Statistica Sinica **33** (2023), 2161-2184
doi:https://doi.org/10.5705/ss.202021.0175

PROPENSITY SCORE WEIGHTING ANALYSIS OF SURVIVAL OUTCOMES USING PSEUDO-OBSERVATIONS

Shuxi Zeng[1], Fan Li[1], Liangyuan Hu[2] and Fan Li[3]

[1]*Duke University,* [2]*Rutgers School of Public Health and* [3]*Yale School of Public Health*

Abstract: Survival outcomes are common in comparative effectiveness studies and require unique handling, because they are usually incompletely observed owing to right-censoring. A "once for all" approach for causal inference with survival outcomes constructs pseudo-observations and allows standard methods such as propensity score weighting to proceed as if the outcomes are completely observed. For a general class of model-free causal estimands with survival outcomes on user-specified target populations, we develop corresponding propensity score weighting estimators based on such pseudo-observations and establish their asymptotic properties. In particular, using the functional delta method and the von Mises expansion, we derive a new closed-form variance of the weighting estimator that takes into account the uncertainty due to both the pseudo-observation calculation and the propensity score estimation. This allows for a valid and computationally efficient inference, without resampling. We also prove the optimal efficiency property of the overlap weights within the class of balancing weights for survival outcomes. The proposed methods are applicable to both binary and multiple treatments. Extensive simulations are conducted to explore the operating characteristics of the proposed method versus other commonly used alternatives. We apply the proposed method to compare the causal effects of three popular treatment approaches for prostate cancer patients.

Key words and phrases: Balancing weights, causal inference, multiple treatments, overlap weights, survival analysis.

1. Introduction

Survival or time-to-event outcomes are common in comparative effectiveness research and require unique handling, because they are usually incompletely observed owing to right-censoring. In observational studies, a popular approach to drawing causal inferences with survival outcomes is to combine standard survival estimators with propensity score methods (Rosenbaum and Rubin (1983)). For example, one can construct a Kaplan–Meier estimator on an inverse probabil-

Corresponding author: Fan Li, Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520-8034, USA. E-mail: fan.f.li@yale.edu.

ity weighted sample to adjust for measured confounding (Robins and Finkelstein (2000); Hubbard, Van Der Laan and Robins (2000)). Another common approach combines the Cox model with inverse probability weighting (IPW) to estimate the hazard ratio (Austin and Stuart (2017)) or the counterfactual survival curves (Cole and Hernán (2004)). Combining a causal inference with the Cox model introduces two limitations. First, the Cox model assumes proportional hazards in the target population, the violation of which leads to biased causal estimates. Second, the target estimand is usually the hazard ratio, the interpretation of which can be opaque owing to the built-in selection bias (Hernán (2010)). In contrast, estimands based on a survival probability or restricted mean survival time are free of model assumptions and have a natural causal interpretation (Mao et al. (2018)).

To analyze observational studies with survival outcomes, an attractive alternative approach is to combine causal inference methods with *pseudo-observations* (Andersen, Klein and Rosthøj (2003)). Each pseudo-observation is constructed based on a jackknife statistic, and is interpreted as the individual contribution to the target estimate from a complete sample without censoring. This approach addresses censoring in a "once for all" manner, and allows standard methods to proceed as if the outcomes are completely observed (Andersen, Hansen and Klein (2004)). To this end, one can perform a direct confounding adjustment using an outcome regression with pseudo-observations, and derive causal estimators using the g-formula (Robins (1986)). Another approach is to combine propensity score weighting with pseudo-observations. Andersen, Syriopoulou and Parner (2017) considered an IPW estimator to estimate the causal risk difference and difference in restricted mean survival time. Their approach has since been further extended to doubly robust estimation with survival and recurrent event outcomes (Wang (2018); Su, Platt and Plante (2020)).

Despite its simplicity and versatility, several open questions in propensity score weighting with pseudo-observations remain to be addressed. First, pseudo-observations require computing a jackknife statistic for each unit, which poses computational challenges to resampling-based variance estimation under propensity score weighting (Andersen, Syriopoulou and Parner (2017)). On the other hand, failing to account for the uncertainty when estimating the propensity scores and jackknifing can lead to inaccurate and often conservative variance estimates. Second, the IPW estimator with pseudo-observations corresponds to a target population represented by the study sample, but interpreting such a population is often questionable in the case of a convenience sample (Li, Thomas and Li (2019)). Moreover, the inverse probability weights are prone to a lack of covariate overlap,

and engender causal estimates with excessive variance, even when combined with an outcome regression (Mao, Li and Greene (2019)). Li, Morgan and Zaslavsky (2018) proposed a general class of balancing weights (which includes the IPW as a special case) for defining target estimands on user-specified target populations. In particular, the overlap weights emphasize the target population with the most covariate overlap and best clinical equipoise, and are theoretically shown to provide the most efficient causal contrasts. However, the theory of overlap weights has thus far focused on noncensored outcomes, and its optimal variance property is unclear with survival outcomes. Third, many comparative effectiveness studies involve multiple treatments, which can exacerbate the consequence of a lack of overlap when the IPW is considered (Yang et al. (2016)). While overlap weights (Li and Li (2019)) improve the bias and efficiency over those of the IPW with noncensored outcomes, extensions to censored survival outcomes remain limited, with the exception of the work of Cheng et al. (2022) for binary treatments.

