Published in final edited form as:

J Biopharm Stat. 2019 ; 29(4): 592–605. doi:10.1080/10543406.2019.1633659.

Bias-adjusted Kaplan-Meier Survival Curves for Marginal Treatment Effect in Observational Studies

Xiaofei Wang^{1,*}, Fangfang Bai², Herbert Pang³, Stephen L George¹

¹Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

²School of Statistics, University of International Business and Economics, Beijing, 100029, China

³Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China

Summary:

For time-to-event outcomes, the Kaplan-Meier estimator is commonly used to estimate survival functions of treatment groups and to compute marginal treatment effects, such as the difference in survival rates between treatments at a landmark time. The derived estimates of marginal treatment effect are uniformly consistent under general conditions when data are from randomized clinical trials. For data from observational studies, however, these statistical quantities are often biased due to treatment-selection bias. Propensity score-based methods estimate the survival function by adjusting for the disparity of propensity scores between treatment groups. Unfortunately, mis-specification of the regression model can lead to biased estimates. Using an empirical likelihood (EL) method in which the moments of the covariate distribution of treatment groups are constrained to equality, we obtain consistent estimates of the survival functions and the marginal treatment effect. Equating moments of the covariate distribution between treatment groups simulates the covariate distribution that would have been obtained if the patients had been randomized to these treatment groups. We establish the consistency and the asymptotic limiting distribution of the proposed EL estimators. We demonstrate that the proposed estimator is robust to model mis-specification. Simulation is used to study the finite sample properties of the proposed estimator. The proposed estimator is applied to a lung cancer observational study to compare two surgical procedures in treating early stage lung cancer patients.

Keywords

Balance of Covariates Distribution; Kaplan-Meier Curve; Marginal Treatment Effect; Observational Study; Subsets of Covariates; Survival Function; Time-To-Event Outcome

1. Introduction

Kaplan-Meier Curves and Measures of Treatment Effect

In cancer and other diseases, time to an event, such as time to death or time to treatment failure, is often used as the clinical outcome for measuring treatment effect. For example, in an observational CALGB lung cancer study (Nwogu et al., 2015), video-assisted thoracic

* xiaofei.wang@duke.edu.

surgery (VATS) and open lobectomy (Open) are two surgical procedures to be compared for their efficacy in treating early stage non-small cell lung cancer patients. The hazard ratio, often estimated via a Cox proportional hazards model, provides a relative measure of treatment effect. Absolute measures of treatment effect, including the difference in survival rates, are useful for evaluating treatment effect when the proportional hazard assumption is violated. Long-term survival effects as in the CALGB VATS vs. Open study and delayed treatment effects in immunotherapy trials are among the reasons that non-proportional hazards exist between treatment regimens in cancer research (Putter et al., 2005; Hoos et al., 2010; Chen, 2013). In these cases, the hazard ratio is no longer a suitable measure for treatment effect and an alternative measure is the difference in survival rates at a landmark time. In this article, we focus on using the Kaplan-Meier estimator (Kaplan and Meier, 1958) to estimate the overall survivals of two or more treatments and compare the difference in survival probabilities at a landmark time. With the CALGB VATS vs. Open study data, we estimate the Kaplan-Meier curves of VATS and Open and compare the survival probabilities of the two surgical procedures at a fixed landmark time.

As is often the case, a randomized clinical trial is not a viable option for comparing treatment regimens due to the high cost in time and resources or to the lack of equipoise for randomization. Real world evidence based on health care records and diseases registries, such as the CALGB VATS vs. Open study, is increasingly used to assess the safety and effectiveness of drugs and devices (Sherman et al., 2016). In the VATS vs. Open example, there are no randomized clinical trials comparing VATS with Open. The CALGB study was designed as a prospective registry trial to evaluate the relative efficacy of the two surgical procedures and standard Kaplan-Meier estimators were used to characterize the survival rates over time (Nwogu et al., 2015). Unfortunately, the Kaplan-Meier estimator based on observational data can be misleading due to unbalanced distribution of baseline covariates between treatment groups. For the VATS vs. Open comparison, it is known that surgeons tend to treat patients with smaller tumors with VATS and patients with larger tumors or more complicated medical conditions with Open. The survival benefit seen in the Kaplan-Meier estimates for VATS relative to Open seen in this study may well be attributed to healthier patients receiving VATS.

Inverse Probability of Treatment Weighting (IPTW) Estimators for Kaplan-Meier Curve

When estimating treatment effects with data from observational studies, it is important to use statistical methods to remove or adjust the effect of treatment-selection bias. Propensity score methods are among the most commonly used statistical methods to minimize the effect of this bias (Austin (2007), Austin et al. (2007), Austin (2011), Deb et al. (2016), Yao et al. (2017)). Once the propensity score is estimated, several methods, including matching on the propensity score (Rosenbaum and Rubin, 1985), stratification on the propensity score (Rosenbaum and Rubin, 1984), and inverse probability of treatment weighting (IPTW) using the propensity score (Rosenbaum, 1987), can be used to remove the confounding effects. Extensive reviews of these methods can be found in D'Agostino (1998) and Lunceford and Davidian (2004).

For survival endpoints, Xie and Liu (2005) proposed an adjusted Kaplan-Meier estimator of the survival function by incorporating IPTWs based on propensity score. Cole and Hernán (2004) proposed a similar inverse weighting of the survival function based on propensity score with a slightly different weight function. Although these methods are relatively easy to implement, several practical difficulties are associated with IPTW methods.

One obvious problem associated with IPTW is that a propensity score close to zero will give an extremely large weight leading to numerical instability and an inflated variance estimate. Another important but often ignored problem for these methods is that the propensity score must be correctly estimated. Defining a propensity score model is difficult, especially when the association of the covariates with treatment selection is not a linear function and involves higher order terms or interactions. It has been found that slight mis-specification of the propensity score model can result in substantial bias of the estimated treatment effects (e.g. Smith and Todd (2005); Kang and Schafer (2007)). The papers from Austin et al. (2007) and Imai and Ratkovic (2014) also demonstrate the importance of correct specification of the propensity score model.

