*Biostatistics* (2021), **0**, 0, *pp.* 1–27
doi:10.1093/biostatistics/"Biostatistics˙final version"

# Causal Inference for Recurrent Event Data using Pseudo-Observations

Chien-Lin Su*, Robert W. Platt, Jean-François Plante
*Department of Epidemiology, Biostatistics and Occupational Health, McGill University*
*Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montréal*
*Department of Decision Sciences, HEC Montréal, Montréal, Quebec, Canada*

chien-lin.su@mail.mcgill.ca

## Summary

Recurrent event data are commonly encountered in observational studies where each subject may experience a particular event repeatedly over time. In this article, we aim to compare cumulative rate functions of two groups when treatment assignment may depend on the unbalanced distribution of confounders. Several estimators based on pseudo-observations are proposed to adjust for the confounding effects, namely inverse probability of treatment weighting estimator, regression model-based estimators and doubly robust estimators . The proposed marginal regression estimator and doubly robust estimators based on pseudo-observations are shown to be consistent and asymptotically normal. A bootstrap approach is proposed for the variance estimation of the proposed estimators. Model diagnostic plots of residuals are presented to assess the goodness-of-fit for the proposed regression models. A family of adjusted two-sample pseudo-score tests is proposed to compare two cumulative rate functions. Simulation studies are conducted to assess finite sample performance of the proposed method. The proposed technique is demonstrated through an application to a hospital readmission data set.

*Key words*: Recurrent event data; Cumulative rate function; Pseudo-observations; Inverse probability of treatment weighting; Doubly robust estimator; Two-sample pseudo-score tests.

## 1. Introduction

In clinical studies, patients may experience a same type of event of interest repeatedly over time, which are referred to as recurrent events. Examples of recurrent events include recurrent asthma attacks in children (**?**) and repeated hospital readmission for colorectal cancer patients (**?**). The investigators are often interested in the effect of covariates on the recurrent events and in the comparison of cumulative rate functions (CRF) or mean functions among groups receiving different treatments. In randomized trial, subjects are randomly assigned to different treatment groups with no systematic difference between covariate factors. In such studies, the nonparametric Nelson-Aalen (NA) estimator (**?**) is commonly used to estimate the CRFs for specific groups, and the two-sample pseudo-score tests proposed by **?** can be applied to test the null hypothesis that the CRFs for two treatment groups are identical. For non-randomized trials or observational studies, confounding often occurs due to a dependence of the treatment assignment on the subjects' baseline characteristics and prognosis. The NA estimator and the two-sample pseudo-score tests may be biased and unreliable due to confounding effects arising from the possibly different distributions of subjects' baseline characteristics in the different treatment groups.

An illustrating example in this work is the hospital readmission dataset of colon cancer patients obtained from Hospital de Bellvitge in Barcelona, Spain, which were originally analyzed by **?**, but also available in the R library `frailtypack` (**?**). For each of the 403 colorectal cancer patients who were treated or not treated with chemotherapy, the times between each admission and readmission to the hospital were observed. Potential confounders include patients' sex and the Duke's staging classification of the tumor. A direct comparison of CRFs between the treated and untreated groups may not be valid as the unbalanced distribution of confounders such as gender or Duke's staging classification. Suitable adjustment for confounding effects is thus needed in order to make valid causal conclusions.

A milestone in causal inference is the emergence of the potential outcome framework, first raised by **?**. The main idea is to consider all possible observable and counterfactual outcomes simultaneously. The focus is often concentrated on estimating the average causal effect (ACE) in the whole population, i.e. the difference between the mean outcome over the target population if all subjects were in the treatment group and the mean outcome if all subjects were in the control group. Several methods have been proposed to estimate the ACE in the presence of confounders. They fall into three categories: (a) estimators using inverse probability of treatment weighting (IPTW) based on propensity score (PS) based on a model for the treatment assignment; (b) outcome regression (OR) model where estimators are often built on a standard regression model for the conditional expectation of the outcome given treatment and confounders; and (c) doubly robust (DR) estimators that include both OR and PS models simultaneously, and are robust to model misspecification as they are valid when at least one of the PS or OR model is correct. Those approaches have been widely discussed in the literature (**???**).

**?** proposed the pseudo-observations approach to model the state probabilities in multistate models. Consequently, pseudo-observations can be treated as complete data and be used for regression analysis such as generalized linear model (GLM). Many applications based on pseudo-observations are discussed in the literature including the estimation of: the restricted mean survival times (**?**), the survival function at a fixed time point (**?**) and the cumulative incidence function in the framework of competing risks (**?**). Additionally, their large sample properties have also been thoroughly investigated by **?**, **?** and **?**. Recently, **?** introduced how pseudo-observations can be utilized to estimate the ACE in the context of competing risks.

To estimate an ACE of interest in the context of recurrent event data, **?** investigated causal inference for randomized trials and all-or-none compliance. They proposed a complier average causal effect (CACE) which is the difference between the average numbers of recurrences in the treatment and control groups within the compliers. **?** studied the effect of omitting covariates in both the marginal rate model and the partially conditional rate model. The partially conditional rate model induces confounding through conditioning on the event history which leads to a biased estimate of treatment effect. This induced confounding by the conditioning is well-known in causal inference (**?**). However, those approaches mentioned above are valid only if the assumed regression model with confounding factors as covariates is correctly specified, which is however never known in practice.

In this work, we aim to develop DR estimators espically using the pseudo-observations to compare the CRFs of two groups and estimating the ACE for recurrent events in the presence of confounders, which has not been studied in the literature. We aim to fill this gap. The proposed DR estimators based on pseudo-observations assess some advantages. First, DR estimators are robust to model misspecification as they are based on a combination of IPTW and the regression model-based adjustment approaches. Second, the parameters related to OR model in the DR estimators are easily to estimate as pseudo-observations can be straightforward used for regression analysis such as GLM without censoring issue. Third, the proposed DR estimators can be easily constructed and implemented via standard software. Specifically, a model for treatment assignment with confounders as covariates is fitted to treatment assignment and confounder data. Given a set of time points, the pseudo-observations for each subject are then generated from the nonparametric NA estimator for recurrent event data. The IPTW estimator for the ACE is the difference between the weighted means of pseudo-observations from the two groups. The ACE can also be estimated via the G-formula approach in which two OR models are considered with treatment and confounders as covariates. While the first model assumes a semiparametric multiplicative rate (SMR) model on the event rate of recurrent process, the second one treats the pseudo-observations of CRF calculated at a set of time points as responses in a GLM. Finally, the DR estimators are constructed by combining both the IPTW estimator and the G-formula estimator obtained from either the SMR model, GLM or Super Learner (SL) approach. To test the null hypothesis that the CRFs for the treated and untreated groups are identical, adjusted versions of the two sample pseudo-score tests are proposed.

The remainder of the paper is organized as follows. In Section 2, we formalize the ACE parameters of interest and use a naive NA estimator for estimation. In Section 3, we propose several estimators including IPTW estimator, G-formula estimators and DR estimators. The asymptotic properties of the proposed regression estimator and DR estimators based on pseudo-observations are established, and a procedure to estimate variance with the bootstrap is also provided. Graphical model diagnosis based on (pseudo-)residuals to assess the adequacy of the proposed OR models are also presented. A family of adjusted two-sample pseudo-score tests is proposed in Section 4. Section 5 reports some simulation results. The analysis of a real dataset is provided in Section 6, and some concluding remarks are given in Section 7. All proofs, extra simulation tables and additional figures are provided in the online supplementary materials. Also, the R codes are deposited to github (https://github.com/ChienLinSu/CIRED-PO).

## 2. Notations and formulation of the problem

Consider a clinical trial of total duration $\tau$ in which $n$ patients are assigned to receive one of two treatments. For subject $i$ ($i = 1, ..., n$), let $Z_i$ be a dichotomous treatment indicator and $\mathbf{X}_i$ a $p$-dimensional vector of confounders. We denote observed values of $Z_i$ by $z_i$ and let $\mathbf{z} = (z_1, ..., z_n)$ be the vector of treatment assignments for the whole sample where $z_i = 1$ if subject $i$ is in the treatment group and $z_i = 0$ otherwise. Under the stable unit treatment value assumption (SUTVA) of **?**, subjects' outcomes are independent of the treatment assigned to other patients. Under treatment assignment $z$, we define $\tilde{N}_i^z(t)$, the potential outcomes for the number of events observed by time $t \in [0, \tau]$, and $C_i^z$, the potential right censoring time. We assume that the outcome for subject $i$ depends only on their treatment and not that received by other patients. In addition, let $\tilde{Y}_i^z(t) = I(C_i^z \geqslant t)$ be the potential outcomes for the "at risk" function indicating whether subject $i$ is under observation at time $t$. With two treatments, $z \in \{0,1\}$, and all these elements indexed with $z$ are defined for both treatments. Throughout this paper, we assume (i) independence of vectors $(\tilde{N}_i^z(\cdot), \tilde{Y}_i^z(\cdot), C_i^z, \mathbf{X}_i, Z_i)$, $i = 1, ..., n$, which are also identically distributed; (ii) random assignment where $\tilde{N}_i^1(\cdot), \tilde{Y}_i^1(\cdot), C_i^1, \tilde{N}_i^0(\cdot), \tilde{Y}_i^0(\cdot), C_i^0$ are independent of $Z_i$ conditional on $\mathbf{X}_i$, and (iii) censoring at random, meaning that censoring mechanisms $C_i^z$ are independent of the recurrent event processes $\tilde{N}_i^z$ given confounders $\mathbf{X}_i$. The focus of this work is to estimate the average causal effect (ACE) at a specific time $t \in [0, \tau]$, which is defined as the difference in the average number of recurrent events observed by time $t$ for patients in the treated and untreated groups, i.e., we consider $\theta(t) = E[\tilde{N}^1(t)] - E[\tilde{N}^0(t)] \equiv \Lambda^1(t) - \Lambda^0(t)$, where $\Lambda^z(t) = E[\tilde{N}^z(t)]$. When the occurrence rate of events is conditional on the event history $\mathcal{F}_H^z(t) = \{\tilde{N}^z(u) : 0 \leqslant u \leqslant t\}$, $\Lambda^z(t)$ is called a mean function (MF), but otherwise, it is a cumulative rate function (CRF). **?** showed that the estimated parameters in the Cox model cannot be interpreted causally. The problem stems from conditioning on the event history, namely that the hazard function for an individual at time $t$ implies that he survived up to that point. By conditioning on a so-called collider, noncausal pathways may get activated and commonly used effect estimates may not be interpreted causally as short-term risks (**?**). To avoid a similar issue, we consider the occurrence rate of events at time $t$ unconditionally on the event history $\mathcal{F}_H^z(t)$ throughout this paper.