We address the above questions. We consider a general multiple-treatment setup and extend the balancing weights in Li, Morgan and Zaslavsky (2018) and Li and Li (2019) to analyze survival outcomes in observational studies based on pseudo-observations. We develop new asymptotic variance expressions for causal effect estimators that account for the variability when estimating propensity scores and constructing pseudo-observations. In contrast to existing variance expressions developed for propensity score weighting estimators (Lunceford and Davidian (2004); Mao et al. (2018)), our new asymptotic variance expression is developed based on the functional delta method and the von Mises expansion (Graw, Gerds and Schumacher (2009); Jacobsen and Martinussen (2016); Overgaard, Parner and Pedersen (2017)), which are uniquely required in this context, because the pseudo-observations are themselves estimated using jackknifing. Such asymptotic results also enable a valid and computationally efficient inference, without resampling. Based on the new asymptotic variance expression, we further prove that overlap weights lead to the most efficient survival causal estimators, expanding the theoretical underpinnings of overlap weights to causal survival analysis. We carry out simulations to evaluate and compare a range of commonly used weighting estimators. Finally, we apply the proposed method to data from the National Cancer Database to estimate the causal effects of three treatments on the mortality of patients with high-risk localized prostate cancer.

2. Propensity Score Weighting with Survival Outcomes

2.1. Time-to-event outcomes, causal estimands, and assumptions

We consider a sample of N units drawn from a population. Let $Z_i \in \mathcal{J} = \{1, 2, \ldots, J\}, J \geq 2$ denote the assigned treatment. Each unit has a set of potential outcomes $\{T_i(j), j \in \mathcal{J}\}$, measuring the counterfactual survival time mapped to each treatment. We similarly define $\{C_i(j), j \in \mathcal{J}\}$ as a set of potential censoring times. Under the stable unit treatment value assumption (SUTVA), $T_i = \sum_{j\in\mathcal{J}} \mathbf{1}\{Z_i = j\}T_i(j)$ and $C_i = \sum_{j\in\mathcal{J}} \mathbf{1}\{Z_i = j\}C_i(j)$. Because of right-censoring, we might only observe the lower bound of the survival time for some units. We express the observed failure time as $\widetilde{T}_i = T_i \wedge C_i$, the censoring indicator as $\Delta_i = \mathbf{1}\{T_i \leq C_i\}$, and the p-dimensional time-invariant pre-treatment covariates, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})' \in \mathcal{X}$. In summary, we observe the tuple $\mathcal{O}_i = (Z_i, \mathbf{X}_i, \widetilde{T}_i, \Delta_i)$ for each unit. We define the generalized propensity score, $e_j(\mathbf{X}_i) = \Pr(Z_i = j|\mathbf{X}_i)$, as the probability of receiving treatment j, given baseline covariates (Imbens (2000). Our results are presented for general and finite $J \geq 2$.

The causal estimands of interest are based on two typical transformations of the potential survival times: (i) the at-risk function, $\nu_1(T_i(j); t) = \mathbf{1}\{T_i(j) \geq t\}$, and (ii) the truncation function, $\nu_2(T_i(j); t) = T_i(j) \wedge t$, where t is a given time point of interest. The identity function is implied by $\nu_2(T_i(j); \infty) = T_i(j)$. To simplify the discussion, we use $k \in \{1, 2\}$ to index the choice of the transformation function v. We further define $m_j^k(\mathbf{X}; t) = \mathbb{E}\{\nu_k(T_i(j); t)|\mathbf{X}\}$ as the conditional expectation of the transformed potential survival outcome, and the pairwise conditional causal effect at time t as $\tau_{j,j'}^k(\mathbf{X}; t) = m_j^k(\mathbf{X}; t) - m_{j'}^k(\mathbf{X}; t)$, for $j \neq j' \in \mathcal{J}$. We are interested in the conditional causal effect, averaged over a *target population*. We assume the study sample is drawn from a population with covariate density $f(\mathbf{X})$ (with respect to a measure $\mu(\cdot)$), and represent the target population by the density $g(\mathbf{X})$. The function $h(\mathbf{X}) \propto g(\mathbf{X})/f(\mathbf{X})$ is a tilting function, that re-weights the observed sample to represent the target population. The *pairwise average causal effect* at time t on the target population is defined as

$$\tau_{j,j'}^{k,h}(t) = \frac{\int_{\mathcal{X}} \tau_{j,j'}^k(\mathbf{X}; t) f(\mathbf{X}) h(\mathbf{X}) \mu(d\mathbf{X})}{\int_{\mathcal{X}} f(\mathbf{X}) h(\mathbf{X}) \mu(d\mathbf{X})}, \quad \forall \quad j \neq j' \in \mathcal{J}. \tag{2.1}$$