For these reasons, it is important to develop a robust method for bias correction in estimating treatment effect, one not requiring correct specification of the propensity score model. Hirano et al. (2003) developed nonparametric propensity score estimators to provide valid inference on average treatment effect without relying on parametric assumptions in propensity score estimation. These methods require sieve approximations of unknown conditional functions, as these functions appear explicitly in the semiparametric efficient inference functions (Robins et al., 1994). Asymptotically, the estimators converge to the true propensity score function as the sample size increases, but no finite sample properties of these estimators are known and the implementation of the approach is computationally very difficult. Many recent methods focus on improving covariate balance within propensity score and outcome regression frameworks (Qin and Zhang, 2007; Chan et al., 2012; Han and Wang, 2013; Imai and Ratkovic, 2014). In principle, these methods could be extended to survival data in the CALGB VATS vs. Open study to estimate the survival functions and the absolute marginal treatment effect, but they all require parametric or nonparametric modeling of the propensity score or the outcome model and thus encounter similar theoretical and computational difficulties.

The Proposed Method

It is natural to question whether estimation of these propensity score functions is even necessary, and whether they can be replaced by directly balancing the distribution of baseline covariates.

The goal of this paper is to study an empirical likelihood-based method (e.g. Owen (2001); Qin (2017)) to enforce the equality of the covariate distributions, more specifically the equality of the moments of the covariates between treatment groups. After obtaining the empirical probability mass for each observation, we are able to estimate consistently the survival function for the treatment group and the control group as well as the marginal treatment effects between treatment groups. A Wald-type statistics can be used to test the absolute difference in survival rates between treatment groups at a pre-specified landmark

time. The resulting estimator is a nonparametric estimator for the marginal survival function. We also propose variance estimators for these statistical quantities.

The obvious advantage of the proposed approach is that there is no need to estimate a propensity score in estimating marginal treatment effect. The inverse weighting methods used to estimate the Kaplan-Meier curves require specification of either a propensity score model, an outcome model or both. Consistency of the estimators requires some underlying models to be correctly specified. Since the parameter of interest is the average treatment effect based on the Kaplan-Meier estimates, it is more natural and robust to directly balance the covariate distribution and to bypass the estimation of either the propensity score or the outcome models. The approach of balancing the covariate distribution between treatments has been discussed (e.g. Hellerstein and Imbens (1999); Hainmueller (2011); Chan et al. (2012, 2016)). In particular, Chan et al. (2016) proposed estimation of average treatment effects by empirical balancing calibration weights and demonstrated that the estimator can achieve global efficiency, which means the estimator is able to achieve semiparametric efficiency bounds without the requirement of correct specification of propensity score or outcome regression models. This method is related to ours, but the discussion is generic and does not cover the estimation of survival functions and absolute treatment effect based on survival functions.

The paper will be organized as follows. Section 2 provides notation and a description of the proposed empirical likelihood (EL) based method for estimating marginal survival functions. Section 3 establishes the large sample properties of the estimators for marginal survival function and the absolute difference in survival rates at time t . In Section 4, the finite sample properties of the proposed estimators, relative to the IPTW estimator and the standard Kaplan-Meier estimators are studied via simulations. In Section 5, the proposed method is illustrated using the CALGB lung cancer registry data to compare survival function between VATS and Open. The paper concludes with discussion of various issues in Section 6. The details of the asymptotic properties of the proposed estimators and their proofs are given in the online Supplementary Material.

2. Notation and the Proposed Estimator

Let Z be a $l \times 1$ vector of covariates, and D be an indicator of observed treatment status, where $D = 1$ for the treatment group, and $D = 0$ for the control group. Let T_1 be the survival time if a patient receives the treatment and T_0 be the survival time if a patient receives the control. The corresponding survival functions are denoted by $S_1(t)$ and $S_0(t)$ respectively. Note that, (T_1, T_0) cannot be observed simultaneously for an individual. For any individual in the treatment group, only T_1 can be observed; for any individual in the control group, only T_0 can be observed. Let C be the censoring time. Throughout this paper, independent censorship is assumed, i.e., C is independent of (T_1, T_0) . To estimate the difference between the two groups, strongly ignorable treatment assignment $(T_1, T_0) \perp\!\!\!\perp D|Z$ is assumed. The

parameter of interest is the difference of the survival rates at time t , which is defined by $\Delta(t) = S_1(t) - S_0(t)$.

In real world evidence studies, the covariate distribution between treatment groups are often unbalanced, leading to treatment-selection bias. Consequently, the standard Kaplan-Meier estimate of the survival function is biased in this situation. In order to reduce this selection bias, a new estimator for survival rate based on an empirical likelihood is proposed. For simplicity, we only discuss the estimation of $S_1(t)$; a similar derivation holds for $S_0(t)$.

Denote $T = DT_1 + (1 - D)T_0$. Furthermore, let $Y_1 = T_1 \wedge C, \delta_1 = I(T_1 \leq C)$ and $Y_0 = T_0 \wedge C, \delta_0 = I(T_0 \leq C)$. If $D_i = 1$, the observed data are $(Z_i, Y_{1i}, \delta_{1i})$; if $D_i = 0$, the observed data are $(Z_i, Y_{0i}, \delta_{0i})$.

Let $n_1 = \sum_{i=1}^n D_i$ and $n_0 = \sum_{i=1}^n (1 - D_i)$. For simplicity, denote the n_1 observations with $D_i = 1$ as $(Z_i, Y_{1i}, \delta_{1i})$, $i = 1, \dots, n_1$; the other n_0 observations with $D_i = 0$ as $(Z_i, Y_{0i}, \delta_{0i})$, $i = n_1 + 1, \dots, n$. Let $\bar{N}_1(u) = \sum_{i=1}^{n_1} I(Y_{1i} \leq u, \delta_{1i} = 1)$ and $\bar{Y}_1(u) = \sum_{i=1}^{n_1} I(Y_{1i} \geq u)$, then the Kaplan-Meier estimator of $S_1(t)$ (Kaplan and Meier, 1958) is defined as:

$$\hat{S}_1^{km}(t) = \prod_{u \leq t} \left\{ 1 - \frac{d\bar{N}_1(u)}{\bar{Y}_1(u)} \right\}.$$

However, $\hat{S}_1^{km}(t)$ is biased when there exists treatment-selection bias in the data.

