and intermediate event information collected up to t_0 , where t is a clinically relevant pre-specified time point such that $P(X_{Li} > t \mid T_{Li} \geq t_0) \in (0, 1)$ and $P(T_{Li} \leq t_0, T_{Si} \leq t_0) \in (0, 1)$.

In order to rigorously define the survival and treatment effect quantities we aim to estimate, we consider a potential outcomes framework. Assume there are two treatments, Treatment 1 and Treatment 0 and let $G_i = 1$ or 0 denote the treatment received by individual i . Each individual has two potential outcomes: T_{L1i} , which is the time of the long term event after receiving treatment 1, and T_{L0i} , which is the time of the long term event after receiving treatment 0. However, in reality we only observe one of these outcomes for each patient $T_{Li} = T_{L1i}I(G_i = 1) + T_{L0i}I(G_i = 0)$. Due to censoring, we define $X_{L1i} = \min(T_{L1i}, C_i)$, $\mathbf{X}_{S1i} = \min(\mathbf{T}_{S1i}, C_i)$ and $\delta_{L1i} = I(T_{L1i} \leq C_i)$, $\delta_{S1i} = I(\mathbf{T}_{S1i} \leq C_i)$ or $X_{L0i} = \min(T_{L0i}, C_i)$, $\mathbf{X}_{S0i} = \min(\mathbf{T}_{S0i}, C_i)$ and $\delta_{L0i} = I(T_{L0i} \leq C_i)$, $\delta_{S0i} = I(\mathbf{T}_{S0i} \leq C_i)$. In essence, there are two levels of missing data in this framework. First, since individuals are assigned to only one treatment, only T_{L1i} , \mathbf{T}_{S1i} or T_{L0i} , \mathbf{T}_{S0i} are potentially observable. Second, we are additionally not able to observe T_{L1i} , \mathbf{T}_{S1i} for all individuals with $G_i = 1$ (and similarly T_{L0i} , \mathbf{T}_{S0i} for all individuals with $G_i = 0$) due to censoring.

2.2 Estimation of survival using the Kaplan–Meier estimator

We aim to estimate survival at time t within each treatment group, $S_1(t) = P(T_{L1} > t)$ and $S_0(t) = P(T_{L0} > t)$. To make our assumptions explicit we define:

Assumption A.1 $(T_{L1i}, \mathbf{T}_{S1i}, T_{L0i}, \mathbf{T}_{S0i}, \mathbf{Z}_i) \perp C_i \mid G_i$

Assumption A.2 $(T_{L1i}, T_{L0i}, \mathbf{T}_{S1i}, \mathbf{T}_{S0i}) \perp G_i \mid \mathbf{Z}_i$

Assumption A.1 assumes independent censoring and Assumption A.2 is often referred to as the assumption of no unmeasured confounders (Rosenbaum and Rubin 1983b) or the assumption of strong ignorability (Robins et al. 2000). Without loss of generality, we first focus on estimation of $S_1(t)$. In a randomized clinical trial (RCT) setting, instead of Assumption A.2, one could make the much stronger assumption that $(T_{L1i}, T_{L0i}, \mathbf{T}_{S1i}, \mathbf{T}_{S0i}) \perp G_i$ which would hold due to random treatment assignment. In such a randomized setting, a common nonparametric approach to estimate survival is the Kaplan–Meier (KM) estimate (Kaplan and Meier 1958),

$$\widehat{S}_{KM,j}(t) = \begin{cases} 1 & \text{if } t < t_{1j} \\ \prod_{t_{kj} \leq t} \left[1 - \frac{d_{kj}}{y_{kj}} \right] & \text{if } t \geq t_{1j} \end{cases} \tag{1}$$

where t_{1j}, \dots, t_{Dj} are the distinct observed long term event times in treatment group j , d_{kj} is the number of events at time t_{kj} in treatment group j , and y_{kj} is the number of patients at risk at t_{kj} in treatment group j .

However, in an observation study where treatment is not randomized, one cannot assume that

$(T_{L1i}, T_{L0i}, \mathbf{T}_{S1i}, \mathbf{T}_{S0i}) \perp G_i$. Indeed, individual characteristics that may be associated with treatment may also be associated with the potential outcome. For example, if the exposure of interest was diabetes and the long term outcome was death, individual characteristics such as age, body mass index, gender and diet may be associated with both the likelihood of having diabetes and survival. Analyses which ignore selection

bias (i.e. that the distribution of confounders differ in the two treatment groups) can result in biased estimates of treatment effectiveness, particularly if treatment selection is related to treatment effectiveness or the primary long term event of interest. However, it may be possible to identify such individual characteristics and appropriately adjust methods originally developed for an RCT setting accordingly. Specifically, if \mathbf{Z}_i contains all individual characteristics that may be associated with both treatment and the outcome, then among individuals with the same \mathbf{Z}_i , treatment group and the potential outcomes would be independent (Assumption A.2). Therefore, methods that appropriately account for the differential distribution of \mathbf{Z}_i within treatment groups will lead to valid estimation of the quantities of interest (Rosenbaum and Rubin 1983b).

Methods that take advantage of this assumption to estimate survival and treatment effects in the presence of selection bias include regression adjustment and IPT weighting. IPT weighting involves appropriately weighting estimates or estimating equations by the inverse of the probability of treatment or the propensity score, $W_j(\mathbf{Z}_i) = P(G_i = j | \mathbf{Z}_i)$, the probability of being in treatment group j given individual characteristics. It has been shown that when Assumption A.2 holds and $W_j(\mathbf{Z}_i)$ is known or can be consistently estimated, $T_{\perp Li}, \mathbf{T}_{\perp Si}, T_{\perp Oi}, \mathbf{T}_{\perp Si} \perp G_i | W_j(\mathbf{Z}_i)$ (Rosenbaum and Rubin 1983b, 1984; Hernán et al. 2000). That is, among individuals with the same propensity score, treatment and the potential outcomes are independent. A particular example of an IPT weighted estimator in our setting is the IPT weighted Kaplan–Meier (IPTW KM) estimator (Xie and Liu 2005) of $S_j(t)$:

$$\widehat{S}_{IPTW,j}(t) = \begin{cases} 1 & \text{if } t < t_{1j} \\ \prod_{t_{kj} \leq t} \left[1 - \frac{d_{kj}^w}{y_{kj}^w} \right] & \text{if } t \geq t_{1j} \end{cases}$$

where $d_{kj}^w = \sum_{i: X_{\perp Li} = t_{kj}, \delta_{\perp Li} = 1} \widehat{W}_j(\mathbf{Z}_i)^{-1} \delta_{\perp Li} I(G_i = j)$ and $y_{kj}^w = \sum_{i: X_{\perp Li} \geq t_{kj}} \widehat{W}_j(\mathbf{Z}_i)^{-1} I(G_i = j)$, $W_j(\mathbf{Z}_i) = P(G_i = j | \mathbf{Z}_i)$, and $\widehat{W}_j(\mathbf{Z}_i)$ is the estimated propensity score.