Instead of observing both potential counting processes $\tilde{N}_i^1(t)$ and $\tilde{N}_i^0(t)$ simultaneously, we only observe $\tilde{N}_i(t) = Z_i\tilde{N}_i^1(t) + (1 - Z_i)\tilde{N}_i^0(t)$ for subject $i$ and similarly for the observed at risk process $\tilde{Y}_i(t) = Z_i\tilde{Y}_i^1(t) + (1 - Z_i)\tilde{Y}_i^0(t)$. When the treatment assignment $Z_i$ is independent of the potential processes $(\tilde{N}_i^1(t), \tilde{N}_i^0(t))$ for $i = 1, ..., n$; i.e., $E[\tilde{N}_i(t)|Z_i = 1] = E[\tilde{N}_i^1(t)]$ and $E[\tilde{N}_i(t)|Z_i = 0] = E[\tilde{N}_i^0(t)]$, one might utilize the Nelson-Aalen (NA) estimator (**?**) to estimate the CRF for the treated and untreated groups respectively. That is, $\theta(t)$ can be estimated by $\hat{\theta}_{\mathrm{NA}}(t) = \hat{\Lambda}_{\mathrm{NA}}^1(t) - \hat{\Lambda}_{\mathrm{NA}}^0(t)$ where $\hat{\Lambda}_{\mathrm{NA}}^1(t) = \sum_{i=1}^n \int_0^t \dfrac{Z_i\tilde{Y}_i(s)}{\sum_{j=1}^n Z_j\tilde{Y}_j(s)}\, \mathrm{d}\tilde{N}_i(s)$ and

$\hat{\Lambda}^0_{\mathrm{NA}}(t) = \sum_{i=1}^n \int_0^t \dfrac{(1-Z_i)\tilde{Y}_i(s)}{\sum_{j=1}^n (1-Z_j)\tilde{Y}_j(s)} \mathrm{d}\tilde{N}_i(s)$. However, in observational studies, the existence of confounders prevents straightforward estimation of $\theta(t)$ based on $\hat{\theta}_{\mathrm{NA}}(t)$ because the independence between the assignment $Z_i$ and the potential processes $(\tilde{N}_i^1(t),\tilde{N}_i^0(t))$ does not hold anymore.

## 3. Proposed Methodology

In this section, we propose six different estimators for $\theta(t)$ that account for confounders. Table 1 summarizes the links between the proposed estimators and their corresponding models and censoring assumptions. Basically, estimators based on pseudo-observations require censoring to be independent of all other variables (censoring completely at random); while the conditional regression model for the event times (SMR model) requires only that censoring is independent of the counting process given the covariates (censoring at random).

### 3.1 *IPTW Estimator*

To construct the IPTW estimator for $\theta(t)$, we utilize the pseudo-observations approach to CRF. To be specific, given a time $t$, the CRF-based pseudo-observation for subject $i$ is calculated by $\hat{\Lambda}^i(t) = n\hat{\Lambda}_{\mathrm{NA}}(t) - (n-1)\hat{\Lambda}^{-i}_{\mathrm{NA}}(t)$, where $\hat{\Lambda}_{\mathrm{NA}}(t)$ is the NA estimator calculated with all subjects and $\hat{\Lambda}^{-i}_{\mathrm{NA}}(t)$ is the same estimator obtained when leaving out subject $i$. Illustrations of $\hat{\Lambda}^i(t)$ may be found in Web Appendix C. Simulated data are used to represent different shapes that can occur: functions that are all positive, all negative, or display positive and negative values. In Web Appendix A from the online supplementary materials, we show that

$$E[\hat{\Lambda}^i(t)|Z_i, \boldsymbol{X}_i] \approx E[\tilde{N}_i(t)|Z_i, \boldsymbol{X}_i] = \Lambda(t|Z_i, \boldsymbol{X}_i), \tag{3.1}$$

where $\Lambda(t|Z_i, \boldsymbol{X}_i)$ shows that individual treatments and covariates may influence the expected outcome. We then construct the IPTW estimator based on $\hat{\Lambda}^i(t)$ as defined above. Specifically, we adopt the propensity score (PS) of **?** to balance the confounders between the treated and untreated groups in the sense that a PS-corrected distribution of the confounders would be identical in the two groups. In practice, the PS can be modeled by a logistic regression where we denote the individual probabilities

$$e_i(\boldsymbol{\alpha}) = Pr(Z_i = 1|\mathbf{V}_i) = \frac{\exp(\boldsymbol{\alpha}^T\mathbf{V}_i)}{1+\exp(\boldsymbol{\alpha}^T\mathbf{V}_i)}, \tag{3.2}$$

with $\mathbf{V}_i = (1, \mathbf{X}_i)^T, i = 1, ..., n$ and where $\boldsymbol{\alpha}$ is the $(p{+}1)$-dimensional vector of regression parameters. Using the IPTW approach proposed by **?** as well as (3.1) and (3.2), $\Lambda^z(t)$, $z \in \{0,1\}$ can be estimated based on the pseudo-observations $\hat{\Lambda}^i(t)$. For a fixed time $t$, the IPTW estimator for $\Lambda^z(t)$ can be constructed as $\hat{\Lambda}^1_{\mathrm{IPTW}}(t) = n^{-1}\sum_{i=1}^n \dfrac{Z_i\hat{\Lambda}^i(t)}{e_i(\hat{\boldsymbol{\alpha}})}$ and $\hat{\Lambda}^0_{\mathrm{IPTW}}(t) = n^{-1}\sum_{i=1}^n \dfrac{(1-Z_i)\hat{\Lambda}^i(t)}{1-e_i(\hat{\boldsymbol{\alpha}})}$,where $\hat{\boldsymbol{\alpha}}$ is the estimate for $\boldsymbol{\alpha}$ obtained by fitting the PS model in (3.2). Thus, $\theta(t)$ can be estimated by$\hat{\theta}_{\mathrm{IPTW}}(t) = \hat{\Lambda}^1_{\mathrm{IPTW}}(t) - \hat{\Lambda}^0_{\mathrm{IPTW}}(t)$ which is an unbiased estimate of $\theta(t)$ as long as the logistic regression in (3.2) is correctly specified.

### 3.2 *G-formula Estimators*

Our second strategy is motivated by the G-formula (**?**), where outcome regression (OR) models for the relationship between the outcome of interest, confounders and treatment are used to eliminate the bias directly. We consider two versions thereof.

3.2.1 *Semiparametric multiplicative rate (SMR) estimator* We first consider the following semiparametric multiplicative rate (SMR) model

$$E\big[\mathrm{d}\tilde{N}_i(t)|\boldsymbol{X}_i, Z_i\big] = h\{\gamma Z_i + \boldsymbol{\beta}^\top \boldsymbol{X}_i\}\mathrm{d}\mu(t), \tag{3.3}$$