We use an empirical likelihood approach to address treatment-selection bias. If $D_i = 1$, the observed data are $(Z_i, Y_{1i}, \delta_{1i})$, and the density function is denoted by q_i . If $D_i = 0$, the observed data are $(Z_i, Y_{0i}, \delta_{0i})$, and the density function is denoted by r_i . Let $p_i = q_i^{D_i} r_i^{(1-D_i)}$, $i = 1, \dots, n$. For the observed data, the likelihood function is

$$\mathcal{L} = \prod_{i=1}^n p_i.$$

Since the covariates Z are always observed for both the treatment group and the control group, we can estimate p_i by maximizing the likelihood function subject to the following constraints:

$$p_i \geq 0, \sum_{i=1}^n p_i D_i = 1, \sum_{i=1}^n p_i (1 - D_i) = 1, \quad (2.1)$$

$$\sum_{i=1}^n p_i D_i \phi(z_i) = \sum_{i=1}^n p_i (1 - D_i) \phi(z_i), \quad (2.2)$$

where $\phi(z) = (\phi(z_1), \phi(z_2), \dots, \phi(z_m))^T$ is a $m \times 1$ vector of independent known functions of the covariates z and is bounded.

The first constraint (2.1) reflects the fact that both q_i and r_i are density functions. The second constraint (2.2) enforces the equality of the moments of the covariates between treatment groups. Let $a = E\phi(z)$. Using the Lagrange multiplier algorithm, maximizing $\mathcal{L}$ subjecting to the constraints (2.1) and (2.2), we obtain

$$p_i D_i = \frac{D_i}{n_1 + \gamma^T (\phi(z_i) - a)}, \quad (2.3)$$

$$p_i (1 - D_i) = \frac{1 - D_i}{n_0 - \gamma^T (\phi(z_i) - a)}, \quad i = 1, \dots, n, \quad (2.4)$$

where γ and a satisfy the following estimating equations:

$$\sum_{i=1}^n \frac{D_i (\phi(z_i) - a)}{n_1 + \gamma^T (\phi(z_i) - a)} = 0 \quad (2.5)$$

$$\sum_{i=1}^n \frac{(1 - D_i) (\phi(z_i) - a)}{n_0 - \gamma^T (\phi(z_i) - a)} = 0. \quad (2.6)$$

Then a new estimator of the survival function $S_1(t)$, an empirical likelihood-based (EL-based) estimator $\hat{S}_1(t)$ is:

$$\hat{S}_1(t) = \prod_{u \leq t} \left\{ 1 - \frac{d\bar{N}_1^{EL}(u)}{\bar{Y}_1^{EL}(u)} \right\}, \quad (2.7)$$

where $\bar{N}_1^{EL}(u) = \sum_{i=1}^n p_i D_i I(Y_{1i} \leq u, \delta_{1i} = 1)$ and $\bar{Y}_1^{EL}(u) = \sum_{i=1}^n p_i D_i I(Y_{1i} \geq u)$. The estimator $\hat{S}_1(t)$ removes the selection bias.

The usual choice for $\phi(Z)$ is $\phi(Z) = Z$ or $\phi(Z) = (Z, ZZ^T)$, i.e., adjusting the first moment or the first and second moment of the covariates. For one dimensional variable Z , $\phi(Z) = Z$ forces mean equality between treatment groups and $\phi(Z) = (Z, Z^2, \dots, Z^k)$ ($k \geq 2$) forces the first k -th moment equality between treatment groups. An explanation of why treatment-selection bias can be removed by this approach may be seen in the following. Common distributions are determined completely by a finite number of parameters. The parameters usually can be estimated by moment estimation methods and are often consistent. In our approach, the constraint (2.2) with $\phi(Z) = (Z, Z^2, \dots, Z^k)$ implies that the moments of the covariates between treatment groups are the same. This will lead to equality of the distribution of the covariates between treatment groups. Thus, treatment-selection bias can be removed. A lower order k often is enough to remove most of the selection bias. For example, adjusting on the first moments of the covariates and the second moment of the joint distribution of all paired covariates is often sufficient in most applications.

In the next section, we discuss the large sample properties of $\hat{S}_1(t)$ including consistency and the limiting distribution.

3. Large Sample Properties for $\hat{S}_k(t)$ and $\hat{\Delta}(t)$

Under the assumptions in our paper, there exists a constant $c \in (0, 1)$, such that as $n \rightarrow \infty$, $n_1/n \rightarrow c$. Throughout the remainder of the paper, we let $\|\cdot\|$ denote Euclidean norm and $\psi(Z) = (1, \phi^T(Z))^T$. Furthermore, for any matrix B , B^T denotes the transpose of B and $\xrightarrow{\mathcal{D}}$ denotes “converges in distribution.”

Propensity score $e(Z) = P(D = 1|Z)$ describes the probability of treatment given the covariates Z . As in Rosenbaum and Rubin (1983), we assume that the propensity score is bounded away from 0 and 1, that is, there exists $\varsigma > 0$, such that $\varsigma < P(D = 1|Z) < 1 - \varsigma$ for any Z . We assume that the propensity score $e(Z)$ is contained in the linear space spanned by $\psi(Z)$. Note that this condition is used solely to derive the large sample properties and the proposed estimators do not rely on the value of the estimated propensity score. This condition precludes lengthy technical arguments and theoretical discussions. All the proofs of the theoretical results are given in the online Supplementary Material.