2.3 Landmark estimation of survival in an observational study

In this section we aim to develop the landmark estimation procedure of Parast et al. (2014) in the potential outcomes framework such that bias resulting from selection bias would be eliminated and estimates obtained would provide improved efficiency compared to the IPTW KM estimate by incorporating baseline and intermediate event information. As in Parast et al. (2014), we note that for $t > t_0$, $S_1(t) = P(T_{\perp Li} > t)$ can be expressed as $S_1(t | t_0)S_1(t_0)$, where

$$S_1(t | t_0) = P(T_{\perp Li} > t | T_{\perp Li} > t_0) \quad \text{and} \quad S_1(t_0) = P(T_{\perp Li} > t_0).$$

In essence, we aim to incorporate intermediate event information in estimation of $S_1(t | t_0)$ to improve the efficiency of the overall estimate of $S(t)$, but we desire an approach that (a) does not require that we correctly specify the relationship between the intermediate event and the primary outcome since any specified model is unlikely to hold in practice and (b) accounts for selection bias. Throughout, we assume that t_0 is

pre-selected and fixed, however we discuss the selection of t_0 further in the Discussion. We first focus on obtaining a consistent estimate of $S_1(t | t_0)$ and note that,

$$S_1(t|t_0) = P(T_{L1} > t | T_{L1} > t_0) = E\{P(T_{L1} > t | T_{L1} > t_0, \mathbf{H}_1)\} = E\{S_1(t|t_0, \mathbf{H}_1)\} \tag{2}$$

where $S_1(t|t_0, \mathbf{H}_1) = P(T_{L1} > t | T_{L1} > t_0, \mathbf{H}_1)$ and $\mathbf{H}_1 = \{\mathbf{Z}, I(\mathbf{T}_{S1} \leq t_0), \min(\mathbf{T}_{S1}, t_0)\}$. That is, \mathbf{H}_1 contains all information that is potentially observable up to the landmark time, t_0 , for an individual who has survived to t_0 and could include information on multiple intermediate events and/or covariates with repeated measurements before t_0 , if such data were available. Note that \mathbf{H}_1 is only observable for those with $G_i = 1$ and $X_{L1i} > t_0$. If one were able to obtain a consistent estimate of $S_1(t | t_0, \mathbf{H}_1)$, denoted by $\widehat{S}_1(t|t_0, \mathbf{H}_1)$, then one could estimate $S_1(t | t_0)$ by

$$\widehat{S}_1(t|t_0) = \frac{n^{-1} \sum_{i=1}^n \widehat{W}_1(\mathbf{Z}_i)^{-1} \widehat{S}_1(t|t_0, \mathbf{H}_{1i}) I(G_i = 1) I(X_{L1i} > t_0)}{n^{-1} \sum_{i=1}^n \widehat{W}_1(\mathbf{Z}_i)^{-1} I(G_i = 1) I(X_{L1i} > t_0)}. \tag{3}$$

We will now show that we may obtain such a consistent estimate, $\widehat{S}_1(t|t_0, \mathbf{H}_1)$, of $S_1(t | t_0, \mathbf{H}_1)$ by developing the two-stage procedure in Parast et al. (2014) for use in a setting where selection bias is a concern using IPTW. We first reduce the dimension of \mathbf{H}_1 by approximating $S_1(t | t_0, \mathbf{H}_1)$ with a working semiparametric model, the landmark proportional hazards model (Van Houwelingen and Putter 2012)

$$S_1(t | t_0, \mathbf{H}_1) = \exp \left\{ -\Lambda_0^{t_0}(t) \exp(\boldsymbol{\beta}_1^\top \mathbf{H}_1) \right\}, \quad t > t_0 \tag{4}$$

where $\Lambda_0^{t_0}(\cdot)$ is the unspecified baseline cumulative hazard function for T_{L1i} among $\Omega_{t_0,1} = \{X_{L1i} > t_0, G_i = 1\}$ and $\boldsymbol{\beta}_1$ is an unknown vector of coefficients. Let $\widehat{\boldsymbol{\beta}}_1$ be the maximizer of the IPT weighted log partial likelihood function,

$$\begin{aligned} \widehat{\ell}_{t_0}(\boldsymbol{\beta}_1) &= \sum_{i \in \Omega_{t_0,1}} \delta_{L1i} W_1(\mathbf{Z}_i)^{-1} \\ &\times \left[\boldsymbol{\beta}_1^\top \mathbf{H}_{1i} - \log \left\{ \sum_{j \in \Omega_{t_0,1}} W(\mathbf{Z}_j)^{-1} e^{\boldsymbol{\beta}_1^\top \mathbf{H}_{1j}} I(X_{L1j} > X_{L1i},) \right\} \right]. \end{aligned} \tag{5}$$

In an effort to obtain a final estimate that is robust to model misspecification, we avoid the assumption that this landmark proportional hazards model is correctly specified by focusing only on the resulting risk score $\widehat{U}_{1i} \equiv \widehat{\boldsymbol{\beta}}_1^\top \mathbf{H}_{1i}$. That is, instead of aiming to obtain an estimate of $S_1(t | t_0, \mathbf{H}_1) = P(T_{L1} > t | T_{L1} > t_0, \mathbf{H}_1)$ in (2) and (3), we now change our focus to obtaining an estimate of $S_1(t | t_0, U_1) = P(T_{L1} > t | T_{L1} > t_0, U_1)$ where $U_1 = \boldsymbol{\beta}_{10}^\top \mathbf{H}_1$ and $\boldsymbol{\beta}_{10}$ is the limit of $\widehat{\boldsymbol{\beta}}_1$. Note that the derivation supporting (2) and (3) would still hold when $S_1(t | t_0, \mathbf{H}_1)$ and $\widehat{S}_1(t | t_0, \mathbf{H}_1)$ are replaced by $S_1(t | t_0, U_1)$ and $\widehat{S}_1(t | t_0, \widehat{U}_1)$, a consistent estimate of $S_1(t | t_0, U_1)$, respectively. In this first stage, the working model is essentially used as a tool to reduce the dimension of \mathbf{H} by constructing \widehat{U} .

In the second stage, we derive $\widehat{S}_1(t|t_0, \widehat{U}_1)$, such that an estimate of $S_1(t | t_0)$ can then be obtained as (3) with $\widehat{S}_1(t | t_0, \mathbf{H}_1)$ replaced by $\widehat{S}_1(t | t_0, \widehat{U}_1)$. We propose to use an IPT weighted nonparametric conditional Nelson–Aalen estimator (Beran 1981) based on subjects in $\Omega_{t_0,1}$ to obtain an estimate of $S_1(t | t_0, U_1)$. Specifically for any given t and u , the synthetic data $\{(X_{L1i}, \delta_{L1i}, \widehat{U}_{1i}), i \in \Omega_{t_0,1}\}$ is used to calculate the IPT weighted local constant estimator for the conditional hazard $\Lambda_1(t | t_0, u) = -\log S_1(t | t_0, u)$ as

$$\widehat{\Lambda}_1(t|t_0, u) = \int_{t_0}^t \frac{\sum_{i \in \Omega_{t_0,1}} \widehat{W}_1(\mathbf{Z}_i)^{-1} K_h(\widehat{U}_{1i} - u) dN_i(z)}{\sum_{i \in \Omega_{t_0,1}} \widehat{W}_1(\mathbf{Z}_i)^{-1} K_h(\widehat{U}_{1i} - u) Y_i(z)}$$

where $Y_i(t) = I(T_{L1i} \geq t)$, $N_i(t) = I(T_{L1i} \leq t)\delta_{L1i}$, $K(\cdot)$ is a smooth symmetric density function, $K_h(x) = K(x/h)/h$, and $h = O(n^{-\nu})$ is a bandwidth with $1/2 > \nu > 1/4$. The resulting estimate for $S_1(t | t_0, U_1)$ is $\widehat{S}_1(t | t_0, \widehat{U}_1) = \exp\{-\widehat{\Lambda}_1(t | t_0, \widehat{U}_1)\}$. Finally, $S_1(t | t_0)$ is estimated as (3) with $\widehat{S}_1(t|t_0, \mathbf{H}_1)$ replaced by $\widehat{S}_1(t | t_0, \widehat{U}_{1i}) = \exp\{-\widehat{\Lambda}_1(t|t_0, \widehat{U}_{1i})\}$.