where $\boldsymbol{\beta}$ and $\gamma$ are regression parameters, and $\mu(t)$ is the unspecified baseline rate function. The link function, $h : \mathbb{R} \to \mathbb{R}$ with $h(\cdot) \geqslant 0$, is pre-specified and assumed to be continuous almost everywhere and twice differentiable. Possible link functions include $h(x) = \exp(x)$, $h(x) = 1 + x$ and $h(x) = \log(1 + \exp(x))$. Notice that the proposed model (3.3) is in line with the models in **?** and in subsection 3.3.3 of **?**. Under model (3.3), $\theta(t)$ can be expressed as $\theta(t) = E_{\boldsymbol{X}}\Big[E\big[\tilde{N}(t)|Z = 1, \boldsymbol{X}\big] - E\big[\tilde{N}(t)|Z = 0, \boldsymbol{X}\big]\Big]$, where $E_{\boldsymbol{X}}$ stands for taking expectation with respect to the distribution of $\boldsymbol{X}$ in the whole population. Hence, one can estimate $\theta(t)$ by $\hat{\theta}_{\mathrm{SMR}}(t) = \dfrac{1}{n}\sum_{i=1}^{n}\left[h\{\hat{\gamma}_{\mathrm{SMR}} + \hat{\boldsymbol{\beta}}_{\mathrm{SMR}}^{\top}\mathbf{X}_i\}\hat{\mu}(t) - h\{\hat{\boldsymbol{\beta}}_{\mathrm{SMR}}^{\top}\mathbf{X}_i\}\hat{\mu}(t)\right]$, where $\hat{\boldsymbol{\beta}}_{\mathrm{SMR}}^{*} = (\hat{\gamma}_{\mathrm{SMR}}, \hat{\boldsymbol{\beta}}_{\mathrm{SMR}}, \hat{\mu}(t))$ are estimators of $\boldsymbol{\beta}_{\mathrm{SMR}}^{*} = (\gamma^{*}, \boldsymbol{\beta}^{*}, \mu^{*}(t))$. Note that $\hat{\boldsymbol{\beta}}_{\mathrm{SMR}}^{*}$ can be obtained using results from **?**, in particular their estimating equations (5) and (6) with $K = 1$ since we consider only one type of recurrent events. Asymptotic properties of $\hat{\boldsymbol{\beta}}_{\mathrm{SMR}}^{*}$ proved in Theorem 1 of **?** hold here, and details about inference procedures can be found therein. The validity of the estimator for $\hat{\theta}_{\mathrm{SMR}}(t)$ depends on the correct specification of the SMR model (3.3), which can be assessed by examining the total summation of the residuals for each subject, $\hat{M}_i(t; \hat{\gamma}_{\mathrm{SMR}}, \hat{\boldsymbol{\beta}}_{\mathrm{SMR}}) = \tilde{N}_i(t) - \int_0^t \tilde{Y}_i(u)h\{\hat{\gamma}_{\mathrm{SMR}}Z_i + \hat{\beta}_{\mathrm{SMR}}^{\top}\boldsymbol{X}_i\}d\hat{\mu}(u)$, as proposed by **?**. For a correct model, these residuals should have a mean of approximately zero and be independent of the covariates.

3.2.2 *Pseudo-Observations Estimator* Instead of imposing a certain structure for all time points such as the proportional rates in model (3.3), an alternative strategy consists of modelling the covariate effects directly on the CRF at a finite set of time points using the pseudo-observations approach. As investigated by **?**, estimators based on the pseudo-observations approach are still unbiased for the ACE of interest while the proportional hazard assumption is violated for the Cox model. The same idea applies here and using pseudo-observations for CRF allows avoiding bias when proportional rates are misspecified. The pseudo-observations are evaluated at those time points and used as response in a generalized linear model (GLM) for the covariate effects. Note that the individuals' pseudo-observations in the GLM may not really be interpretable but are rather just devices for estimation. Specifically, denote $\boldsymbol{t} = \{t_1, ..., t_H\}$ as the set of distinct times and define the pseudo-observation for subject $i$ at time $t_h$ as $\hat{\Lambda}^i(t_h)$ where $h = 1, ..., H$ and $i = 1, ..., n$. We then assume a GLM with

$$g(\Lambda^i(t_h)) = \xi^{t_h} + \gamma Z_i + \boldsymbol{\beta}^{\top}\boldsymbol{X}_i, \tag{3.4}$$

where $\xi^{t_h}$ is the intercept term for time $t_h$, $\boldsymbol{\beta}$ and $\gamma$ are regression parameters and $g$ is a link function. Common choices include the cloglog function $g(x) = \log(-\log(x))$ and $g(x) = \log(x)$. In practice, the choice of link function depends on the parameter of interest. For example, one would choose the logarithm function when estimating the cumulative hazard or the cloglog function for estimating the probability of survival at a given time point. Note that when $g(x) = \log(x)$ and $\xi^{t_h} = \log \mu(t_h)$, model (3.4) is equivalent to the SMR model (3.3), but since the estimating strategies are different, a comparison between these two approaches is presented in Section 5. To estimate the unknown parameters $\gamma$, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_H = \{\xi^{t_1}, ..., \xi^{t_H}\}$, we use the generalized estimating equation (GEE) as proposed by **?**. We get the gradient

$$U(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\varrho}_i^{-1}(\boldsymbol{t}, \boldsymbol{\beta}^*; Z_i, \boldsymbol{X}_i)}{\partial \boldsymbol{\beta}^*} \boldsymbol{V}_i^{-1} \Big( \hat{\Lambda}^i(\boldsymbol{t}) - \boldsymbol{\varrho}_i^{-1}(\boldsymbol{t}, \boldsymbol{\beta}^*; Z_i, \boldsymbol{X}_i) \Big), \tag{3.5}$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_H, \gamma, \boldsymbol{\beta})$, $\hat{\Lambda}^i(\boldsymbol{t}) = (\hat{\Lambda}^i(t_1), ..., \hat{\Lambda}^i(t_H))^\top$, $\boldsymbol{\varrho}_i^{-1}(\boldsymbol{t}, \boldsymbol{\beta}^*; Z_i, \boldsymbol{X}_i)$ is a vector of $H$ elements whose $j^{th}$ component is equal to $g^{-1}(\xi^{t_j} + \gamma Z_i + \boldsymbol{\beta}^T \boldsymbol{X}_i)$, and $\boldsymbol{V}_i : H \times H$ is the usual working covariance matrix which may account for the correlation structure inherent to the pseudo-observations as mentioned by **?**. Note that although the pseudo-observations can be negative and the proposed GLM in (3.4) has a log link function, the equation (3.5) still works for estimating $\boldsymbol{\beta}^*$ since it does not use the logarithm of the pseudo-observations. In the simulation studies and the real data analysis presented in a later section, we adopt an independent correlation structure among pseudo-observations. Additional simulations not reported here showed that specifying a correlated matrix brings no advantage. This is in line with **?** and **?** who also suggest to use an independent correlation structure in the context of competing risks. Under model (3.4), $\theta(t), t \in \boldsymbol{t}$ can therefore be estimated by $\hat{\theta}_{\text{PO}}(t) = \frac{1}{n} \sum_{i=1}^{n} \Big[ g^{-1}(\hat{\xi}^t_{\text{PO}} + \hat{\gamma}_{\text{PO}} + \hat{\boldsymbol{\beta}}^T_{\text{PO}} \boldsymbol{X}_i) - g^{-1}(\hat{\xi}^t_{\text{PO}} + \hat{\boldsymbol{\beta}}^T_{\text{PO}} \boldsymbol{X}_i) \Big]$, where $\hat{\boldsymbol{\beta}}^*_{\text{PO}} = (\hat{\xi}^t_{\text{PO}}, \hat{\gamma}_{\text{PO}}, \hat{\boldsymbol{\beta}}_{\text{PO}})$ are estimators obtained from solving equations (3.5) which can be done by using the "geese" function in the R package `geepack` (**?**) using $Z_i$, $\boldsymbol{X}_i$ and a dummy categorical variable for $\boldsymbol{t} = \{t_1, ..., t_H\}$ as covariates.

For a given time $t$, denote by $\boldsymbol{\beta}^*_0 = (\xi^{*t}_{\text{PO}}, \gamma^*_{\text{PO}}, \boldsymbol{\beta}^*_{\text{PO}})$ the true parameters of interest. Theorem 1 in Web Appendix B shows the asymptotic normality of $\hat{\boldsymbol{\beta}}^*_{\mathrm{PO}}$; the proofs based on the ideas of **?** and **?** are found in Web Appendix B. In addition, to assess the fitness of the model (3.4), we adopt the idea of pseudo-residuals proposed by **?** and **?** to the context of recurrent event data. To be specific, we compare the pseudo-observations $\hat{\Lambda}^i(t)$ to the predicted values $\hat{\Lambda}(t|Z_i, \boldsymbol{X}_i)$, yielding the pseudo-residuals $\{\hat{\Lambda}^i(t) - \hat{\Lambda}(t|Z_i, \boldsymbol{X}_i); i = 1, ..., n\}$. If the model fits the data well, no trends should be perceptible in plots of the residuals against a covariate at any given time point. We illustrate both residuals on a real dataset in Section 6.

### 3.3 *Doubly Robust Estimators*

The IPTW estimator $\hat{\theta}_{\mathrm{IPTW}}(t)$ and G-formula estimators $\hat{\theta}_{\mathrm{SMR}}(t)$ and $\hat{\theta}_{\mathrm{PO}}(t)$ are unbiased for the ACE $\theta(t)$ only if the statistical models for the PS from (3.2) and OR model in (3.3) or (3.4) are correctly specified. Doubly robust (DR) estimators, however, are robust to misspecification in the sense that they combine both IPTW and OR estimators while remaining consistent as long as one of those two models is correctly specified (**?**). Following this idea, we propose two DR estimators which are constructed by combining the IPTW estimator $\hat{\theta}_{\mathrm{IPTW}}(t)$ with an OR estimator using either $\hat{\theta}_{\mathrm{SMR}}(t)$ and $\hat{\theta}_{\mathrm{PO}}(t)$. Specifically, the DR estimators for CRFs $\Lambda^1(t)$ and $\Lambda^0(t)$ can be constructed as follows:

$$\begin{aligned}
\hat{\Lambda}^1_{\mathrm{DR}}(t; \hat{\boldsymbol{\alpha}}, \hat{\Theta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{Z_i\hat{\Lambda}^i(t)}{e_i(\hat{\boldsymbol{\alpha}})} - \frac{(Z_i - e_i(\hat{\boldsymbol{\alpha}}))\hat{E}(\tilde{N}(t)|Z=1, \boldsymbol{X}_i)}{e_i(\hat{\boldsymbol{\alpha}})}\Big], \\
\hat{\Lambda}^0_{\mathrm{DR}}(t; \hat{\boldsymbol{\alpha}}, \hat{\Theta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{(1-Z_i)\hat{\Lambda}^i(t)}{1 - e_i(\hat{\boldsymbol{\alpha}})} + \frac{(Z_i - e_i(\hat{\boldsymbol{\alpha}}))\hat{E}(\tilde{N}(t)|Z=0, \boldsymbol{X}_i)}{1 - e_i(\hat{\boldsymbol{\alpha}})}\Big],
\end{aligned} \tag{3.6}$$