THEOREM 1: (Consistency theorem) T_1 is a failure time random variable with continuous survival function $S_1(s)$, and the distribution function is $F_1(s) = 1 - S_1(s)$. The cumulative hazard function is $\Lambda_1(s)$ and $\Lambda_1(s) = \int_0^s dF_1(v)/S_1(v)$. Let $\bar{Y}_1(s) = \sum_{i=1}^n 1_{I(Y_{1i} \geq s)}$ and $t \in (0, \infty]$ such that as $n_1 \rightarrow \infty$, $\bar{Y}_1(t) \xrightarrow{P} \infty$. Assume that $E[\psi(Z)\psi^T(Z)]$ is positive definite and $E\|\psi(Z)\|^3 < \infty$. The EL-based estimator $\hat{S}_1(t)$ is defined in (2.7), is an estimator of $S_1(t)$. Then

$$\sup_{0 \leq s \leq t} |\hat{S}_1(s) - S_1(s)| \xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty.$$

THEOREM 2: (Limiting theorem) Denote $\pi_1(t) = P(Y_1 \geq t)$ and $\mathcal{T}_1 = \{t : \pi_1(t) > 0\}$. Assume that $E[\psi(Z)\psi^T(Z)]$ is positive definite and $E\|\psi(Z)\|^3 < \infty$. Assume as $n_1 \rightarrow \infty$,

$\sup_{0 \leq t < \infty} \left| \frac{\bar{Y}_1(t)}{n_1} - \pi_1(t) \right| \xrightarrow{P} 0$, and $F_1(\cdot)$ is continuous. Then, as $n \rightarrow \infty$, for any $t \in \mathcal{T}_1$,

$$\sqrt{n} \left(\frac{\hat{S}_1(t) - S_1(t)}{S_1(t)} \right) \xrightarrow{\mathcal{D}} N(0, V_1), \quad (3.1)$$

where

$$V_1 = E \left[\int_0^t \frac{1}{\pi_1(s)} \frac{D}{e(Z)} dM_1(s) \right]^2 - B_1 A^{-1} B_1^T$$

$$\text{and } A = E \left[\frac{\psi(Z)\psi^T(Z)}{e(Z)(1-e(Z))} \right], B_1 = E \left[\frac{D\psi^T(Z)}{e^2(Z)} \int_0^t \frac{1}{\pi_1(s)} dM_1(s) \right],$$

$$M_1(s) = I(Y_1 \leq s, \delta_1 = 1) - \int_0^t I(Y_1 \geq s) d\Lambda_1(s).$$

Since V_1 involves the propensity score $e(Z)$ and, in general, we do not know the true propensity score model, a nonparametric bootstrap method is recommended to estimate the variance. If we know the true model for $e(Z)$, then by fitting the model based on the data, an estimate of $e(Z)$ can be obtained, and this estimate can be plugged into V_1 to obtain the estimated variance of $\hat{S}_1(t)$.

We can obtain the large sample properties for the estimate of $S_0(t)$ and the difference $\Delta(t)$ in the same way, where the estimates are defined as :

$$\hat{S}_0(t) = \prod_{u \leq t} \left\{ 1 - \frac{\sum_{i=1}^n p_i(1-D_i)I(Y_{0i} \leq u, \delta_{0i} = 1)}{\sum_{i=1}^n p_i(1-D_i)I(Y_{0i} \geq u)} \right\},$$

and

$$\hat{\Delta}(t) = \hat{S}_1(t) - \hat{S}_0(t).$$

We can obtain the large sample results for $\hat{S}_0(t)$. The derivations are similar and we omit them here.

THEOREM 3: (Limiting theorem) Denote $\pi_0(t) = P(Y_0 \geq t)$ and $\mathcal{T} = \{t : \pi_0(t) > 0\}$. Assume that $E[\psi(Z)\psi^T(Z)]$ is positive definite and $E\|\psi(Z)\|^3 < \infty$. Assume as $n_0 \rightarrow \infty$, $\sup_{0 \leq t < \infty} \left| \frac{\bar{Y}_0(t)}{n_0} - \pi_0(t) \right| \xrightarrow{P} 0$, and $S_0(\cdot)$ is continuous. The cumulative hazard function of T_0 is $\Lambda_0(s)$. Then, as $n \rightarrow \infty$, for any $t \in \mathcal{T}$,

$$\sqrt{n} \left(\frac{\hat{S}_0(t) - S_0(t)}{S_0(t)} \right) \xrightarrow{\mathcal{D}} N(0, V_0),$$

where

$$V_0 = E \left[\frac{1-D}{1-e(Z)} \int_0^t \frac{1}{\pi_0(s)} dM_0(s) \right]^2 - B_0 A^{-1} B_0^T$$

$$\text{and } A = E \left[\frac{\psi(Z)\psi^T(Z)}{e(Z)(1-e(Z))} \right], B_0 = E \left[\frac{(1-D)\psi^T(Z)}{(1-e(Z))^2} \int_0^t \frac{1}{\pi_0(s)} dM_0(s) \right],$$

$$M_0(s) = I(Y_0 \leq s, \delta_0 = 1) - \int_0^t I(Y_0 \geq s) d\Lambda_0(s).$$

Since $\hat{S}_1(t)$ and $\hat{S}_0(t)$ are correlated by some common covariates, the large sample property of $\hat{\Delta}(t) = \hat{S}_1(t) - \hat{S}_0(t)$ is complicated. Its large sample properties are given in the following theorem.

THEOREM 4: (Limiting theorem) $\pi_1(t)$, $\pi_0(t)$, $\Lambda_1(t)$ and $\Lambda_0(t)$ are defined as before.

$\mathcal{T} = \{t : \pi_1(t)\pi_0(t) > 0\}$. Assume that $E[\psi(Z)\psi^T(Z)]$ is positive definite and $E\|\psi(Z)\|^3 < \infty$.