Now that we have proposed an estimation procedure for $S_1(t|t_0)$, an estimate for $S_1(t_0)$ follows similarly from this same two-stage procedure replacing \mathbf{H} with \mathbf{Z} and $\Omega_{t_0,1}$ with $\Omega = \{G_i = 1\}$. Specifically, we can obtain an estimate for $S_1(t_0)$ as

$$\widehat{S}_1(t_0) = \frac{\frac{1}{n_1} \sum \widehat{W}_1(\mathbf{Z}_i)^{-1} \widehat{S}_1(t_0|\mathbf{Z}_i) I(G_i = 1)}{\frac{1}{n_1} \sum \widehat{W}_1(\mathbf{Z}_i)^{-1} I(G_i = 1)} \tag{6}$$

where $\widehat{S}_1(t_0|\mathbf{Z}_i)$ is a consistent estimate of $P(T_{L1} > t_0|\mathbf{Z}_i)$. To obtain $\widehat{S}_1(t_0|\mathbf{Z}_i)$, we use the two stage estimation procedure to obtain a risk score \widehat{U}_{1i}^* in the first stage and smooth over \widehat{U}_{1i}^* to obtain $\widehat{\Lambda}_1(t_0 | \widehat{U}_{1i}^*)$ such that $\widehat{S}_1(t_0 | \widehat{U}_{1i}^*) = \exp\{-\widehat{\Lambda}_1(t_0 | \widehat{U}_{1i}^*)\}$ is a consistent estimator of $S_1(t_0 | U_1^*) = P(T_{L1i} > t_0|U_1^*)$ where $U_1^* = \beta_{10}^{*T} \mathbf{Z}_i$ and β_{10}^* is the limit of $\widehat{\beta}_1^*$, the maximizer of the weighted Cox partial likelihood corresponding to the working model,

$$S_1(t_0|\mathbf{Z}_1) = \exp\{-\Lambda_0(t) \exp(\beta_1^{*T} \mathbf{Z}_1)\}, \tag{7}$$

which uses only \mathbf{Z} where $\Lambda_0(\cdot)$ is the unspecified baseline cumulative hazard function for T_{L1i} , calculated in stage 1.

An estimate for the primary quantity of interest $S_1(t)$ in an observational study incorporating intermediate event and covariate information collected up to t_0 follows as $\widehat{S}_{LM,1}(t) \equiv \widehat{S}_1(t | t_0)\widehat{S}_1(t_0)$ where LM indicates that a landmark time, t_0 , has been used to decompose the estimate into two components. The estimate for $S_0(t)$ follows similarly and is denoted as $\widehat{S}_{LM,0}(t)$. The consistency of $S_j(t)$ follows from the consistency of $\widehat{S}_j(t | t_0)$ and $\widehat{S}_j(t_0)$. The consistency of $\widehat{S}_j(t | t_0)$ for $S_j(t | t_0)$ and $\widehat{S}_j(t_0)$ for $S_j(t_0)$ is ensured by Assumption A.1 and Assumption A.2, the assumption that the propensity scores $\widehat{W}_j(\mathbf{Z}_i)$ are consistent estimates for $W_j(\mathbf{Z}_i)$, the consistency of $\widehat{\beta}_j$ and $\widehat{\beta}_j^*$ for some constants β_{j0} and β_{j0}^* , respectively, even under misspecification of (4) and (7) (Lin 2000; Lin and Wei 1989; Pan and Schaubel 2008),

and the uniform consistency of $\widehat{S}_j(t | t_0, \widehat{U}_j)$ and $\widehat{S}_j(t_0, \mathbf{Z})$ which can be shown using similar arguments as in [Cai et al. \(2010\)](#), [Du and Akritas \(2002\)](#), and [Parast et al. \(2014\)](#) under mild regularity conditions. We discuss the assumption concerning the consistency of $\widehat{W}_j(\mathbf{Z}_i)$ further in the Discussion.

It is worth noting that a similar two-stage approach could be used to gain efficiency even if one only has baseline covariates and no intermediate event information. That is, an estimate of $S(t)$ incorporating only baseline covariate information, \mathbf{Z} , can be obtained as in (6) with t_0 replaced by t . With this approach, no landmarking is used and only a single working model specifying the relationship between T_{Lj} and \mathbf{Z} for $j = 0, 1$ is needed. In our numerical studies, we calculate this estimate to shed light on how much of our observed efficiency gain is due to intermediate event information versus \mathbf{Z} information alone.

3 Estimation of the treatment effect in an observational study

We aim to estimate the average treatment effect (ATE) in terms of a difference in survival at time t . That is, the treatment effect is defined as the risk difference, $\Delta(t) = S_1(t) - S_0(t)$. Using landmark estimation in an observational setting, we may obtain $\widehat{\Delta}_{LM}(t) = \widehat{S}_{LM,1}(t) - \widehat{S}_{LM,0}(t)$ since $\widehat{S}_{LM,j}(t)$ is a consistent estimate of $S_j(t)$. The standard error of $\widehat{\Delta}_{LM}(t)$ can be estimated as $\widehat{\sigma}(\widehat{\Delta}_{LM}(t))$ using a perturbation-resampling procedure as described in Sect. 4. A normal confidence interval (CI) for $\Delta(t)$ may be constructed accordingly. To test the null hypothesis of $H_0 : \Delta(t) = 0$, a Wald-type test may be performed based on $\widehat{Z}_{LM}(t) = \widehat{\Delta}_{LM}(t) / \widehat{\sigma}(\widehat{\Delta}_{LM}(t))$. To examine bias and efficiency in estimation of a treatment effect, we compare this testing procedure to a test based on (1) the KM estimate in an RCT setting, $\widehat{\Delta}_{KM}(t) = \widehat{S}_{KM,1}(t) - \widehat{S}_{KM,0}(t)$, (2) the landmark estimation procedure for an RCT setting $\widehat{\Delta}_{LM}^{RCT}(t) = \widehat{S}_{LM,1}^{RCT}(t) - \widehat{S}_{LM,0}^{RCT}(t)$, and (3) the IPTW KM estimate $\widehat{\Delta}_{IPTW}(t) = \widehat{S}_{IPTW,1}(t) - \widehat{S}_{IPTW,0}(t)$, where $\widehat{S}_{LM,j}^{RCT}(t)$ is the estimate of survival for treatment group j obtained using the landmark estimation procedure for an RCT setting.

4 Variance estimation using perturbation-resampling

To obtain variance estimates, we use a perturbation-resampling method ([Park and Wei 2003](#); [Cai et al. 2005](#); [Tian et al. 2007](#)). Specifically, let $\{\mathbf{V}^{(b)} = (V_1^{(b)}, \dots, V_n^{(b)})^T, b = 1, \dots, B\}$ be $n \times B$ independent copies of a positive random variable U from a known distribution with unit mean and unit variance such as an $\text{Exp}(1)$ distribution. To estimate the variance of our proposed procedure, for $j = 0, 1$, let

$$\begin{aligned} & \widehat{S}_{LM,j}(t|t_0)^{(b)} \\ &= \frac{n_{t_0,j}^{-1} \sum_{i \in \Omega_{t_0,j}} [\widehat{W}_j(\mathbf{Z}_i)^{(b)}]^{-1} \exp \left\{ -\widehat{\Lambda}_j^{(b)} \left(t|t_0, \widehat{U}_{ji}^{(b)} \right) \right\} I(G_i = j) I(X_{Lji} > t_0) V_i^{(b)}}{n_{t_0,j}^{-1} \sum_{i \in \Omega_{t_0,j}} [\widehat{W}_j(\mathbf{Z}_i)^{(b)}]^{-1} I(G_i = j) I(X_{Lji} > t_0) V_i^{(b)}} \end{aligned}$$

where

$$\widehat{\Lambda}_j^{(b)}(t|t_0, \widehat{U}_{ji}^{(b)}) = \int_{t_0}^t \frac{\sum_{i \in \Omega_{t_0, j}} V_i^{(b)} [\widehat{W}_j(\mathbf{Z}_i)^{(b)}]^{-1} K_h(\widehat{U}_{ji}^{(b)} - u) dN_i(z)}{\sum_{i \in \Omega_{t_0, j}} V_i^{(b)} [\widehat{W}_j(\mathbf{Z}_i)^{(b)}]^{-1} K_h(\widehat{U}_{ji}^{(b)} - u) Y_i(z)},$$