where $\hat{\Theta}$ is the estimator related to $\hat{E}(\tilde{N}(t)|Z = z, \boldsymbol{X})$, the estimator for $E(\tilde{N}(t)|Z = z, \boldsymbol{X})$, $z \in \{0, 1\}$, obtained from one of the OR models. Thus, the SMR-based DR estimator for $\theta(t)$ is

$$\hat{\theta}^{\mathrm{SMR}}_{\mathrm{DR}}(t) = \hat{\Lambda}^1_{\mathrm{DR}}(t; \hat{\boldsymbol{\alpha}}, \hat{\Theta}_{\mathrm{SMR}}) - \hat{\Lambda}^0_{\mathrm{DR}}(t; \hat{\boldsymbol{\alpha}}, \hat{\Theta}_{\mathrm{SMR}}) \tag{3.7}$$

in which $\hat{\Theta}_{\mathrm{SMR}} = (\hat{\gamma}_{\mathrm{SMR}}, \hat{\boldsymbol{\beta}}_{\mathrm{SMR}}, \hat{\mu}(t))$ and $\hat{E}(\tilde{N}(t)|Z = 0, \boldsymbol{X}_i)$ and $\hat{E}(\tilde{N}(t)|Z = 1, \boldsymbol{X}_i)$ are substituted by $h\{\hat{\boldsymbol{\beta}}^{\top}_{\mathrm{SMR}}\mathbf{X}_i\}\hat{\mu}(t)$ and $h\{\hat{\gamma}_{\mathrm{SMR}} + \hat{\boldsymbol{\beta}}^{\top}_{\mathrm{SMR}}\mathbf{X}_i\}\hat{\mu}(t)$, respectively, both obtained from the

SMR model in (3.3). Another alternative DR estimator relies on pseudo-observations, and

$$\hat{\theta}_{\mathrm{DR}}^{\mathrm{PO}}(t) = \hat{\Lambda}_{\mathrm{DR}}^{1}(t; \hat{\boldsymbol{\alpha}}, \hat{\Theta}_{\mathrm{PO}}) - \hat{\Lambda}_{\mathrm{DR}}^{0}(t; \hat{\boldsymbol{\alpha}}, \hat{\Theta}_{\mathrm{PO}}) \tag{3.8}$$

in which $\hat{\Theta}_{\mathrm{PO}} = (\hat{\xi}_{\mathrm{PO}}^{t}, \hat{\gamma}_{\mathrm{PO}}, \hat{\boldsymbol{\beta}}_{\mathrm{PO}})$ and $\hat{E}(\tilde{N}(t)|Z = 0, \boldsymbol{X}_i)$ and $\hat{E}(\tilde{N}(t)|Z = 1, \boldsymbol{X}_i)$ are replaced by $g^{-1}(\hat{\xi}_{\mathrm{PO}}^{t} + \hat{\boldsymbol{\beta}}_{\mathrm{PO}}^{\top}\boldsymbol{X}_i)$ and $g^{-1}(\hat{\xi}_{\mathrm{PO}}^{t} + \hat{\gamma}_{\mathrm{PO}} + \hat{\boldsymbol{\beta}}_{\mathrm{PO}}^{\top}\boldsymbol{X}_i)$, respectively, both coming from the GLM in (3.4) which used pseudo-observations. The consistency and normality for DR estimators (3.7) and (3.8) as well as their proofs are summarized in Web Appendix B.

We also investigate the DR estimator based on pseudo-observations when (3.2) and (3.4) are nonparametrically estimated at slower convergence rates which can be reached by considering the Super Learner (SL) estimators (**?**). That is, we consider $\hat{\theta}_{\mathrm{DR}}^{\mathrm{SL}}(t) = \hat{\Lambda}_{\mathrm{DR}}^{1}(t; \hat{\Theta}_{\mathrm{SL}}) - \hat{\Lambda}_{\mathrm{DR}}^{0}(t; \hat{\Theta}_{\mathrm{SL}})$ in which $\hat{\Theta}_{\mathrm{SL}}$ denote the SL estimators for (3.2) and (3.4). For illustration, we only consider algorithms, `SL.knn`, `SL.glm`, `SL.mean` and `SL.randomForest`, from the R package `Super Learner` (**?**) to investigate the performance of $\hat{\theta}_{\mathrm{DR}}^{\mathrm{SL}}$ in Section 5.

### 3.4 *Variance Estimation*

The variance formulae for regression model-based estimators $\hat{\theta}_{\mathrm{SMR}}(t)$ and $\hat{\theta}_{\mathrm{PO}}(t)$ may be calculated based on the delta method using the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\mathrm{SMR}}^{*}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PO}}^{*}$, respectively. Such calculations are however not straightforward as they involve complicated formulae. Due to the complexity of the variance formulae for $\hat{\theta}_{\mathrm{DR}}^{\mathrm{SMR}}(t)$ and $\hat{\theta}_{\mathrm{DR}}^{\mathrm{PO}}(t)$ as shown in Web Appendix B, it may not be straightforward to calculate the variance. In addition, **?** showed that a bootstrap approach estimator results in better performances in terms of smaller standard error and approximately correct coverage rate when using the IPTW approach for survival outcomes. Therefore, to avoid the issues mentioned above, we propose to estimate the variances of the estimated ACEs using nonparametric bootstrap which has also been adopted by **?** to obtain confidence limits for the ACE of interest in the context of competing risks. That is, we first resample $n$ subjects with replacement from the original data in order to obtain a bootstrap sample. Second, we recalculate the PS and the NA estimator based on a bootstrap sample. Third, we calculate the pseudo-observations based on the recalculated NA estimator and the proposed estimators are applied on a bootstrap sample. Finally, the above procedure is repeated $B$ times. The variances are computed empirically from the $B$ estimates.

## 4. Two-Sample Tests

We now consider developing tests for $H_0 : \Lambda^1(t) = \Lambda^0(t), 0 < t \leqslant \tau$, versus $H_1 : \Lambda^1(t) \neq \Lambda^0(t)$, for some $0 < t \leqslant \tau$, where $\Lambda^z(t), z = 0, 1$ are the CRFs for the untreated and treated groups respectively. In the absence of confounders, **?** investigated a family of pseudo-score test statistics for the null hypothesis $H_0$. The test statistics studied by **?** are based on $U^{\mathrm{CLN}}(t) = \int_0^t Q(u) d\hat{\theta}_{\mathrm{NA}}(u)$, where $Q(u) = \{\tilde{Y}_{0\cdot}(u)\tilde{Y}_{1\cdot}(u)a(u)\}/\tilde{Y}_{\cdot\cdot}(u)$, $a(u)$ is a fixed weight function such as 1 or $t - u$, $\tilde{Y}_{z\cdot}(u) = \sum_{i:Z_i=z} \tilde{Y}_i(u)$ is the size of the risk set at time $u$ for treated $(z = 1)$ and untreated $(z = 0)$ groups, and $\tilde{Y}_{\cdot\cdot}(u) = \tilde{Y}_{0\cdot}(u) + \tilde{Y}_{1\cdot}(u)$ is the total number of individuals at risk at time $u$ in the whole sample. **?** proposed standardized form of the test statistic using a variance estimate $\hat{V}_{\mathrm{P}}(t)$ based on a Poisson process assumption, and an alternative variance estimate $\hat{V}_{\mathrm{R}}(t)$ robust to a departure from that assumption. Under $H_0$, both $U_{\mathrm{P}}^{\mathrm{CLN}}(t) = [U^{\mathrm{CLN}}(t)]^2/\hat{V}_{\mathrm{P}}(t)$ and $U_{\mathrm{R}}^{\mathrm{CLN}}(t) = [U^{\mathrm{CLN}}(t)]^2/\hat{V}_{\mathrm{R}}(t)$ are asymptotically $\chi^2(1)$. These tests assume random assignment and can therefore not be directly performed when the groups are unbalanced due to confounding. Consequences of ignoring the failure of this assumption are shown in Section 5. To fix the imbalance, we exploit the weighted log-rank test for statistical comparison of survival functions proposed by **?** and define three adjusted versions of the two-sample pseudo-score tests in which components in $Q(u)$ are re-weighted and $\theta(t)$ is estimated by a DR estimate from $\hat{\theta}_{\mathrm{DR}}^{\mathrm{SMR}}(t)$, $\hat{\theta}_{\mathrm{DR}}^{\mathrm{PO}}(t)$ or $\hat{\theta}_{\mathrm{DR}}^{\mathrm{SL}}(t)$. Specifically, let $w_i^*(u) = \tilde{Y}_{z\cdot}(u) w_i(\hat{\boldsymbol{\alpha}}) / \sum_{i:Z_i=z} w_i(\hat{\boldsymbol{\alpha}})$ be the weight function at time $u$ for individual $i$ in the untreated $(z = 0)$ or treated $(z = 1)$ groups, where $w_i(\hat{\boldsymbol{\alpha}}) = Z_i/e_i(\hat{\boldsymbol{\alpha}}) + (1 - Z_i)/(1 - e_i(\hat{\boldsymbol{\alpha}}))$. Hence, the weights $w_i^*(u)$ are proportional to the number of individuals at risk for a given time $u$ in each group. We then propose the following three adjusted pseudo-score test statistics $U_{\mathrm{DR}}^{\mathrm{SMR}}(t) = \frac{\int_0^t Q^*(u) d\hat{\theta}_{\mathrm{DR}}^{\mathrm{SMR}}(u)}{\hat{\sigma}_{\mathrm{DR}}^{\mathrm{SMR}}(t)}$, $U_{\mathrm{DR}}^{\mathrm{PO}}(t) = \frac{\int_0^t Q^*(u) d\hat{\theta}_{\mathrm{DR}}^{\mathrm{PO}}(u)}{\hat{\sigma}_{\mathrm{DR}}^{\mathrm{PO}}(t)}$, and $U_{\mathrm{DR}}^{\mathrm{SL}}(t) = \frac{\int_0^t Q^*(u) d\hat{\theta}_{\mathrm{DR}}^{\mathrm{SL}}(u)}{\hat{\sigma}_{\mathrm{DR}}^{\mathrm{SL}}(t)}$ where $Q^*(u) = \tilde{Y}_{0\cdot}^*(u)\tilde{Y}_{1\cdot}^*(u)a(u)/\tilde{Y}_{\cdot\cdot}^*(u)$, $\tilde{Y}_{z\cdot}^*(u) = \sum_{i:Z_i=z, C_i \geqslant u} w_i^*(u)$ for $z = 0, 1$, $\tilde{Y}_{\cdot\cdot}^*(u) = \tilde{Y}_{0\cdot}^*(u) + \tilde{Y}_{1\cdot}^*(u)$ is the weighted number of individuals at risk in the combined sample at time $u$, and the denominators are the estimated standard errors of their respective numerators based on a nonparametric bootstrap approach. Under $H_0$, the test statistics $U_{\mathrm{DR}}^{\mathrm{SMR}}(t)$, $U_{\mathrm{DR}}^{\mathrm{PO}}(t)$ and $U_{\mathrm{DR}}^{\mathrm{SL}}(t)$ converge asymptotically to a standard normal for any fixed time point $t$; thus, the null hypothesis is rejected at level $\alpha$ if the absolute value of the chosen test statistic exceeds $z_{\alpha/2}$.