Assume as $n \rightarrow \infty$, $\sup_{0 \leq t < \infty} \left| \frac{\bar{Y}_0(t)}{n_0} - \pi_0(t) \right| \xrightarrow{P} 0$, $\sup_{0 \leq t < \infty} \left| \frac{\bar{Y}_1(t)}{n_1} - \pi_1(t) \right| \xrightarrow{P} 0$. Both $S_1(\cdot)$ and $S_0(t)$ are continuous. Then, as $n \rightarrow \infty$, for any $t \in \mathcal{T}$,

$$\sqrt{n}(\hat{\Delta}(t) - \Delta(t)) \xrightarrow{\mathcal{D}} N(0, V),$$

where $B = S_1(t)B_1 + S_0(t)B_0$, V_0 , B_1 , B_0 are defined in Theorem 2 and Theorem 3 and

$$V = S_1^2(t)V_1 + S_0^2(t)V_0 + BA^{-1}B^T - 2S_1(t)BA^{-1}B_1^T - 2S_0(t)BA^{-1}B_0^T.$$

4. Simulation Studies

To illustrate the performance of the proposed EL method, we compare it with two existing methods: the inverse probability of treatment weighting (IPTW) by Xie and Liu (2005) and

the standard Kaplan-Meier method (KM). Our first simulation study is an example with a large selection bias and the second is one with a moderate selection bias.

Simulation study 1

First, we study a situation with a large selection bias. Consider covariates $Z = (Z_1, Z_2)^T$, where $Z_1 \sim U(0, 1)$, a uniform distribution on $[0, 1]$ and Z_2 follows a exponential distribution with mean 1. The true propensity score $e(Z)$ is

$$P(D = 1|Z) = \frac{\exp(0.8\exp(Z_1) - 1.4Z_2^{1.1})}{1 + \exp(0.8\exp(Z_1) - 1.4Z_2^{1.1})}.$$

In this model, the treatment variable is strongly related to the covariates and a large selection bias exists. For the treatment group, the hazard function of T_1 is $\lambda_1(t|Z) = \lambda_{10}(t)\exp(Z^{*T}\beta_1)$, where $Z^* = (Z_1, Z_2, Z_2)^T$ and $\lambda_{10}(t) = t\exp(-2)$. For the control group, the hazard function of T_0 is $\lambda_0(t|Z) = \lambda_{00}(t)\exp(Z^T\beta_0)$, where $Z = (Z_1, Z_2)^T$ and $\lambda_{00}(t) = t\exp(-2)$. The observed survival time is $T = DT_1 + (1 - D)T_0$. The censoring variable C follows a uniform distribution on $[0, c]$. The true value of parameters are $\beta_1 = (3, 1)^T$, $\beta_0 = (2, 3)^T$ and $c = 6.6$.

For a fixed point $t^* = 1.18$, we estimate the survival rates for the treatment group and the control group respectively and the difference in survival rates at t^* . Note that the choice of $t^* = 1.18$ is arbitrary; the choice of a different time point would produce similar results. The IPTW method requires specification of a model for the propensity score $e(Z)$. Since the true model is unknown, there is a possibility that it is misspecified. In the simulation, we incorrectly model the propensity score $e(Z) = P(D = 1|Z)$ directly on the covariates Z rather than Z^* by a logistic regression model. In our simulation, the sample size is 600 with $n_0 = n_1 = 300$. The simulation results given in Table 1 are based on 1000 simulations.

In the simulation tables, *bias* is the relative bias, *true* means the true value of the parameters, *se* is the standard deviation of the estimators in the simulation, *sd* is the mean of the standard deviation of the estimators, and *RMSE* is the root means square error. “*IPTW Correct*” refers to the IPTW estimator when the propensity score $e(Z)$ is correctly specified and “*IPTW Incorrect*” refers to the IPTW estimator when the propensity score $e(Z)$ is misspecified.

From Table 1, we can see that if selection bias is ignored, the standard Kaplan-Meier estimator is very biased; the relative bias of the difference $\Delta(t)$ is 0.912. When the propensity score model is correct, the IPTW method effectively removes bias and performs well. However, when the propensity score model is incorrect, the IPTW method is biased. For $\phi(z) = z$, adjusting only the first moment of the covariates, our proposed EL estimator removes completely the bias.

In Figure 1, we give the plots of the entire survival function $S_1(t)$, $S_0(t)$ and their difference $\Delta(t)$. The plots are based on 300 samples of size $n = 600$. The upper curves represents the curve of the survival function $S_1(t)$ of T_1 , and the lower curves represents the curve of the survival function $S_0(t)$ of T_0 . The difference in survival rates denotes $\Delta(t) = S_1(t) - S_0(t)$. The dot-dash line is obtained by the EL method, the dash line from the standard method, and the solid line represents the true curve. Figures 1 shows that the EL method yields satisfactory estimates of the true survival function and their difference, and they are much better than those from the standard KM estimators.

Simulation study 2

We next consider the situation when there is moderate selection bias in the data.

All the settings are identical to those in simulation study 1 except that the true propensity score $e(Z)$ is

$$P(D = 1|Z) = \frac{\exp(0.3\exp(Z_1) - 0.5Z_2^{1.1})}{1 + \exp(0.3\exp(Z_1) - 0.5Z_2^{1.1})}.$$

The simulation results are given in Table 2. The number of simulations is 1000. *CVP* is the 95% empirical coverage. As expected, the standard Kaplan-Meier estimator is biased, reflecting the moderate selection bias in the data. When treatment assignment model is correctly specified, the IPTW estimator performs well. There is always the possibility of model mis-specification, and when the treatment assignment model is misspecified, the IPTW estimator for the difference $\Delta(t)$ remains biased. Furthermore, the 95% empirical coverage for $S_1(t)$ and $S_0(t)$ is far from the nominal coverage.

For the EL estimator, we consider two cases. The first case $\phi(Z) = Z$ corresponds to adjusting only for the first moment of the covariates. The results shows that the proposed EL estimator is unbiased and the sd is not very far from the se. The 95% empirical coverages for the three parameters $S_1(t)$, $S_0(t)$ and $\Delta(t)$ are very close to the nominal coverage. The second case

$\phi(Z) = (Z, \quad ZZ^T)$ corresponds to adjusting for the first and second moments of the covariates. In this case, the EL estimator is unbiased and is more efficient than that of the first case with smaller standard deviation (sd). These results also demonstrate that adjusting a finite order moment of the covariates is ordinarily sufficient to remove the bias.