$\widehat{U}_{ji}^{(b)} = \widehat{\beta}_j^{(b)} \mathbf{H}_{ji}$ and $\widehat{\beta}_j^{(b)}$ is the solution to (5) but with additional weights $V_i^{(b)}$ and $\widehat{W}_j(\mathbf{Z}_i)^{(b)} = \widehat{P}^{(b)}(G_i = j | \mathbf{Z}_i)$ where $\widehat{P}^{(b)}(G_i = j | \mathbf{Z}_i)$ is obtained using weights $V_i^{(b)}$. For example, if $\widehat{W}_j(\mathbf{Z}_i)$ is estimated using logistic regression, the perturbed version is estimated using weighted logistic regression with weights $V_i^{(b)}$. Similarly, $\widehat{S}_j^{(b)}(t_0)$ can be obtained by replacing $\mathbf{H}_i = \mathbf{Z}_i$ throughout and using all patients $\{G_i = j\}$. We now let $\widehat{S}_{LM, j}^{(b)}(t) \equiv \widehat{S}_j^{(b)}(t | t_0) \widehat{S}_j^{(b)}(t_0)$ and estimate the variance of $\widehat{S}_{LM, j}(t)$ as the empirical variance of $\{\widehat{S}_{LM, j}^{(b)}(t), b = 1, \dots, B\}$. This procedure can be used to obtain $\widehat{\Delta}_{LM}^{(b)}(t) = \widehat{S}_{LM, 1}^{(b)}(t) - \widehat{S}_{LM, 0}^{(b)}(t)$ for $b = 1, \dots, B$. Then one can estimate $\widehat{\sigma}(\widehat{\Delta}_{LM}(t))$ as the empirical variance of $\{\widehat{\Delta}_{LM}^{(b)}(t), b = 1, \dots, B\}$. In the numerical examples, we use this approach to obtain variance estimates for the standard KM estimator, the IPTW KM estimator, and the RCT version of the landmark estimator as well.

To construct $100(1 - \alpha)\%$ confidence intervals, one can either use the empirical percentiles of the perturbed samples (i.e., $100\alpha/2^{th}$ and $100(1 - \alpha/2)^{th}$ percentiles) or a normal approximation (i.e. $\widehat{S}_{LM, j}(t) \pm c\widehat{\sigma}_{LM, j}(t)$ where $\widehat{\sigma}_{LM, j}(t)$ is the empirical variance of $\{\widehat{S}_{LM, j}^{(b)}(t), b = 1, \dots, B\}$ and c is the $100(1 - \alpha/2)^{th}$ percentile of the standard normal distribution). The validity of the perturbation-resampling procedure can be shown using similar arguments as in Cai et al. (2010) and Zhao et al. (2010) since the distribution of $\sqrt{n}\{\widehat{S}_{LM, j}(t) - S_j(t)\}$ can be approximated by the distribution of $\sqrt{n}\{\widehat{S}_{LM, j}^{(b)}(t) - \widehat{S}_{LM, j}(t)\}$ conditional on the observed data.

5 Simulation study

We conducted simulation studies to examine the finite sample properties of the proposed estimation procedures. For illustration, $t_0 = 1$ year and $t = 2$ years i.e. we are interested in the probability of survival past 2 years. In all simulations, $W_j(\mathbf{Z}_i)$ is estimated using logistic regression, $n=2000$ for each treatment group and results summarize 1000 replications. The single baseline covariate Z was generated from a $N(1, 2)$ distribution in the treatment group and from a $N(0.5, 2)$ distribution in the control group. Censoring, C , was generated from a mixed distribution where $C = BC_1 + (1 - B)C_2, B \sim Bernoulli(0.5), C_1 \sim Exp(0.5),$ and $C_2 \sim Exp(0.9)$. In all settings, Assumption A.1 (censoring is independent of potential outcomes and Z) and Assumption A.2 (treatment group is independent of potential outcomes given Z) hold.

In simulation setting (i), there is no treatment effect, event times for the single intermediate event are generated as $T_S = \exp\{-Y + \epsilon_S\}$ where $Y \sim N(0.7, 4)$ and $\epsilon_S \sim N(0, 0.49)$ in both groups, and survival times are generated as $T_L = T_S + \exp\{(-2Z + \epsilon_L)/8\}$ where $\epsilon_L \sim N(1, 2.25)$ in both groups. That is, there is selection

bias through Z . This two-part distribution was selected to reflect a situation where the model describing the relationship between T_L , T_S and Z would be difficult to correctly specify. Note that in these simulations T_S occurs before T_L but our method does not require this to be true. In this setting, $P(T_{L1i} > t) = P(T_{L0i} > t) = 0.436$, about 61 % of individuals in the treatment group are censored, 63 % of individuals in the control group are censored, 39 % of individuals in the treatment group survive past t_0 , 41 % of individuals in the control group survive past t_0 , of those that survive past t_0 , 52 and 54 % have the intermediate event before t_0 in the treatment and control groups, respectively. In simulation setting (ii), there is a moderate treatment effect, event times for the single intermediate event and survival times for the control group are generated as in setting (i), event times for the single intermediate event in the treatment group are generated as $T_S = \exp\{-Y + \epsilon_S\}$ where $Y \sim N(0.7, 4)$ and $\epsilon_S \sim N(0.1, 0.49)$, and survival times in the treatment group are generated as $T_L = T_S + \exp\{(-1.5Z + \epsilon_L)/8\}$ where $\epsilon_L \sim N(2, 2.25)$. That is, treatment prolongs survival. In this setting, $P(T_{L0i} > t) = 0.436$, $P(T_{L1i} > t) = 0.483$, about 64 % of individuals in the treatment group are censored, 63 % of individuals in the control group are censored, 43 % of individuals in the treatment group survive past t_0 , 41 % of individuals in the control group survive past t_0 , of those that survive past t_0 , 55 and 54 % have the intermediate event before t_0 in the treatment and control groups, respectively.

In each setting, we estimate $S_j(t)$ in each group using the Kaplan–Meier estimate, $\widehat{S}_{KM,j}(t)$, the IPTW KM estimate, $\widehat{S}_{IPTW,j}(t)$, the landmark estimator developed in an RCT setting, $\widehat{S}_{LM,j}^{RCT}(t)$, and the landmark estimator proposed here, $\widehat{S}_{LM,j}(t)$. We summarize our simulation results in terms of the average estimate, bias, empirical standard error (the standard deviation of the 1000 estimates), average standard error (the average of the 1000 standard error estimates), mean squared error (the average of the 1000 squared bias estimates), relative efficiency (relative to the IPTW KM estimate), and coverage of the truth for the 1000 95 % confidence intervals. Table 1 shows the performance of the resulting survival estimates for the control group ($j = 0$), and for the treatment group ($j = 1$) in setting (i) and (ii). Note that only the distribution of the treatment group differs in setting (i) and (ii) and therefore the distribution of the control group is the same in both settings. Results show that estimates obtained using the standard Kaplan–Meier and the landmark estimation procedure for the randomized setting are biased, as expected. Estimates obtained using either the IPTW Kaplan–Meier estimate or the proposed landmark estimation procedure have very small bias and the proposed landmark estimation procedure provides improved efficiency with respect to the MSE ranging from 16–23 %. For the proposed landmark estimation procedure, standard error estimates obtained using perturbation-resampling procedure are close to the empirical estimates and coverage levels are close to the nominal 0.95 level. Table 2 shows the performance of the treatment effect estimates, $\widehat{\Delta}_{KM}(t)$, $\widehat{\Delta}_{IPTW}(t)$, $\widehat{\Delta}_{LM}^{RCT}(t)$, and $\widehat{\Delta}_{LM}(t)$ in both settings. Unweighted estimates, $\widehat{\Delta}_{KM}(t)$ and $\widehat{\Delta}_{LM}^{RCT}(t)$ have large bias in both settings, Type 1 error rates much larger than 0.05 in the null setting, and poor power in the moderate treatment effect setting. Both IPT weighted estimates have very small bias and Type 1 error rates close to 0.05 in the null setting. In terms of treatment effect estimation, the proposed landmark estimation procedure provides increased efficiency (24–28 %) compared to the IPTW KM estimate and improved power in setting (ii) (0.439 vs. 0.525). We also obtained