## 5. Simulations

Several simulation studies are conducted to evaluate the proposed estimators. For subject $i, i = 1, ..., n$, the data generating process has three independent covariates $(X_{i1}, X_{i2}, X_{i3})$, a bernoulli with mean 0.5, a uniform on $(0, 1)$ and a standard normal respectively. The propensity score (PS) model (3.2) has $\mathrm{logit}(e_i(\boldsymbol{\alpha})) = 0.2 + 0.4X_{i1} + 0.6X_{i2} + X_{i3}$. This yields approximately 64% of treated subjects and 36% of untreated individuals based on 500 simulated data. Event times for subjects are generated from the homogeneous Poisson processes (HPP): $E[d\tilde{N}_i(t)|X_{i1}, Z_i, \eta_i] = \eta_i \exp\{\gamma Z_i + \beta X_{i1}\} d\mu(t)$, where $\eta_i$ is a subject-specific Gamma frailty with $\mathrm{E}(\eta_i) = 1$ and $\mathrm{Var}(\eta_i) = \sigma^2$ inducing a positive correlation among the within-subject events. A large $\sigma^2$ implies a high positive correlation among event times, and $\sigma^2 = 0$ yields $\eta_i = 1$, which induces independence for event times within subjects. We set $\beta = -\log(5)$ and $\gamma = \log(0.8)$. A study duration of $\tau = 1$ is employed, and the censoring time for each subject is independently generated from a Uniform(0,$\tau$), which is also independent of the event processes. We consider $n \in \{100, 300\}$, $\sigma^2 \in \{0, 0.25\}$ and $\mu(t) \in \{5t, 20t\}$. With $\sigma^2 = 0$ and $\mu(t) = 5t$, the expected number of events per subject are 1.1 and 1.6 in the treated and untreated group respectively. They become 4.5 and 6.4 if $\mu(t) = 20t$. The variances of estimators are estimated based on $B = 200$ bootstrap samples.

Table 1 in Web Appendix C presents the performances of the marginal estimators $\hat{\gamma}$ and $\hat{\beta}$ obtained from SMR (3.3) and GLM (3.4). For GLM, we assume an independent correlation structure among pseudo-observations calculated from 4 or 10 time points which are either $j/5$ for $j = 1, \ldots, 4$ or $j/10$ for $j = 1, \ldots, 10$. For each estimator, we report the average bias (Bias), the empirical standard error (ESE), the average of the standard error estimator (SEE) and the empirical coverage rate (CR) of 95% confidence interval. Overall, the performance of marginal estimators is reasonable. Compared to the estimates obtained from the SMR model, the estimators based on pseudo-observations have slightly higher ESE (SEE) especially for estimators using only 4 time points. Indeed, the SMR model uses all data information between times 0 and 1 whereas the GLM based on pseudo-observations only 4 or 10 time points between times 0 and 1. However, ESEs (SEEs) from both models are close when the sample size $n$ or the baseline rate function $\mu(t)$ increases. From now on, to obtain better performance, the subsequent analyses related to pseudo-observations are conducted based on 10 time points.

Table 2 and 3 in Web Appendix C show the performance of estimators $\hat{\theta}_{\text{SMR}}(t)$ and $\hat{\theta}_{\text{PO}}(t)$ when the baseline rate functions are $\mu(t) = 5t$ and $20t$, respectively. We obtain that $\hat{\theta}_{\text{PO}}$ has higher ESE and SEE than $\hat{\theta}_{\text{SMR}}$, but that gap decreases as $n$ increases while the CR is getting closer to the 95% nominal level . As $\sigma^2$ increases, so does the ESE and SEE of estimators $\hat{\theta}_{\text{PO}}$ and $\hat{\theta}_{\text{SMR}}$. We also present the mean square error ratio (MSER), which is the ratio of MSE of $\hat{\theta}_{\text{PO}}$ to the MSE of $\hat{\theta}_{\text{SMR}}$. We observe values of MSER greater than 1, meaning that the MSE of $\hat{\theta}_{\text{SMR}}$ is smaller, but this relative advantage seems to decrease as $n$ increases.

Next, we examine the robustness of the IPTW estimator $\hat{\theta}_{\text{IPTW}}(t)$ and the DR estimators $\hat{\theta}_{\text{DR}}^{\text{SMR}}$, $\hat{\theta}_{\text{DR}}^{\text{PO}}$ and $\hat{\theta}_{\text{DR}}^{\text{SL}}$. For comparison purposes, the naive NA estimator is evaluated as well. We report the MSER of each estimator with the MSE of $\hat{\theta}_{\text{DR}}^{\text{SL}}$ as a reference. We estimate $\theta(t)$ at $t = 0.4$ and $0.8$ for three scenarios: (a) scenario (C,C) where PS model and OR model in HPP are both correct, (b) scenario (C,N) where the PS model is correctly specified but the OR model is not, and (c) scenario (N,C) if which the OR model is correctly specified but not the PS model. While the incorrectly specified PS model omits covariate $X_{i3}$, the incorrect OR model in HPP includes covariates $X_{i2}$ rather than $X_{i1}$. Table 2 presents the results with $\mu(t) = 5t$ and $\sigma^2 = 0.25$. The bias of $\hat{\theta}_{\mathrm{NA}}$ is obvious with a CR furthest from 95% nominal level in all scenarios especially for $n = 300$. For scenario (C,C), $\hat{\theta}_{\mathrm{IPTW}}$, $\hat{\theta}^{\mathrm{SMR}}_{\mathrm{DR}}$, $\hat{\theta}^{\mathrm{PO}}_{\mathrm{DR}}$ and $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ are unbiased and CRs are consistent with the 95% nominal level. Moreover, $\hat{\theta}^{\mathrm{SMR}}_{\mathrm{DR}}$, $\hat{\theta}^{\mathrm{PO}}_{\mathrm{DR}}$ and $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ have similar performance as $n$ increases. Their ESE and SEE are slightly lower than that of $\hat{\theta}_{\mathrm{IPTW}}$ especially for $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$. MSER values imply that $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ has the smallest MSE among all DR estimators. As expected, those estimators improve as $n$ increases from 100 to 300. For scenario (C,N), $\hat{\theta}_{\mathrm{IPTW}}$, $\hat{\theta}^{\mathrm{SMR}}_{\mathrm{DR}}$, $\hat{\theta}^{\mathrm{PO}}_{\mathrm{DR}}$ and $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ display little bias with reasonable CRs even when $n = 100$. ESE and SEE are slightly higher for $\hat{\theta}_{\mathrm{IPTW}}$ compared to DR estimators. In addition, $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ has the smallest MSE among all DR estimators. For scenario (N,C), $\hat{\theta}_{\mathrm{IPTW}}$ is biased as the PS model is incorrectly specified. Three DR estimators $\hat{\theta}^{\mathrm{SMR}}_{\mathrm{DR}}$, $\hat{\theta}^{\mathrm{PO}}_{\mathrm{DR}}$ and $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ are unbiased with a CR consistent with 95% nominal level, and $\hat{\theta}^{\mathrm{SL}}_{\mathrm{DR}}$ has the smallest MSE. From the simulation results of three scenarios, researchers would focus more on getting correct model for OR as the DR estimators tend to have smaller ESE and SEE once the OR model is correctly specified. In Section 7, we give a concrete set of recommendations in practice on which estimators to prefer for which situations.