In simulation studies, we know the true propensity score function, so for simplicity we estimate $\hat{e}(z)$ according to the true propensity score model and then plug $\hat{e}(z)$ into the variance expressions in Section 3 to obtain an estimate of the variance. As noted previously, when the true treatment assignment model is unknown in data analysis, one can use a bootstrap method to estimate the variance. To demonstrate the performance of the bootstrap variance estimator, we give the result for the EL method adjusting for Z . With 300 bootstrap replications resampled from the data, the corresponding results are listed in Table 2, denoted as “EL adjust $z(\text{boot})$ ”. Compared to the simulated standard error, the bootstrap method

estimates the variance of the proposed estimator quite well. The results at other time points are similar but not shown here.

As shown in Table 2, there are no substantial differences in the performance of the EL estimator when either the first moments of Z or both the first and the second moments of Z are used. Additional simulations show that the EL estimator is more efficient than the KM estimator when there is no selection bias, and the results hold across a range of treatment assignment proportions $P(D = 1)$, time points t^* and sample sizes n . This finding suggests that the EL estimator is also an efficient method for covariate adjustment in randomized clinical trials where selection bias does not exist.

5. Data Example

In this section, we re-analyze the observational data extracted from the CALGB lung cancer registry (and originally reported by Nwogu et al. (2015)). There exists selection bias in the data and the proposed EL method is used to compare the two surgical procedures (VATS and Open) in treating non-small cell lung cancer.

The study cohort contains the patients who enrolled between October 2004 and June 2010 from 15 institutions contributing to the CALGB lung cancer registry. There are 519 observations collected from fifteen sites. Rhode Island only performed Open lobectomy (59 patients), all of which were performed by one surgeon. Cedars-Sinai only performed VATS lobectomy (92 patients), 85% of which were performed by one surgeon. Since the two institutions have $e(z)$ very close to 0 or 1, we exclude the observations from the two institutions in our analysis.

The observed covariates include age, gender, performance status, tumor size, insurance, race, any medical co-morbidities, pathologic stage, any symptoms and history of smoking and some other variables. The difference in covariates between the two groups is evaluated by a Fisher's exact test for binary variables and Wilcoxon rank sum test for continuous variables. An initial check for the balance of the covariates in the two groups reveals that the covariates which cause selection bias are tumor size, race, insurance, pathologic stage, history of smoking and any symptoms. The outcome of interest is disease-free survival time, time from the surgery to the initial disease recurrence, a commonly used endpoint in lung cancer studies.

As in the simulation study, we give the results for three estimators: the standard Kaplan-Meier (KM) estimator, the IPTW estimator and the EL estimator. For the IPTW estimator, we model the propensity score $e(Z)$ by a logistic model. An EL method adjusting for all the six covariates is used. The EL estimator is based on the first moment constraint of the covariates. The variance of the EL estimator is estimated by the bootstrap method with 300 bootstraps.

The results for the point estimates at times $(t_1, t_2, t_3) = (36, 48, 56)$, are given in Table 3, where $est.t_1$, $est.t_2$, and $est.t_3$ denote the estimator obtained by the three methods and $sd.t_1$, $sd.t_2$, and $sd.t_3$ denote the standard deviation derived by the bootstrap method. We also give the Z-ratio

for the difference in survival rates between VATS and Open, which are denoted as Z_{t_1} , Z_{t_2} and Z_{t_3} respectively. At $t_1 = 36$, the difference is significant by the unadjusted Kaplan-Meier estimator and it is non-significant for the IPTW estimator and EL estimator; and at $(t_2, t_3) = (48, 56)$, all three estimators are non-significant with a large bias in the Kaplan-Meier estimator. Thus the unadjusted estimator overestimates the survival difference between VATS and Open. The Kaplan-Meier estimator ignoring the bias yields different conclusions from the IPTW method or the EL method. Furthermore, the EL estimator of $\Delta(t)$ is more efficient than the IPTW estimator.

Figure 2 gives the estimated survival curves. There appears to be a difference between VATS and Open from the Kaplan-Meier (KM) estimator but no difference based on IPTW estimator and the EL estimator. The difference between VATS and Open by the EL method is smaller than that obtained by the IPTW method.

6. Discussion

In this paper, we propose an empirical likelihood (EL) based estimator that balances the covariate distribution by forcing equality of moments of the covariates. As the case for the Kaplan-Meier estimator, our EL estimator is a nonparametric method. Unlike Cole and Hernán (2004) and Xie and Liu (2005), the EL estimator allows counterfactual estimation on survival functions without the requirement of a correct specification of the propensity score model, an elusive goal with no effective tool to assess the lack of fit. The EL method involves estimating the EL-based weights through solving a system of equations for moment constraints and there exist cases where no valid solutions of these equations can be found. As with other propensity score-based methods, this issue can be mitigated by collecting more data or by ensuring significant data overlap between treatment groups. The EL method mimics randomized clinical trials through the constraints of the moments between treatment groups. The validity of the EL estimator has been proven under the assumption that the propensity score is contained in the linear space spanned by $\psi(Z)$. This assumption simplifies the proofs of the theorems, but the calculation of the EL estimator does not involve an estimation of the propensity score.

In principle, one could prove these theorems without the propensity score assumption, but additional conditions would need to be added to obtain the large sample properties. First, the propensity score could be approximated by a fully nonparametric sieve estimate. Then we could add some smooth conditions on the propensity score and some other nonparametric conditions. Next, based on nonparametric theory, a similar but different statement to this assumption could be obtained. Finally, the large sample properties could be derived. However, for the estimator to be consistent, the number of the constraints would go to infinity. This is impossible in applications. When doing data analysis, the first and second moment constraints usually are enough to remove the selection bias. This point is also elaborated in Chan et al. (2016), who used a nonparametric assumption on propensity score. Furthermore, the smooth conditions on propensity score are also difficult to check in applications. That makes the proof of the theorems more complicated and requires the

addition of other conditions which are equally difficult to check. For these reasons we added the propensity score assumptions.