Table 1 Resulting survival estimates, $\widehat{S}_{KM,j}(t)$, $\widehat{S}_{IPTW,j}(t)$, $\widehat{S}_{LM,j}^{RCT}(t)$, and $\widehat{S}_{LM,j}(t)$ for $j = 0$ and 1 and corresponding bias, empirical standard error (ESE), average standard error (ASE), mean squared error (MSE), relative efficiency (RE) for the unbiased estimates only with respect to the IPTW KM estimator, and coverage (of 95 % confidence intervals) in the null treatment effect setting (i) and moderate treatment effect setting (ii); note that the control estimates in the moderate treatment effect setting are the same as the control estimates in the null treatment effect setting

	$\widehat{S}_{KM,0}(t)$	$\widehat{S}_{IPTW,0}(t)$	$\widehat{S}_{LM,0}^{RCT}(t)$	$\widehat{S}_{LM,0}(t)$
Null treatment effect setting (i): Control				
Truth	0.4363	0.4363	0.4363	0.4363
Estimate	0.4553	0.4365	0.4564	0.4377
Bias	0.0189	0.0002	0.0201	0.0013
ESE	0.0180	0.0178	0.0161	0.0160
ASE	0.0180	0.0177	0.0163	0.0160
MSE	0.0007	0.0003	0.0007	0.0003
RE	–	1.0000	–	1.2267
Coverage	0.8110	0.9450	0.7670	0.9480
	$\widehat{S}_{KM,1}(t)$	$\widehat{S}_{IPTW,1}(t)$	$\widehat{S}_{LM,1}^{RCT}(t)$	$\widehat{S}_{LM,1}(t)$
Null treatment effect setting (i): Treatment				
Truth	0.4363	0.4363	0.4363	0.4363
Estimate	0.4180	0.4367	0.4188	0.4375
Bias	–0.0184	0.0004	–0.0176	0.0012
ESE	0.0178	0.0182	0.0166	0.0168
ASE	0.0177	0.0180	0.0161	0.0162
MSE	0.0007	0.0003	0.0006	0.0003
RE	–	1.0000	–	1.1617
Coverage	0.8160	0.9410	0.7890	0.9430
Moderate treatment effect setting (ii): Treatment				
Truth	0.4830	0.4830	0.4830	0.4830
Estimate	0.4652	0.4832	0.4660	0.4840
Bias	–0.0178	0.0002	–0.0170	0.0010
ESE	0.0181	0.0183	0.0164	0.0164
ASE	0.0183	0.0185	0.0164	0.0165
MSE	0.0006	0.0003	0.0006	0.0003
RE	–	1.0000	–	1.2340
Coverage	0.8370	0.9460	0.8130	0.9520

estimates of survival and treatment effect using the two-stage approach incorporating baseline Z information only and observed efficiency gains in terms of MSE ranging from 5 to 7 % demonstrating that the efficiency gains of 24–28 % observed using the proposed approach can be attributed to incorporating *both* baseline and intermediate event information.

Table 2 Resulting treatment effect estimates, $\widehat{\Delta}_{KM}(t)$, $\widehat{\Delta}_{IPTW}(t)$, $\widehat{\Delta}_{LM}^{RCT}(t)$, and $\widehat{\Delta}_{LM}(t)$ and corresponding bias, empirical standard error (ESE), average standard error (ASE), mean squared error(MSE), relative efficiency (RE) for the unbiased estimates only with respect to the IPTW KM estimator, and Type 1 error in the null treatment effect setting (i) and power in the moderate treatment effect setting (ii)

	$\widehat{\Delta}_{KM}(t)$	$\widehat{\Delta}_{IPTW}(t)$	$\widehat{\Delta}_{LM}^{RCT}(t)$	$\widehat{\Delta}_{LM}(t)$
Setting (i): Null treatment effect				
Truth	0	0	0	0
Estimate	-0.0373	0.0002	-0.0376	-0.0002
Bias	-0.0373	0.0002	-0.0376	-0.0002
ESE	0.0253	0.0253	0.0227	0.0227
ASE	0.0253	0.0250	0.0229	0.0225
MSE	0.0020	0.0006	0.0019	0.0005
RE	-	1.0000	-	1.2419
Type 1 error	0.313	0.055	0.371	0.058
Setting (ii): Moderate treatment effect				
Truth	0.0467	0.0467	0.0467	0.0467
Estimate	0.0099	0.0466	0.0096	0.0463
Bias	-0.0368	-0.0001	-0.0371	-0.0004
ESE	0.0253	0.0251	0.0223	0.0222
ASE	0.0257	0.0254	0.0231	0.0227
MSE	0.0020	0.0006	0.0019	0.0005
RE	-	1.0000	-	1.2855
Power	0.072	0.439	0.065	0.525

6 Example

We illustrate the proposed procedures using a dataset from the acquired immunodeficiency syndrome (AIDS) Clinical Trial Group (ACTG) Protocol 175 (Hammer et al. 1996). This dataset consists of 2464 patients with human immunodeficiency virus (HIV) randomized to 4 different treatments: zidovudine only, zidovudine + didanosine, zidovudine + zalcitabine, and didanosine only. In the original study the goal was to compare the relative effectiveness of these four treatment conditions on time until progression to AIDS (measured by a 50% decline in CD4 cell counts) or death with death alone being a secondary end point. In this paper, we aim to examine the effect of previous antiretroviral treatment on time until death using patients from ACTG 175. Prior use of antiretrovirals (ART) was measured at baseline in the study for all study participants and it has been shown that previous antiretroviral therapy is highly predictive of survival. Results on the direction can vary. While ART itself is generally associated with improved survival rates, individuals receiving or who have received ART tend to be sicker which means that selection bias can result in that group showing worse survival (Hammer et al. 1997; Patel et al. 2008; Bhatta et al. 2013; Wood et al. 2003; Mocroft et al. 1999). Thus, it is important for any analysis involving prior antiretroviral to appropriately take into account the fact that patients who are anti-

retroviral naïve and patients with prior antiretroviral may differ from one another on a number of important baseline characteristics. We aim here to understand how such characteristics might bias the estimated relationship between prior antiretrovirals and survival in patients with HIV. Specifically, our analysis compares survival among the 1065 individuals who were antiretroviral naïve (Group 0) versus the 1399 individuals with prior antiretroviral therapy (Group 1).

Our long term event of interest, T_L , is the time from treatment randomization to death and intermediate event information consists of two intermediate events, $\mathbf{T}_S = (T_{S1}, T_{S2})^T$ where T_{S1} = time from randomization to an AIDS-defining event e.g. pneumocystis pneumonia and T_{S2} = time from randomization to a 50 % decline in CD4. If a patient experienced multiple intermediate events of one kind, for example multiple AIDS-defining events, the earliest occurrence of the event was used. For illustration, $t_0 = 1$ year and $t = 2.5$ years. Among individuals with prior antiretroviral therapy experience, 15.7 % were censored before 2.5 years while 31.8 % of antiretroviral therapy naïve individuals were censored before 2.5 years. Seventy individuals (5.0 %) with prior antiretroviral therapy experience and 25 (2.3 %) antiretroviral therapy naïve individuals experienced a decrease in CD4 count of at least 50 % within the first year of the study and survived past the first year, respectively. Twenty-seven (1.9 %) individuals with prior antiretroviral therapy experience and 9 (0.8 %) antiretroviral therapy naïve individuals experienced an AIDS defining event within the first year of the study and survived past the first year, respectively.