Table 3 presents simulation results of the test statistics given in Section 4. We set $a(u) = 1$ and examine the empirical type I error rates and power of these test statistics. We re-express the parameter in HPP as $\gamma = \log \psi$ and set $\beta = -\log(5)$. The parameter $\psi$ represents the treatment effect on the CRF of the treatment group. Under $H_0 : \Lambda^1(t) = \Lambda^0(t), 0 < t \leqslant 1$, event times for subjects are generated with $\psi = 1$ and $\mu(t) = 5t$ and $20t$, respectively. Table 3 summarizes type I error rates under the scenarios (a), (b) and (c) described in the previous paragraph. We observe that both tests $U^{\mathrm{CLN}}_{\mathrm{P}}(t)$ and $U^{\mathrm{CLN}}_{\mathrm{R}}(t)$ proposed by **?** have inflated type I error rates increasing along with larger value of $\mu(t)$, particularly for $U^{\mathrm{CLN}}_{\mathrm{P}}(t)$. The inflated rates are caused by the fact that $\hat{\theta}_{\mathrm{NA}}$ is a biased estimator (see Table 2) due to confounding, and the value of $Q(u)$ without adjustment for the risk sets $\tilde{Y}_{z\cdot}(u), z = 0, 1$ is larger than the adjusted $Q^*(u)$. These two factors lead to $U^{\mathrm{CLN}}(t)$ having a large value and then easily rejecting $H_0$, even with robust variance estimation. Note that two tests $U_{\mathrm{P}}^{\mathrm{CLN}}(t)$ and $U_{\mathrm{R}}^{\mathrm{CLN}}(t)$ can be performed via the `mcfDiff.test` function in R package `reda` (?). As expected, the proposed tests $U_{\mathrm{DR}}^{\mathrm{SMR}}(t)$, $U_{\mathrm{DR}}^{\mathrm{PO}}(t)$ and $U_{\mathrm{DR}}^{\mathrm{SL}}(t)$ generally have similar and satisfactory performance for all scenarios. This is not surprising given the robustness of the DR estimators for $\theta(t)$ observed in Table 2. Overall, the error rates are consistent with the 0.05 nominal level as $n$ increases.

Table 4 of Web Appendix C shows the empirical power for test statistics under the alternative hypothesis where $\psi = 1.5$. When $\sigma^2 = 0$, the power of $U_{\mathrm{DR}}^{\mathrm{SMR}}(t)$, $U_{\mathrm{DR}}^{\mathrm{PO}}(t)$ and $U_{\mathrm{DR}}^{\mathrm{SL}}(t)$ are similar and comparable in all scenario. As expected, power increases as $n$ gets larger for each value of $\mu(t)$. The power also increases with $\mu(t)$. We observe that power is higher in scenario (N,C) compared to the other two scenarios. This result may be induced by the relatively smaller estimated standard errors $\hat{\sigma}_{\mathrm{DR}}^{\mathrm{SMR}}(t)$, $\hat{\sigma}_{\mathrm{DR}}^{\mathrm{PO}}(t)$ and $\hat{\sigma}_{\mathrm{DR}}^{\mathrm{SL}}(t)$ while their respective numerators are similar, resulting in the test statistics $U_{\mathrm{DR}}^{\mathrm{SMR}}(t)$, $U_{\mathrm{DR}}^{\mathrm{PO}}(t)$ and $U_{\mathrm{DR}}^{\mathrm{SL}}(t)$ can more easily reject $H_0$. As $\sigma^2$ increases from 0 to 0.25, the power of the three proposed tests decreases in each scenario. Note that the power of the analyses based on $U_{\mathrm{P}}^{\mathrm{CLN}}(t)$ and $U_{\mathrm{R}}^{\mathrm{CLN}}(t)$ are uninterpretable given the serious inflation of the type I error rate observed in Table 3.

## 6. Real Data Analysis

We apply the proposed methodology to a hospital readmission dataset for colorectal cancer patients. This dataset is available from the R package `frailtypack` (?). Each of 403 patients were followed up for a period of time, and the hospital readmission times for each patient were recorded. Time 0 corresponds to the first hospital admission due to colon cancer. Among 403 patients, 199 patients had no readmission, 150 patients had one or two readmissions and others patients had up to 22 readmissions. Patients were treated or not treated with chemotherapy, a decision that could be influenced by potential confounders including sex (male or female) and Duke's stage (combined in 3 groups: stages A-B, stage C, stage D).

The upper portion of Table 4 shows the estimated parameters for the propensity score (PS). We observe that males have a higher probability to be assigned to the treated group given the same Duke stage. Patients with Duke's stage A-B (the baseline group) or Duke's stage D also tend to be assigned to the treated group. The estimated weights $w_i(\hat{\boldsymbol{\alpha}}) = Z_i/e_i(\hat{\boldsymbol{\alpha}})+(1-Z_i)/(1-e_i(\hat{\boldsymbol{\alpha}}))$ for both treated and untreated groups are shown in Figure 2 of Web Appendix C. The boxplots indicate that the weights behave well for both groups with higher variation in the untreated group. The bottom panel of Table 4 presents the parameters of the OR based on SMR model (3.3) and GLM (3.4) with $g(x) = \log(x)$ and $\xi^t = \log \mu(t)$. Pseudo-observations $\hat{\Lambda}_i(t)$ are calculated for each patient at all 367 observed event times from the hospital readmission data. We observe that chemotherapy reduces the risk of hospital readmission under both OR models with significant p-values at the 10% level for both and at the 5% level under GLM (3.4). The coefficient estimate for females reveals that they have a lower readmission rate than males and is statistically significant in both OR models. The estimated coefficient for Duke's stage D is significant, which implies that patients at the highest stage of the disease have intensive hospital readmission. Overall, the fitted results from the SMR model are consistent with the findings based on the GLM, but they display slightly lower standard errors. The results described above are compatible with the findings of **?** in terms of significance and direction of the effects. This is not surprising as they also assumed that covariates and treatment have proportional effects on the occurrence rate of counting process although they considered a parametric model.

Figure 3 of Web Appendix C is used to assess the adequency of GLM (3.4) with $g(x) = \log(x)$ and $\xi^t = \log \mu(t)$. It shows boxplots of the pseudo-residuals for both treatment groups at given several time points. Since the pseudo-residuals fluctuate around zero at any given time point, the plots support the adequacy of the proposed GLM. The variation of pseudo-residuals increases as time increases, especially for the untreated group. An additional set of pseudo-residual plots with sex as a covariate is presented in Figure 4 of Web Appendix C. Therein, the pseudo-residuals vary around 0 and the variation increases with time, especially for male patients. Figure 5 of Web Appendix C presents residuals $\hat{M}_i(C_i; \hat{\gamma}_{\mathrm{SMR}}, \hat{\boldsymbol{\beta}}_{\mathrm{SMR}})$ plotted against the covariate sex and treatment respectively. The residual boxplots indicate that the proposed SMR model (3.3) is suitable for the readmission data since residuals symmetrically fluctuate around zero. The variation of residulas tend to be higher in male group and untreated group, which is consistent with the findings based on pseudo-residuals.

Figure 1 shows the estimated CRFs stratified by the treatment. Since they are so similar, we omited some curves and present only the estimates based on pseudo-observations along with the NA estimator. In the untreated group, the DR estimators (plain and longdash lines) are similar to the NA estimator (dotted line) while the estimator based on GLM is slightly higher. In the treated group, the curves based on the DR estimators are higher than that of the NA estimator, indicating a lower treatment effect after the adjustments. That is, the DR estimators reduce the confounding effect of sex and Duke's stage, and therefore provide a better estimation of CRF for the treated group than that of the NA estimator. Note that in both groups, the DR estimator with SL is overlapping with the DR estimator based on GLM.

To test $H_0 : \Lambda^1(t) = \Lambda^0(t)$, we conduct several tests based on chosen time points. We applied the two-sample tests described in Section 4 based on $B = 200$ bootstrap samples. Table 5 in Web Appendix C dispays the results. The test statistic $U_{\mathrm{R}}^{\mathrm{CLN}}(t)$ implies that the difference of CRFs between untreated and treated groups is identifiable after $t = 560.6$ at the 5% level. However, our proposed test statistics show that the difference of CRFs between two groups never reaches significance at the 5% level. This might be explained by the fact that our proposed test statistics are able to provide precise comparison between two CRFs as they are constructed based on robust DR estimators, which could reduce or eliminate the confounding biases. Also, the test results are consistent with our findings as shown in Figure 1 that the CRFs of treated group become higher after adjustment which shrinks the difference between two CRFs.

## 7. Discussion

In this work, we propose several estimators for the difference in the CRF of two groups whose assignment to a treatment may depend on confounders. Our proposals include IPTW estimator, OR model estimators and DR estimators. The proposed DR estimators are based on a combination of PS and OR models, and they are robust in the sense that they are consistent whenever either one of these models is correctly specified. Asymptotic properties of the estimators are discussed and the normality of the regression marginal estimators based on pseudo-observations are derived. To assess the adequency of the two proposed OR models, we develop two graphical model diagnosis tools. We also propose adjusted two-sample tests to compare two CRFs . Simulation studies show that the proposed methods perform well for finite sample scenarios. The proposed methodology is applied to a recurrent hospital readmission dataset for colorectal cancer patients.