The class of estimands (2.1) is transitive in the sense that $\tau_{j,j'}^{k,h}(t) = \tau_{j,j''}^{k,h}(t) + \tau_{j'',j'}^{k,h}(t)$. Different choices of the function ν_k lead to estimands on different scales. When $k = 1$, we refer to estimand (2.1) as the survival probability causal effect

(SPCE). This estimand represents the causal risk difference, and contrasts the potential survival probabilities at time t among the target population. When $k = 2$, estimand (2.1) is referred to as the restricted average causal effect (RACE), which compares the mean potential survival times restricted by t. When $t = \infty$, this estimand becomes the average survival causal effect (ASCE) that compares the unrestricted mean potential survival times. However, because the observed data do not contain information beyond the maximum follow-up time $t_{\max}$, one can at most identify $\tau_{j,j'}^{k=2,h}(t_{\max})$. With a sufficiently long follow-up time, as in our data example, a practical solution is to estimate $\tau_{j,j'}^{k=2,h}(t_{\max})$ (RACE at time $t_{\max}$) as an approximation of the ASCE. In this sense, the following inferential details for the RACE still apply to the ASCE. In our simulations, we also examine the accuracy of this strategy when estimating the ASCE. Finally, when $J = 2$, the estimands in (2.1) reduce to those in Mao et al. (2018) for binary treatments.

To identify the estimands in (2.1), we maintain several assumptions. For each $j \in \mathcal{J}$, we assume the following: (A1) weak unconfoundedness: $T_i(j) \perp\!\!\!\perp \mathbf{1}\{Z_i = j\}|\mathbf{X}_i$; (A2) overlap: $0 < e_j(\mathbf{X}) < 1$, for any $\mathbf{X} \in \mathcal{X}$; and (A3) completely independent censoring: $\{T_i(j), Z_i, \mathbf{X}_i\} \perp\!\!\!\perp C_i(j)$. Assumptions (A1) and (A2) are the usual no unmeasured confounding and positivity conditions typically invoked for multiple treatments (Imbens (2000); Yang et al. (2016)), and allow us to identify $\tau_j^{k,h}(t)$ in the absence of censoring. Assumption (A3) assumes that censoring is independent of all remaining variables, and is introduced for now as a convenient technical device to establish our main results. (A3) often holds, for example, when the failure times are subject only to administrative right censoring. We relax this assumption in Sections 3 and 4 to enable identification under a weaker condition that assumes (A4) covariate-dependent censoring: $T_i(j) \perp\!\!\!\perp C_i(j)|\mathbf{X}_i, Z_i = j$.

2.2. Balancing weights with pseudo-observations

We now introduce balancing weights to estimate the causal estimands (2.1). Write $f_j(\mathbf{X}) = f(\mathbf{X}|Z = j)$ as the conditional density of the covariates among the treatment group j over $\mathcal{X}$. It follows immediately that $f_j(\mathbf{X}) \propto f(\mathbf{X})e_j(\mathbf{X})$. For any prespecified tilting function $h(\mathbf{X})$, we weight the group-specific density to the target population density using the following balancing weights, up to a proportionality constant:

$$w_j^h(\mathbf{X}) \propto \frac{g(\mathbf{X})}{f_j(\mathbf{X})} \propto \frac{f(\mathbf{X})h(\mathbf{X})}{f(\mathbf{X})e_j(\mathbf{X})} = \frac{h(\mathbf{X})}{e_j(\mathbf{X})}, \quad \forall \ \ j \in \mathcal{J}. \tag{2.2}$$

The weights $\{w_j^h(\mathbf{X}) : j \in \mathcal{J}\}$ balance the weighted distributions of the pretreatment covariates toward the corresponding target population distribution, that is, $f_j(\mathbf{X})w_j^h(\mathbf{X}) \propto g(\mathbf{X})$, for all $j \in \mathcal{J}$.