We estimate the average treatment effect using IPTW estimated using a logistic regression model that included all available baseline covariates. Namely, we aim to balance patients in our two exposure groups on the available observed covariates: the mean of two baseline CD4 counts, Karnofsky score, age at randomization, weight, symptomatic status, and treatment group to which they were originally randomized. Assessing balance is an important step given our requirement that our weights are consistently estimated. While it can be difficult to ensure that this assumption holds in practice, achieving balance in the two groups after weighting provides a good indication that bias in the treatment effect estimate due to observed covariates will be minimized (Stuart et al. 2013; Harder et al. 2010; Marcus et al. 2008). Table 3 shows balance for the groups before and after IPT weighting where we evaluate balance between the two prior therapy groups on the observed baseline covariates by examining a balance metric that summarizes the differences between the two univariate distributions of each baseline covariate, the absolute standardized mean difference (ASMD). For each covariate, the ASMD is the absolute value of the Group 1 mean minus the Group 0 mean divided by the pooled sample (Group 0 and 1) standard deviation. Sufficient balance is achieved when $ASMD < 0.10$ for all baseline covariates (Austin 2007; Austin and Stuart 2015; Austin 2009; Normand et al. 2001; Hankey and Myers 1971). The unweighted portion of the table shows that there were three notable differences between two prior therapy groups. Specifically, individuals in the antiretroviral therapy experienced group are more likely to have a lower average CD4 count at baseline ($ASMD = 0.316$), a lower mean weight ($ASMD = 0.154$) and a higher mean age at randomization ($ASMD = 0.181$). These characteristics associated with antiretroviral therapy experience are also known to be highly associated with survival among individuals with HIV. For example, patients who are older and skinnier

Table 3 Balance tables when using (a) no weights and (b) IPTW obtained using logistic regression across all covariates for the group with more than 52 weeks of antiretroviral therapy experience (Group 1) and the group with no antiretroviral therapy experience (naïve) (Group 0) where SD denotes standard deviation and ASMD denotes the absolute standardized mean difference

Covariate	Group 1 Mean	Group 1 SD	Group 0 Mean	Group 0 SD	ASMD	p-value
(a) Unweighted						
Mean of two baseline CD4 counts	337.445	102.884	371.110	108.124	0.316	0.000
Karnofsky score	0.565	0.496	0.617	0.486	0.105	0.010
Age at randomization	35.585	8.767	34.014	8.435	0.181	0.000
Weight (kg)	74.018	12.591	76.085	14.293	0.154	0.000
Symptomatic status	0.191	0.393	0.160	0.366	0.082	0.042
Assigned to Zidovudine	0.249	0.433	0.251	0.434	0.003	0.944
Assigned to Zidovudine and Didanosine	0.250	0.433	0.247	0.431	0.007	0.854
Assigned to Zidovudine and Zalcitabine	0.249	0.432	0.251	0.434	0.005	0.912
(b) IPT weighted						
Mean of two baseline CD4 counts	352.352	107.574	352.696	105.570	0.003	0.939
Karnofsky score	0.585	0.493	0.585	0.493	0.001	0.981
Age at randomization	34.927	8.604	34.905	8.846	0.003	0.951
Weight (kg)	74.913	12.951	74.885	13.808	0.002	0.959
Symptomatic status	0.176	0.381	0.177	0.382	0.001	0.986
Assigned to Zidovudine	0.250	0.433	0.251	0.434	0.002	0.969
Assigned to Zidovudine and Didanosine	0.248	0.432	0.246	0.431	0.004	0.930
Assigned to Zidovudine and Zalcitabine	0.251	0.434	0.252	0.434	0.001	0.979

are likely to be less healthy than other patients and at a higher risk of death. After IPT weighting, the two prior therapy groups appear balanced on all covariates. After weighting, there are no meaningful differences between the two prior therapy groups (all ASMD's below 0.005). Given this, the IPTW obtained using logistic regression are used to estimate survival and the effect of prior antiretroviral therapy in both the IPTW KM estimates and the IPTW landmark estimation procedure.

Figure 1 displays the unweighted KM estimate of survival in each group. As shown, without adjustment, individuals in the prior antiretroviral group appear to have worse

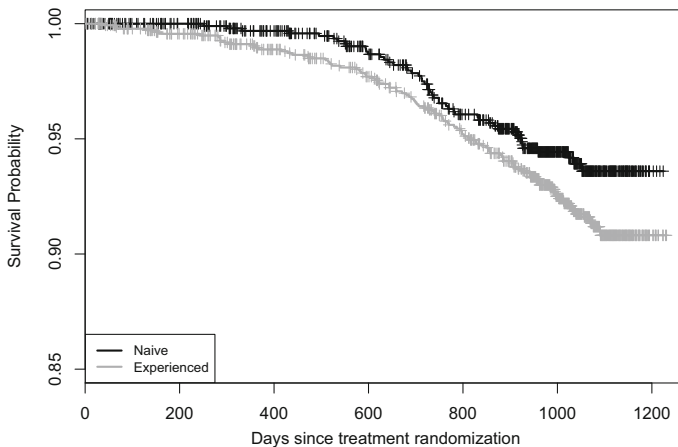


Fig. 1 Kaplan–Meier estimate of survival for antiretroviral naïve group (*black line*) and antiretroviral experienced group (*grey line*)

survival than those in the naïve group. Table 4 shows the estimated 2.5-year survival for each prior therapy group and the estimated treatment effects comparing the two groups using four different methods: unweighted KM, IPTW KM, RCT landmark estimation, and IPTW landmark estimation. As shown, the unweighted KM estimates show 2.5-year survival in the naïve therapy and prior therapy groups to be 0.953 and 0.937, respectively ($p = 0.1034$). The RCT landmark estimation supports similar survival estimates for the two groups. After weighting by the IPTW in either the KM or landmark estimation approach, we find the survival estimates for the two groups to be more similar, 0.948 versus 0.945, with IPTW KM and 0.950 and 0.947 with IPTW landmark estimation. This shift in survival estimates shows a clear connection to the imbalances in the baseline covariates between the two groups. Without IPTW adjustment, the group of individuals with prior ART appeared to have worse survival because they also tended to have higher CD4 counts, lower weight, and higher mean age. After proper adjustment for the imbalances, no significant differences are found between the two groups (p -values for both IPTW methods >0.50). Nonetheless, the IPT weighted landmark estimation procedure is roughly 16% more efficient than that from the IPTW KM estimate, showing one example of the types of increases in precision that might be gained by using landmark estimation in observational studies.

To shed light on whether this observed efficiency gain is due to the incorporation of *both* intermediate event information and baseline covariate information, we compared our estimates to those obtained using the two-stage procedure with only baseline covariates. The estimate of survival in the antiretroviral naïve group was 0.949 (SE = 0.008), the estimate of survival in the antiretroviral experienced group was 0.946 (SE = 0.006), and the estimate of the treatment effect in terms of the difference in survival was -0.0027 (SE = 0.01). The gains in efficiency using only baseline information compared to the IPTW KM estimate were thus about 3% for the survival estimates and 6% for the treatment effect estimate. Comparing these efficiency gains to those obtained using intermediate event information and baseline covariate information (5–10% for

Table 4 Resulting estimates of (a) $S(t)$ and (b) $\Delta(t)$ for $t = 2.5$ years in two exposure groups from ACTG Protocol 175 using the Kaplan–Meier estimator, $\widehat{S}_{KM,j}(t)$, $\widehat{\Delta}_{KM}(t)$; the IPTW KM estimator, $\widehat{S}_{IPTW,j}(t)$, $\widehat{\Delta}_{IPTW}(t)$; the landmark estimator from an RCT setting, $\widehat{S}_{LM,j}^{RCT}(t)$, $\widehat{\Delta}_{LM}^{RCT}(t)$; and the proposed landmark estimator, $\widehat{S}_{LM,j}(t)$, $\widehat{\Delta}_{LM}(t)$, with corresponding standard error from the perturbation-resampling method (SE), and relative efficiency (RE) for the unbiased estimates only with respect to the IPTW KM estimator, and corresponding p-values in (b), where $j = 0$ indicates antiretroviral naïve and $j = 1$ indicates antiretroviral experienced

(a) Survival estimates				
	$\widehat{S}_{KM,0}(t)$	$\widehat{S}_{IPTW,0}(t)$	$\widehat{S}_{LM,0}^{RCT}(t)$	$\widehat{S}_{LM,0}(t)$
Estimate	0.9530	0.9477	0.9556	0.9504
SE	0.0071	0.0081	0.0066	0.0075
RE	–	1	–	1.0861
	$\widehat{S}_{KM,1}(t)$	$\widehat{S}_{IPTW,1}(t)$	$\widehat{S}_{LM,1}^{RCT}(t)$	$\widehat{S}_{LM,1}(t)$
Estimate	0.9370	0.9449	0.9391	0.9467
SE	0.0067	0.0060	0.0064	0.0057
RE	–	1	–	1.0478
(b) Treatment effect estimates				
	$\widehat{\Delta}_{KM}(t)$	$\widehat{\Delta}_{IPTW}(t)$	$\widehat{\Delta}_{LM}^{RCT}(t)$	$\widehat{\Delta}_{LM}(t)$
Estimate	–0.0161	–0.0028	–0.0165	–0.0038
SE	0.0099	0.0103	0.0093	0.0095
RE	–	1	–	1.1565
p-value	0.1034	0.7841	0.0742	0.6933

survival, 16 % for treatment effect) demonstrates that in this particular application, the use of intermediate event information leads to improved efficiency over just using baseline measures.