Examples of pseudo-observations for CRF in Figure 1 of Web Appendix C show that pseudo-observations can take negative values, a behaviour akin to pseudo-observations for survival functions which are not necessarily within $(0, 1)$ as showed in **?**. This does not affect the consistency of the proposed DR estimators constructed based on the asymptotic property in (3.1). In addition, to avoid the possible collider issues as mentioned in Section 2, we consider the marginal rate model, SMR, with occurrence rate of recurrent events unconditional on the event history. According to **?**, marginal rate models are often preferred in practice because they provide more direct practical interpretations for identifying risk factors when comparing to the models conditioned on the event history.

As we showed in simulation studies, using 10 equally spaced time points yields good performances for the estimation of marginal parameters in (3.4) using pseudo-observations of CRF. A similar idea is recommended by **?** for who choosing 5 to 10 equally spaced time points works well for parameter estimation based on pseudo-observations of survival function. It might improve the efficiency of marginal estimator when using large number of points as mentioned by **?** and studied by **?**. However, the tradeoff between using a large versus small number of time points is that it may be time-consuming to solve the estimating equation (3.5) when the number of time points is large since it requires a large number of parameters in the intercept term $\boldsymbol{\beta}_H$. To obtain smooth looking curves of CRF for the readmission dataset, we utilize all observed event times as the time set to calculate pseudo-observations instead of using 10 time points. In practice, one could first adopt 10 equally spaced time points to estimate marginal parameters and plot the estimated CRF. If the estimated CRF does not look smooth enough, one can use all observed times as a time set to calculate the pseudo-observations and then obtain a smoother looking estimated CRF.

The simulation studies show that the IPTW estimator for $\theta(t)$ could be biased with small sample size although it is asymptotically unbiased. We thus suggest that researchers adopt that estimator only when the sample size is large ($n = 300$ from our simulation studies) to avoid the finite-sample bias caused by high variation of weights. The true form of PS model is almost never known in observational studies, which leads to bias estimation of $\theta(t)$ depending on the unknown extent of model misspecification. When the OR model is correctly specified, two estimators $\hat{\theta}_{\text{SMR}}(t)$ and $\hat{\theta}_{\text{PO}}(t)$ have smaller ESE and SEE compared to IPTW estimator and three DR estimators. We suggest that researchers use the G-formula estimators when the OR model can be correctly specified. In real-life settings, the true nature of the relationship between exposure and confounders with respect to the outcome is however never known. Model misspecification will result in biased estimators for $\theta(t)$. To mitigate the effects of misspecification of the PS or OR model, researchers can use the DR estimators which remain consistent if at least one of the PS or OR model is correctly specified. Besides, we would recommend that researchers use the DR estimator with SL approach as it provides substantially better performance.

In this work, we consider the situation where the treatment is independent of all other variables. However, some observational studies may just allow conditional independence of the post-treatment variables given pre-treatment variables. This will be an interesting topic in future work. Extending current work to admit time-varying confounders will also be future projects.

## References

Table 1. List of different models used in the definition of the proposed estimators and their corresponding censoring assumption (shown in the columns).

| | IPTW | G-formula | | Doubly robust | | |
|---|---|---|---|---|---|---|
| **Model** | $\hat{\theta}_{\text{IPTW}}$ | $\hat{\theta}_{\text{SMR}}$ | $\hat{\theta}_{\text{PO}}$ | $\hat{\theta}^{\text{SMR}}_{\text{DR}}$ | $\hat{\theta}^{\text{PO}}_{\text{DR}}$ | $\hat{\theta}^{\text{SL}}_{\text{DR}}$ |
| Logistic regression for PS | X | | | X | X | |
| Semiparametric multiplicative rate (SMR) | | X | | X | | |
| Generalized linear model (GLM) | | | X | | X | |
| Super Learner for PS | | | | | | X |
| Super learner for GLM | | | | | | X |
| **Censoring assumption** | | | | | | |
| Censoring completely at random | X | | X | | X | X |
| Censoring at random | | X | | X | | |

PO: pseudo-observations; DR: doubly robust; IPTW: inverse probability of treatment weighting; SL: super learner; Censoring completely at random: censoring is independent of all other variables; Censoring at random: censoring is independent of the outcome given the covariates

Table 2. Simulation summaries for doubly robust estimator based on 500 replications with $\mu(t) = 5t$ and $\sigma^2 = 0.25$.

| | | | **Estimators** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\theta(0.4) = -0.24$ | | | | | $\theta(0.8) = -0.48$ | | | | |
| $n$ | (PS,Reg) | | $\hat{\theta}_{\text{NA}}$ | $\hat{\theta}_{\text{IPTW}}$ | $\hat{\theta}_{\text{DR}}^{\text{SMR}}$ | $\hat{\theta}_{\text{DR}}^{\text{PO}}$ | $\hat{\theta}_{\text{DR}}^{\text{SL}}$ | $\hat{\theta}_{\text{NA}}$ | $\hat{\theta}_{\text{IPTW}}$ | $\hat{\theta}_{\text{DR}}^{\text{SMR}}$ | $\hat{\theta}_{\text{DR}}^{\text{PO}}$ | $\hat{\theta}_{\text{DR}}^{\text{SL}}$ |
| 100 | (C,C) | Bias | -0.135 | -0.015 | -0.008 | -0.008 | 0.010 | -0.254 | -0.047 | -0.034 | -0.031 | 0.003 |
| | | ESE | 0.334 | 0.447 | 0.399 | 0.403 | 0.335 | 0.706 | 0.838 | 0.775 | 0.782 | 0.669 |
| | | SEE | 0.339 | 0.429 | 0.385 | 0.389 | 0.324 | 0.674 | 0.859 | 0.760 | 0.773 | 0.652 |
| | | CR | 0.918 | 0.942 | 0.938 | 0.938 | 0.930 | 0.914 | 0.952 | 0.948 | 0.949 | 0.945 |
| | | MSER | 1.267 | 1.754 | 1.411 | 1.441 | 1(ref) | 1.221 | 1.741 | 1.362 | 1.407 | 1(ref) |
| | (C,N) | Bias | -0.115 | 0.028 | 0.016 | 0.011 | -0.058 | -0.212 | 0.010 | -0.013 | -0.026 | -0.090 |
| | | ESE | 0.343 | 0.373 | 0.369 | 0.371 | 0.352 | 0.688 | 0.776 | 0.750 | 0.763 | 0.730 |
| | | SEE | 0.334 | 0.383 | 0.372 | 0.378 | 0.333 | 0.668 | 0.795 | 0.790 | 0.792 | 0.712 |
| | | CR | 0.916 | 0.936 | 0.940 | 0.944 | 0.936 | 0.922 | 0.942 | 0.950 | 0.953 | 0.940 |
| | | MSER | 1.096 | 1.289 | 1.219 | 1.254 | 1(ref) | 0.953 | 1.227 | 1.211 | 1.217 | 1(ref) |
| | (N,C) | Bias | -0.125 | -0.048 | -0.020 | -0.020 | 0.013 | -0.232 | -0.071 | -0.044 | -0.045 | -0.011 |
| | | ESE | 0.342 | 0.319 | 0.308 | 0.308 | 0.303 | 0.713 | 0.686 | 0.664 | 0.665 | 0.615 |
| | | SEE | 0.337 | 0.310 | 0.298 | 0.298 | 0.289 | 0.668 | 0.678 | 0.594 | 0.595 | 0.585 |
| | | CR | 0.915 | 0.912 | 0.940 | 0.942 | 0.935 | 0.906 | 0.914 | 0.940 | 0.940 | 0.938 |
| | | MSER | 1.554 | 1.181 | 1.072 | 1.072 | 1(ref) | 1.462 | 1.357 | 1.038 | 1.041 | 1(ref) |
| 300 | (C,C) | Bias | -0.143 | 0.001 | 0.009 | 0.009 | -0.005 | -0.300 | 0.005 | -0.001 | -0.001 | -0.007 |
| | | ESE | 0.214 | 0.209 | 0.198 | 0.198 | 0.196 | 0.421 | 0.425 | 0.419 | 0.420 | 0.415 |
| | | SEE | 0.196 | 0.203 | 0.194 | 0.194 | 0.190 | 0.398 | 0.419 | 0.411 | 0.411 | 0.406 |
| | | CR | 0.874 | 0.946 | 0.950 | 0.948 | 0.949 | 0.870 | 0.936 | 0.948 | 0.948 | 0.946 |
| | | MSER | 1.639 | 1.139 | 1.027 | 1.027 | 1(ref) | 1.493 | 1.054 | 1.012 | 1.012 | 1(ref) |
| | (C,N) | Bias | -0.131 | -0.007 | -0.001 | -0.001 | -0.002 | -0.275 | -0.004 | 0.007 | 0.007 | 0.002 |
| | | ESE | 0.204 | 0.192 | 0.188 | 0.189 | 0.184 | 0.389 | 0.413 | 0.409 | 0.411 | 0.409 |
| | | SEE | 0.196 | 0.209 | 0.204 | 0.204 | 0.197 | 0.393 | 0.420 | 0.413 | 0.412 | 0.408 |
| | | CR | 0.888 | 0.960 | 0.956 | 0.950 | 0.956 | 0.896 | 0.950 | 0.960 | 0.956 | 0.955 |
| | | MSER | 1.474 | 1.157 | 1.078 | 1.078 | 1(ref) | 1.411 | 1.079 | 1.049 | 1.037 | 1(ref) |
| | (N,C) | Bias | -0.111 | -0.012 | -0.001 | -0.001 | -0.002 | -0.231 | -0.015 | -0.002 | -0.002 | -0.003 |
| | | ESE | 0.208 | 0.173 | 0.162 | 0.162 | 0.159 | 0.417 | 0.343 | 0.331 | 0.332 | 0.329 |
| | | SEE | 0.194 | 0.169 | 0.160 | 0.160 | 0.157 | 0.391 | 0.344 | 0.330 | 0.331 | 0.327 |
| | | CR | 0.908 | 0.916 | 0.950 | 0.950 | 0.948 | 0.886 | 0.915 | 0.950 | 0.950 | 0.948 |
| | | MSER | 2.028 | 1.166 | 1.041 | 1.041 | 1(ref) | 1.943 | 1.113 | 1.018 | 1.028 | 1(ref) |

Bias: bias of parameter estimator; ESE: empirical standard error of the parameter estimator; SEE: mean of the standard error estimator; CR: coverage rate of the 95% confidence interval; MSER: mean square error ratio NA: Nelson-Aalen estimator; IPTW: inverse probability of treatment weighting estimator; SMR: semiparametric multiplicative rate; PO: pseudo-observations; DR: doubly robust; SL: super learner; ref: reference; $B = 200$ for standard error estimation(SEE).