The method can be easily extended to more than two treatment arms. The EL method can be applied to continuous outcomes or binary outcomes. It can also be modified to estimate average treatment effect in ATT, where ATT is the mean difference of two potential outcomes over the subpopulation of individuals who received the active treatment (Heckman et al., 1997; Imai and Ratkovic, 2014).

In principle, one can extend the estimator of Xie and Liu (2005) to a double-robustness estimator (augmented inverse probability weighted method, AIPW) using the general method suggested by Robins et al. (1994) (see also Lunceford and Davidian (2004); Tsiatis (2006)). The AIPW estimator (augmented inverse probability weighted method) is doubly robust in that the estimator is consistent when either the propensity score model is correctly specified or the model for the outcome as a function of the covariates is correctly specified. However, Kang and Schafer (2007) conducted a set of simulation studies to study the performance of propensity score weighting methods. They found that mis-specification of a propensity score model can negatively affect the performance of various weighting methods. In particular, they showed that although the doubly robust estimator of Robins et al. (1994) provides a consistent estimate of the treatment effect if either the outcome model or the propensity score model is correct, the performance of the doubly robust estimator can deteriorate when both models are slightly misspecified. In this paper, we have not discussed the semiparametric efficiency of the empirical likelihood method, though our simulation demonstrated efficiency compared to the propensity score-based method. Some authors, including Hirano et al. (2003), Qin and Zhang (2007), Hainmueller (2011), Chan et al. (2012), Chan et al. (2016), have discussed the semiparametric efficiency of empirical likelihood-based estimators in different but related settings. Whether our Kaplan-Meier estimator reaches the semiparametric bound and at what rate are topics for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank the two reviewers for their valuable comments that have led to significant improvement of the manuscript. Xiaofei Wang's work was supported by NIA R21-AG042894 and NCI P01-CA142538. Fangfang Bai's work was supported by National Natural Science Foundation of China (NSFC) (11501104) and by Program for Young Excellent Talents, UIBE (17YQ06). Herbert Pang's work was supported by NIA R21-AG042894.

REFERENCES

- Austin PC (2007). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* 26, 3078–94. [PubMed: 17187347]
- Austin PC (2011). A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research* 46, 119–151. [PubMed: 22287812]

- Austin PC, Grootendorst P, and Anderson GM (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in Medicine* 26, 734–753. [PubMed: 16708349]
- Chan KCG et al. (2012). Uniform improvement of empirical likelihood for missing response problem. *Electronic Journal of Statistics* 6, 289–302.
- Chan KCG, Yam SCP, and Zhang Z (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society* 78, 673. [PubMed: 27346982]
- Chen TT (2013). Statistical issues and challenges in immuno-oncology. *Journal for Immunotherapy of Cancer* 1, 18. [PubMed: 24829754]
- Cole SR and Hernán MA (2004). Adjusted survival curves with inverse probability weights. *Computer Methods & Programs in Biomedicine* 75, 45–49. [PubMed: 15158046]
- D'Agostino RB (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17, 2265–2281. [PubMed: 9802183]
- Deb S, Austin PC, Tu JV, Ko DT, Mazer CD, Kiss A, and Fremes SE (2016). A review of propensity-score methods and their use in cardiovascular research. *Canadian Journal of Cardiology* 32, 259–265. [PubMed: 26315351]
- Hainmueller J (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* page mpr025.
- Han P and Wang L (2013). Estimation with missing data: beyond double robustness. *Biometrika* 100, 417–430.
- Heckman JJ, Ichimura H, and Todd PE (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies* 64, 605–654.
- Hellerstein JK and Imbens GW (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics* 81, 1–14.
- Hirano K, Imbens GW, and Ridder G (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.
- Hoos A, Eggermont AM, Janetzki S, Hodi FS, Ibrahim R, Anderson A, Humphrey R, Blumenstein B, Old L, and Wolchok J (2010). Improved endpoints for cancer immunotherapy trials. *Journal of the National Cancer Institute*
- Imai K and Ratkovic M (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 243–263.
- Kang JDY and Schafer JL (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22, 574–580.
- Kaplan EL and Meier P (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457–481.
- Lunceford JK and Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23, 2937–60. [PubMed: 15351954]
- Nwogu CE, D'Cunha J, Pang H, Gu L, Wang X, Richards WG, Veit LJ, Demmy TL, Sugarbaker DJ, and Kohman LJ (2015). VATS lobectomy has better perioperative outcomes than Open lobectomy: CALGB 31001, an ancillary analysis of CALGB 140202 (Alliance). *Annals of Thoracic Surgery* 99, 399–405. [PubMed: 25499481]
- Owen AB (2001). *Empirical likelihood* CRC press.
- Putter H, Sasako M, Hartgrink H, Van de Velde C, and Van Houwelingen J (2005). Long-term survival with non-proportional hazards: results from the dutch gastric cancer trial. *Statistics in Medicine* 24, 2807–2821. [PubMed: 16134134]
- Qin J (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond* Springer.
- Qin J and Lawless J (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* pages 300–325.

- Qin J and Zhang B (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69, 101–122.
- Ridgeway G and McCaffrey DF (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22, 540–543.
- Robins JM, Rotnitzky A, and Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89, 846–866.
- Rosenbaum PR (1987). Model-based direct adjustment. *Journal of the American Statistical Association* 82, 387–394.
- Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Rosenbaum PR and Rubin DB (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–524.
- Rosenbaum PR and Rubin DB (1985). The bias due to incomplete matching. *Biometrics* pages 103–116. [PubMed: 4005368]
- Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA, et al. (2016). Real-world evidence-what is it and what can it tell us. *N Engl J Med* 375, 2293–2297. [PubMed: 27959688]
- Smith JA and Todd PE (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125, 305–353.
- Tsiatis A (2006). *Semiparametric theory and missing data* New York: Springer-Verlag.
- White H (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society* 50, 1–25.
- Xie J and Liu C (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 24, 3089–3110. [PubMed: 16189810]
- Yao XI, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH, Berry MF, Schrag D, and Pang HH (2017). Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. *JNCI: Journal of the National Cancer Institute* 109, djw323.