Among the 1399 individuals with prior antiretroviral therapy in the trial, 476 individuals had 1–52 weeks of prior antiretroviral therapy and 923 individuals had over 52 weeks of prior antiretroviral therapy. Because the effect of antiretroviral therapy on survival may be quite different for those with extended prior therapy, we provide an additional analysis in the Supplementary Material comparing individuals who were antiretroviral naïve to those who had over 52 weeks of prior antiretroviral therapy. We discuss this further in the Discussion.

7 Discussion

In this paper we have developed the landmark estimation procedure of Parast et al. (2014) for use in an observational setting. It is particularly important to account for the possibility of selection bias in an observational setting when treatment is not randomized since failure to do so may lead to biased estimates. Our simulation study shows that the use of the landmark estimation procedure from an RCT setting in

the presence of selection bias does lead to biased estimates of survival and treatment effects. Furthermore, our proposed extension leads to unbiased estimates and improved efficiency compared to the unbiased IPTW KM estimator.

In addition to providing improved efficiency, our approach is robust to model misspecification of (4) and (7). If one were to assume that the outcome models (4) and (7) were correctly specified, one could simply use these models to obtain the desired survival probabilities and average over the observed covariate patterns. If these models are indeed correct, this approach would likely be more efficient than our proposed approach. However, if these models are not correct, this approach may lead to biased estimates. While such robustness is desirable, it is important to note that our proposed approach does still rely on the consistency of our IPTW. The literature is now rich with methods available to estimate IPTW (McCaffrey et al. 2004; van der Laan 2014; Breiman et al. 1984; Hill 2011; Imai and Ratkovic 2014; Liaw and Wiener 2002). In all applications of IPTW, concerns about the treatment assignment model being misspecified arise. In practice, parametric methods, such as logistic regression, tend to be used to model the treatment assignment indicator and estimate associated probabilities. However, generalized boosted models (GBM) and other machine learning techniques like the super learner have been proposed as an alternative for IPTW estimation as a way to minimize bias from incorrect assumptions about the form of the model used (McCaffrey et al. 2004; van der Laan 2014; Imbens 2000; Robins et al. 2000). These methods eliminate reliance on a simple parametric logistic regression model and do not require the researcher to determine which covariates and interactions should be included in the model. It has been shown that the resulting weights from these approaches yield more precise treatment effect estimates and lower mean squared error than traditional logistic regression methods and other alternative machine learning techniques (Harder et al. 2010; Lee et al. 2010). As a sensitivity analysis, we examined weights and resulting balance from GBM applied to the AIDS Clinical Trial dataset from Sect. 6 and our results were similar. In our approach, there is a trade-off between using parametric models like the logistic regression model and machine learner methods like GBM when constructing the IPTW. When utilizing logistic regression, the perturbation-resampling procedure described in Sect. 4 is straightforward and easily applied as it only involves fitting a weighted logistic model for each iteration while with GBM and other machine learners, perturbation-resampling becomes computationally intensive and in some cases infeasible if the machine learner method cannot incorporate the weights $V_i^{(b)}$. Future work is still needed to develop best practices for perturbation-resampling with machine learning procedures. In light of these trade-offs, we suggest that IPTW estimated by logistic regression models be utilized as long as balance between treatment and comparison groups has been obtained (e.g., a sign that bias from observed covariates in the treatment effect estimate should be limited) to allow for efficient use of the perturbation resampling approach. In contrast, if poor balance is obtained using parametric models we suggest the use of more state of the art methods like GBM to estimate the IPTW at the expense of foregoing perturbation-resampling of the IPTW.

We illustrated our proposed method using an AIDS clinical trial dataset and examined the effect of prior antiretroviral therapy on survival. We performed two analyses, one using all individuals and dichotomizing into prior therapy naïve compared to prior

therapy experienced, and a second analysis removing individuals with 1-52 weeks of prior antiretroviral therapy (presented in the Supplementary Material). However, it would be of interest to instead use the actual time of prior antiretroviral therapy experience rather than dichotomizing into two groups. The use of methodology that allows for treatment effect estimation with a continuous treatment, rather than dichotomous treatment groups, would be applicable in this example (Imai and Van Dyk 2004; Hirano and Imbens 2004; Zhu et al. 2015). Furthermore, future development of a landmark estimation procedure that can accommodate continuous treatment would be warranted.

A limitation of our proposed method is the required strong assumption that there are no unmeasured confounders in the model for the IPTW (Assumption A.2). In practice, one could consider sensitivity analyses to examine how sensitive the observed findings might be to violations of this assumption (Griffin et al. 2013; Rosenbaum and Rubin 1983a; Higashi et al. 2005).

A second limitation of our proposed method is the assumption that t_0 is pre-selected and fixed. There are several issues to consider when selecting t_0 . First, as was shown in Parast et al. (2014), the gain in efficiency that is observed when using a landmark estimation approach that incorporates intermediate event information is due to both the correlation between T_L and $\{T_S, Z\}$ and censoring. If there was weak correlation or very little censoring between t_0 and t , we would not expect to gain much efficiency using this approach. Second, if t_0 is chosen to be too close to baseline (or time of treatment initiation), we would not expect to observe many intermediate events between baseline and t_0 and thus incorporating intermediate event information is unlikely to lead to large gains in efficiency. On the other hand, if t_0 is chosen to be too close to t , then the subgroup with $X_L > t_0$ may be very small and thus we may also expect only small gains in efficiency and/or potentially small bias due to smoothing over a small sample. In the Supplementary Material we present simulation results across a range of t_0 (fixing $t = 2$) and results from the example across a range of t_0 (fixing $t = 2.5$ years). While the results for the example show that our findings are quite robust to the choice of t_0 , the results for the simulation study do demonstrate variability in relative efficiency with respect to the choice of t_0 . For example, when $t_0 = 0.5$ in the moderate treatment effect setting, the relative efficiency with respect to the IPTW KM estimator is almost 27% but when $t_0 = 1.5$ in this setting, the relative efficiency is less than 6%. Future work on the selection of t_0 either by examining efficiency across a range to identify the optimal t_0 (accounting for the selection procedure when making inference) or by considering a combination of multiple landmark times would be very useful in practice.

An R package implementing the methods described here, called `landest`, is available on CRAN.