Table 3. Empirical type I error rates for test statistics based on 500 replications.

| | | | $\sigma^2 = 0$ | | | | | $\sigma^2 = 0.25$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (PS,Reg) | $\mu(t)$ | $n$ | $U_{\mathrm{P}}^{\mathrm{CLN}}$ | $U_{\mathrm{R}}^{\mathrm{CLN}}$ | $U_{\mathrm{DR}}^{\mathrm{SMR}}$ | $U_{\mathrm{DR}}^{\mathrm{PO}}$ | $U_{\mathrm{DR}}^{\mathrm{SL}}$ | $U_{\mathrm{P}}^{\mathrm{CLN}}$ | $U_{\mathrm{R}}^{\mathrm{CLN}}$ | $U_{\mathrm{DR}}^{\mathrm{SMR}}$ | $U_{\mathrm{DR}}^{\mathrm{PO}}$ | $U_{\mathrm{DR}}^{\mathrm{SL}}$ |
| (C,C) | 5t | 100 | 22.0 | 8.2 | 4.5 | 4.5 | 5.5 | 31.0 | 9.2 | 5.5 | 5.5 | 5.7 |
| | 5t | 200 | 22.0 | 10.8 | 4.9 | 5.0 | 5.2 | 25.0 | 9.6 | 5.2 | 5.2 | 4.9 |
| | 5t | 300 | 28.2 | 12.8 | 4.8 | 4.8 | 5.1 | 30.2 | 10.2 | 5.3 | 5.3 | 5.3 |
| | 20t | 100 | 41.6 | 12.0 | 5.5 | 5.5 | 4.6 | 51.0 | 9.4 | 4.8 | 4.7 | 4.8 |
| | 20t | 200 | 51.4 | 15.0 | 4.8 | 4.6 | 4.8 | 57.0 | 11.8 | 5.5 | 5.5 | 5.3 |
| | 20t | 300 | 58.8 | 18.6 | 5.1 | 4.9 | 5.2 | 58.4 | 13.2 | 5.2 | 5.3 | 5.1 |
| (C,N) | 5t | 100 | 18.0 | 9.0 | 4.5 | 4.5 | 4.5 | 23.2 | 8.6 | 5.1 | 5.1 | 5.2 |
| | 5t | 200 | 24.4 | 11.4 | 5.3 | 5.5 | 5.2 | 30.0 | 11.6 | 6.4 | 6.4 | 5.6 |
| | 5t | 300 | 27.8 | 13.4 | 4.6 | 4.6 | 4.8 | 34.0 | 10.6 | 5.4 | 5.2 | 5.2 |
| | 20t | 100 | 41.6 | 12.0 | 5.4 | 5.5 | 5.3 | 54.2 | 9.8 | 5.5 | 5.6 | 5.5 |
| | 20t | 200 | 48.2 | 14.6 | 5.2 | 5.3 | 5.4 | 55.4 | 12.2 | 5.8 | 5.8 | 5.4 |
| | 20t | 300 | 56.2 | 21.6 | 5.4 | 5.4 | 5.3 | 61.0 | 12.8 | 5.4 | 5.5 | 5.3 |
| (N,C) | 5t | 100 | 19.4 | 10.6 | 5.5 | 4.5 | 5.4 | 26.6 | 7.6 | 5.1 | 5.1 | 5.2 |
| | 5t | 200 | 25.6 | 12.0 | 5.2 | 5.2 | 5.4 | 30.0 | 10.0 | 4.6 | 4.6 | 5.4 |
| | 5t | 300 | 26.4 | 11.8 | 4.8 | 5.8 | 5.6 | 34.0 | 11.6 | 5.1 | 5.2 | 5.2 |
| | 20t | 100 | 40.2 | 11.4 | 5.0 | 5.0 | 5.2 | 46.6 | 9.2 | 4.5 | 4.5 | 4.6 |
| | 20t | 200 | 55.2 | 17.8 | 5.6 | 5.5 | 5.4 | 55.0 | 10.3 | 5.4 | 5.4 | 5.3 |
| | 20t | 300 | 59.6 | 20.6 | 5.5 | 5.5 | 5.4 | 60.8 | 15.8 | 5.3 | 5.3 | 5.4 |

The numbers for type I error rates are multiplied by 100. SMR: semiparametric multiplicative rate; PO: pseudo-observations; DR: doubly robust; SL: super learner; $U_{\mathrm{P}}^{\mathrm{CLN}}$: variance estimate is based on Poisson process assumption; $U_{\mathrm{R}}^{\mathrm{CLN}}$: variance estimate is based on robust to a departure from Poisson process assumption.

Table 4. Parameter estimates for hospital readmission data.

| | Logistic Model for PS | | | | | |
|---|---|---|---|---|---|---|
| **Parameter** | Est. | SEE | P-value | | | |
| Intercept | 1.024 | 0.186 | <0.001 | | | |
| Female | -0.256 | 0.220 | 0.245 | | | |
| DukeC | -1.771 | 0.245 | <0.001 | | | |
| DukeD | -0.563 | 0.288 | 0.051 | | | |
| | **Outcome Regression (OR) model** | | | | | |
| | **SMR Model** | | | **GLM Model** | | |
| **Parameter** | Est. | SEE | P-value | Est. | SEE | P-value |
| Chemo | -0.266 | 0.158 | 0.092 | -0.522 | 0.250 | 0.036 |
| Female | -0.495 | 0.166 | 0.003 | -0.427 | 0.209 | 0.041 |
| DukeC | 0.384 | 0.228 | 0.093 | 0.337 | 0.254 | 0.184 |
| DukeD | 1.514 | 0.218 | <0.001 | 1.230 | 0.240 | <0.001 |

Est.: Parameter estimate; SEE: Standard error estimate; PS: propensity score; SMR: semiparametric multiplicative rate; GLM: generalized linear model

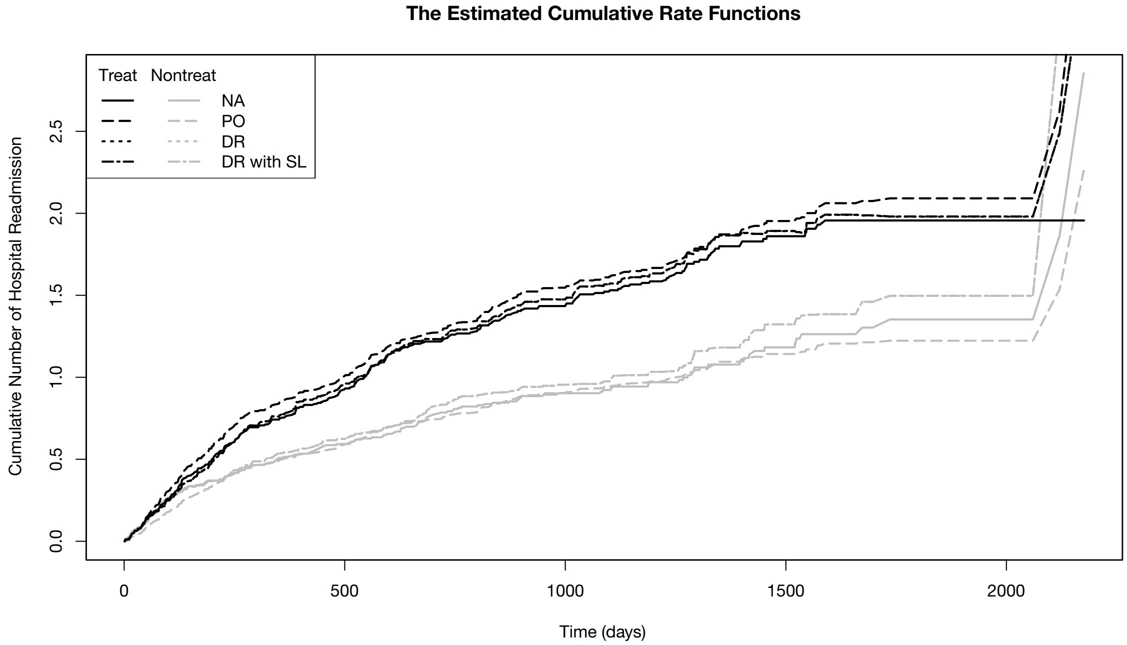

Fig. 1. The estimated cumulative rate functions for treat and nontreat groups from hospital readmission data using the Nelson-Aalen (NA) estimator, pseudo-observations (PO) estimator, doubly robust (DR) estimator based on pseudo-observations (PO) and doubly robust (DR) estimator based on pseudo-observations (PO) using Super Learner (SL).

[]