Figure 1.
Simulation study 1 – estimates of survival functions.

Figure 2.
Survival function estimates for CALGB lung cancer data.

Table 1

Results of simulation study 1 (large selection bias)

method	para	true	est.hat	bias	se	sd	RMSE
KM	$S_1(t)$	0.504	0.626	0.242	0.03	0.029	0.059
KM	$S_0(t)$	0.265	0.169	0.362	0.023	0.023	0.132
KM	$\Delta(t)$	0.239	0.457	0.912	0.038	0.037	0.833
IPTW Correct	$S_1(t)$	0.504	0.52	0.032	0.036	0.032	0.002
IPTW Correct	$S_0(t)$	0.265	0.267	0.007	0.034	0.036	0.001
IPTW Correct	$\Delta(t)$	0.239	0.253	0.06	0.047	0.048	0.006
IPTW Incorrect	$S_1(t)$	0.504	0.557	0.104	0.036	0.033	0.012
IPTW Incorrect	$S_0(t)$	0.265	0.234	0.116	0.032	0.034	0.014
IPTW Incorrect	$\Delta(t)$	0.239	0.322	0.349	0.047	0.048	0.124
EL adjust for Z	$S_1(t)$	0.504	0.502	0.003	0.037	0.035	0.001
EL adjust for Z	$S_0(t)$	0.265	0.268	0.013	0.039	0.034	0.002
EL adjust for Z	$\Delta(t)$	0.239	0.234	0.021	0.05	0.05	0.003

bias: the relative bias; true: the true value of the parameters; se: the standard deviation of the estimators in the simulation; sd: the mean of the standard deviation of the estimators; RMSE is the root means square error; IPTW Correct: refers to the IPTW estimator when the propensity score $e(Z)$ is correctly specified; IPTW Incorrect: refers to the IPTW estimator when the propensity score $e(Z)$ is misspecified.

Table 2

Results of simulation study 2 (moderate selection bias)

method	para	true	est.hat	bias	se	sd	RMSE	CVP
KM	$S_1(t)$	0.504	0.566	0.122	0.031	0.03	0.126	43.6
KM	$S_0(t)$	0.265	0.223	0.161	0.025	0.025	0.163	60.3
KM	$\Delta(t)$	0.239	0.343	0.437	0.04	0.039	0.439	26.1
IPTW correct	$S_1(t)$	0.504	0.507	0.005	0.029	0.03	0.029	95.6
IPTW correct	$S_0(t)$	0.265	0.266	0.003	0.028	0.029	0.028	96
IPTW correct	$\Delta(t)$	0.239	0.241	0.008	0.038	0.042	0.039	97.5
IPTW incorrect	$S_1(t)$	0.504	0.521	0.034	0.03	0.03	0.045	91.1
IPTW incorrect	$S_0(t)$	0.265	0.254	0.04	0.028	0.029	0.049	93.5
IPTW incorrect	$\Delta(t)$	0.239	0.266	0.116	0.039	0.041	0.123	91.2
EL adjust for Z	$S_1(t)$	0.504	0.504	0	0.027	0.029	0.027	96.5
EL adjust for Z	$S_0(t)$	0.265	0.266	0.002	0.028	0.028	0.028	94.9
EL adjust for Z	$\Delta(t)$	0.239	0.238	0.002	0.036	0.038	0.036	95.6
EL adjust for $Z(\text{boot})$	$S_1(t)$	0.504	0.503	0.002	0.027	0.027	0.028	94.4
EL adjust for $Z(\text{boot})$	$S_0(t)$	0.265	0.264	0.003	0.028	0.028	0.028	95.3
EL adjust for $Z(\text{boot})$	$\Delta(t)$	0.239	0.239	0.001	0.037	0.036	0.037	94.3
EL adjust for (Z, ZZ^T)	$S_1(t)$	0.504	0.505	0.003	0.027	0.026	0.028	93.5
EL adjust for (Z, ZZ^T)	$S_0(t)$	0.265	0.261	0.015	0.026	0.026	0.03	95.2
EL adjust for (Z, ZZ^T)	$\Delta(t)$	0.239	0.244	0.023	0.033	0.033	0.041	94.0

EL adjust $z(\text{boot})$: the variance is estimated by a bootstrapped method; CVP: the 95% empirical coverage

Table 3Lung cancer example: Disease-free survival at (t_1, t_2, t_3)

method	para	est. t_1	sd. t_1	Z_{t_1}	est. t_2	sd. t_2	Z_{t_2}	est. t_3	sd. t_3	Z_{t_3}
KM	$S_1(t)$	0.637	0.035		0.543	0.037		0.493	0.039	
KM	$S_0(t)$	0.534	0.038		0.468	0.038		0.438	0.039	
KM	$\Delta(t)$	0.104	0.052	2	0.075	0.053	1.42	0.055	0.055	1
IPTW	$S_1(t)$	0.568	0.039		0.483	0.039		0.443	0.04	
IPTW	$S_0(t)$	0.577	0.049		0.511	0.051		0.485	0.052	
IPTW	$\Delta(t)$	-0.01	0.063	0.15	-0.028	0.064	0.44	-0.041	0.066	0.63
EL	$S_1(t)$	0.573	0.043		0.489	0.04		0.448	0.041	
EL	$S_0(t)$	0.57	0.042		0.503	0.043		0.474	0.043	
EL	$\Delta(t)$	0.003	0.061	0.05	-0.013	0.061	0.22	-0.026	0.056	0.46