References

- Amato DA (1988) A generalized Kaplan-Meier estimator for heterogeneous populations. *Commun Stat* 17(1):263–286
- Austin PC (2007) The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 26(16):3078–3094

- Austin PC (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 28:3083–3107
- Austin PC, Stuart EA (2015) Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 34(28):3661–3679
- Bai X, Tsiatis AA, O'Brien SM (2013) Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics* 69(4):830–839
- Beran R (1981) Nonparametric regression with randomly censored survival data. Technical report, University of California Berkeley
- Bhatta L, Klouman E, Deuba K, Shrestha R, Karki DK, Ekstrom AM, Ahmed LA (2013) Survival on antiretroviral treatment among adult HIV-infected patients in Nepal: a retrospective cohort study in far-western region, 2006–2011. *BMC Infect Dis* 13(1):604
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton
- Cai T, Tian L, Wei LJ (2005) Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika* 92(3):619–632
- Cai T, Tian L, Uno H, Solomon S, Wei L (2010) Calibrating parametric subject-specific risk estimation. *Biometrika* 97(2):389–404
- Chen PY, Tsiatis AA (2001) Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* 57(4):1030–1038
- Cook R, Lawless J (2001) Some comments on efficiency gains from auxiliary information for right-censored data. *J Stat Plan Inference* 96(1):191–202
- Cox DR (1972) Regression models and life tables (with discussion). *J R Stat Soc* 34:187–220
- Du Y, Akritas M (2002) Uniform strong representation of the conditional Kaplan-Meier process. *Math Methods Stat* 11(2):152–182
- Faucett CL, Schenker N, Taylor JM (2002) Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* 58(1):37–47
- Fine J, Jiang H, Chappell R (2001) On semi-competing risks data. *Biometrika* 88(4):907–919
- Finkelstein DM, Schoenfeld DA (1994) Analysing survival in the presence of an auxiliary variable. *Stat Med* 13(17):1747–1754
- Fleming TR, Prentice RL, Pepe MS, Glidden D (1994) Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* 13(9):955–968
- Garcia TP, Ma Y, Yin G (2011) Efficiency improvement in a class of survival models through model-free covariate incorporation. *Lifetime Data Anal* 17(4):552–565
- Gray R (1994) A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* 81(3):527–539
- Griffin BA, Eibner C, Bird CE, Jewell A, Margolis K, Shih R, Slaughter ME, Whitsel EA, Allison M, Escarce JJ (2013) The relationship between urban sprawl and coronary heart disease in women. *Health Place* 20:51–61
- Hammer S, Katzenstein D, Hughes M, Gundacker H, Schooley R, Haubrich R, Henry W, Lederman M, Phair J, Niu M et al (1996) A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New Engl J Med* 335(15):1081–1090
- Hammer SM, Squires KE, Hughes MD, Grimes JM, Demeter LM, Currier JS, Eron JJ Jr, Feinberg JE, Balfour HH Jr, Deyton LR et al (1997) A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New Engl J Med* 337(11):725–733
- Hankey BF, Myers MH (1971) Evaluating differences in survival between two groups of patients. *J Chron Dis* 24(9):523–531
- Harder VS, Stuart EA, Anthony JC (2010) Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods* 15(3):234
- Hernán MÁ, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 11(5):561–570
- Higashi T, Shekelle PG, Adams JL, Kamberg CJ, Roth CP, Solomon DH, Reuben DB, Chiang L, MacLean CH, Chang JT et al (2005) Quality of care is associated with survival in vulnerable older patients. *Ann Intern Med* 143(4):274–281
- Hill JL (2011) Bayesian nonparametric modeling for causal inference. *J Comput Gr Stat* 20(1):217–240

- Hirano K, Imbens GW (2004) The propensity score with continuous treatments. *Applied bayesian modeling and causal inference from incomplete-data perspectives: an essential journey with donald rubin's statistical family*. Wiley, New York, pp 73–84
- Imai K, Ratkovic M (2014) Covariate balancing propensity score. *J R Stat Soc* 76(1):243–263
- Imai K, Van Dyk DA (2004) Causal inference with general treatment regimes. *J Am Stat Assoc* 99(467):854–866
- Imbens GW (2000) The role of the propensity score in estimating dose-response functions. *Biometrika* 87(3):706–710
- Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53(282):457–481
- Lagakos S (1988) The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika* 75(1):156–160
- Lagakos S, Schoenfeld D (1984) Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* 40:1037–1048
- Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. *Stat Med* 29(3):337–346
- Li Y, Taylor JM, Little RJ (2011) A shrinkage approach for estimating a treatment effect using intermediate biomarker data in clinical trials. *Biometrics* 67(4):1434–1441
- Liaw A, Wiener M (2002) Classification and regression by randomforest. *R News* 2(3):18–22
- Lin D (2000) On fitting cox's proportional hazards models to survey data. *Biometrika* 87(1):37–47
- Lin D, Wei L (1989) The robust inference for the cox proportional hazards model. *J Am Stat Assoc* 84:1074–1078
- Lu X, Tsiatis A (2008) Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* 95(3):679–694
- Marcus SM, Siddique J, Ten Have TR, Gibbons RD, Stuart E, Normand SLT (2008) Balancing treatment comparisons in longitudinal studies. *Psychiatr Ann* 38(12):805
- McCaffrey DF, Ridgeway G, Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 9(4):403
- McCroft A, Madge S, Johnson AM, Lazzarin A, Clumeck N, Goebel FD, Viard JP, Gatell J, Blaxhult A, Lundgren JD et al (1999) A comparison of exposure groups in the eurosida study: starting highly active antiretroviral therapy (HAART), response to HAART, and survival. *JAIDS* 22(4):369–378
- Murray S, Tsiatis A (1996) Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 52:137–151
- Murray S, Tsiatis AA (2001) Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. *Lifetime Data Anal* 7(2):125–141
- Nieto FJ, Coresh J (1996) Adjusting survival curves for confounders: a review and a new method. *Am J Epidemiol* 143(10):1059–1068
- Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ (2001) Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol* 54(4):387–398
- Pan Q, Schaubel DE (2008) Proportional hazards models based on biased samples and estimated selection probabilities. *Can J Stat* 36(1):111–127
- Parast L, Tian L, Cai T (2014) Landmark estimation of survival and treatment effect in a randomized clinical trial. *J Am Stat Assoc* 109(505):384–394
- Park Y, Wei LJ (2003) Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* 90:717–723
- Patel K, Williams PL, Seeger JD, McIntosh K, Van Dyke RB, Seage GR et al (2008) Long-term effectiveness of highly active antiretroviral therapy on the survival of children and adolescents with HIV infection: a 10-year follow-up study. *Clin Infect Dis* 46(4):507–515
- Robins JM, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–560
- Rosenbaum PR, Rubin DB (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc* 45:212–218
- Rosenbaum PR, Rubin DB (1983b) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
- Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 79(387):516–524

- Rotnitzky A, Robins J (2005) Inverse probability weighted estimation in survival analysis. *Encycl Biostat* 4:2619–2625
- Stuart EA, Lee BK, Leacy FP (2013) Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol* 66(8):S84–S90
- Therneau TM (2000) *Modeling survival data: extending the Cox model*. Springer, New York
- Thomsen BL, Keiding N, Altman DG (1991) A note on the calculation of expected survival, illustrated by the survival of liver transplant patients. *Stat Med* 10(5):733–738
- Tian L, Cai T, Goetghebeur E, Wei L (2007) Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* 94(2):297–311
- Tian L, Cai T, Zhao L, Wei LJ (2012) On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. *Biostatistics* 13(2):256–273
- van der Laan MJ (2014) Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int J Biostat* 10(1):29–57
- Van Houwelingen J, Putter H (2012) *Dynamic prediction in clinical survival analysis*. CRC Press, New York
- Wood E, Hogg RS, Yip B, Harrigan PR, O’Shaughnessy MV, Montaner JS (2003) Is there a baseline CD4 cell count that precludes a survival response to modern antiretroviral therapy? *AIDS* 17(5):711–720
- Xie J, Liu C (2005) Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med* 24(20):3089–3110
- Zhang M (2015) Robust methods to improve efficiency and reduce bias in estimating survival curves in randomized clinical trials. *Lifetime Data Anal* 2014:1–19
- Zhang M, Schaubel DE (2012a) Contrasting treatment-specific survival using double-robust estimators. *Stat Med* 31(30):4255–4268
- Zhang M, Schaubel DE (2012b) Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics* 68(4):999–1009
- Zhang M, Tsiatis AA, Davidian M (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 64(3):707–715
- Zhao L, Cai T, Tian L, Uno H, Solomon S, Wei L (2010) Stratifying subjects for treatment selection with censored event time data from a comparative study. *Harvard University Biostatistics Working Paper Series*, p 122
- Zhu Y, Coffman DL, Ghosh D (2015) A boosting algorithm for estimating generalized propensity scores with continuous treatments. *J Causal Inference* 3(1):25–40