To apply the balancing weights to survival outcomes subject to right-censoring, we first construct the pseudo-observations (Andersen, Klein and Rosthøj (2003)). For a given time t, we generically define $\theta^k(t) = \mathbb{E}\{\nu_k(T_i;t)\}$ as a population parameter. The pseudo-observation for each unit is written as $\widehat{\theta}_i^k(t) = N\widehat{\theta}^k(t) - (N-1)\widehat{\theta}_{-i}^k(t)$, where $\widehat{\theta}^k(t)$ is the consistent estimator of $\theta^k(t)$, and $\widehat{\theta}_{-i}^k(t)$ is the corresponding estimator with unit i left out. For the transformation ν_k $(k = 1, 2)$, we employ the Kaplan–Meier estimator to construct $\theta^k(t)$, given by $\widehat{S}(t) = \prod_{\widetilde{T}_i \le t}\{1 - (dN(\widetilde{T}_i)/Y(\widetilde{T}_i))\}$, where $N(t) = \sum_{i=1}^N \mathbf{1}\{\widetilde{T}_i \le t, \Delta_i = 1\}$ is the counting process for the event of interest, and $Y(t) = \sum_{i=1}^N \mathbf{1}\{\widetilde{T}_i \ge t\}$ is the at-risk process. When the interest lies in the survival functions $(k = 1)$, the ith pseudo-observation is estimated as $\widehat{\theta}_i^1(t) = N\widehat{S}(t) - (N-1)\widehat{S}_{-i}(t)$. When the interest lies in the restricted mean survival times $(k = 2)$, the ith pseudo-observation is estimated as $\widehat{\theta}_i^2(t) = N\int_0^t \widehat{S}(u)du - (N-1)\int_0^t \widehat{S}_{-i}(u)du = \int_0^t \widehat{\theta}_i^1(u)du$. The pseudo-observation uses a leave-one-out jackknife approach to address right-censoring, and provides a straightforward unbiased estimator of the functional of uncensored data under the independent censoring assumption (A3). From Graw, Gerds and Schumacher (2009) and Andersen, Syriopoulou and Parner (2017), and under the unconfoundedness assumption (A1), one can show that $\mathbb{E}\{\widehat{\theta}_i^k(t)|\mathbf{X}_i, Z_i = j\} \approx \mathbb{E}\{\nu_k(T_i;t)|\mathbf{X}_i, Z_i = j\} = \mathbb{E}\{\nu_k(T_i(j);t)|\mathbf{X}_i\}$, based on which the g-formula can be used to estimate the pairwise average causal effect on the overall population $(h(\mathbf{X}) = 1)$. For the class of estimands (2.1), we further propose the following nonparametric Hájek-type estimator:

$$\widehat{\tau}_{j,j'}^{k,h}(t) = \frac{\sum_{i=1}^N \mathbf{1}\{Z_i = j\}\widehat{\theta}_i^k(t)w_j^h(\mathbf{X}_i)}{\sum_{i=1}^N \mathbf{1}\{Z_i = j\}w_j^h(\mathbf{X}_i)} - \frac{\sum_{i=1}^N \mathbf{1}\{Z_i = j'\}\widehat{\theta}_i^k(t)w_{j'}^h(\mathbf{X}_i)}{\sum_{i=1}^N \mathbf{1}\{Z_i = j'\}w_{j'}^h(\mathbf{X}_i)}. \quad (2.3)$$

The estimator (2.3) compares the weighted average pseudo-observations in each treatment group. First, without censoring, the ith pseudo-observation is simply a transformation of the observed outcome $\nu_k(T_i;t)$, and (2.3) is identical to the estimator in Li and Li (2019) for complete outcomes. Second, a number of weighting schemes proposed for noncensored outcomes are applicable to (2.3). For example, the IPW estimator considers $h(\mathbf{X}) = 1$ and $w_j^h(\mathbf{X}) = 1/e_j(\mathbf{X})$, corresponding to a target population of the combination of all treatment groups represented by the study sample. In this case, when only $J = 2$ treatments are present, estimator (2.3) reduces to the IPW estimator in

Andersen, Syriopoulou and Parner (2017). When the target population is the group receiving treatment l (similar to the average treatment effects for the treated estimand in binary treatments), the corresponding $h(\mathbf{X}) = e_l(\mathbf{X})$ and the balancing weight is $w_j^h(\mathbf{X}) = e_l(\mathbf{X})/e_j(\mathbf{X})$. The overlap weights (OW) specify $h(\mathbf{X}) = \left\{\sum_{l\in\mathcal{J}} e_l^{-1}(\mathbf{X})\right\}^{-1}$ and $w_j^h(\mathbf{X}) = e_j^{-1}(\mathbf{X})\left\{\sum_{l\in\mathcal{J}} e_l^{-1}(\mathbf{X})\right\}^{-1}$, and correspond to the target population as an intersection of all treatment groups with an optimal covariate overlap (Li and Li (2019)). This overlap population mimics that enrolled in a randomized trial, and emphasizes units with treatment decisions that are most ambiguous. When different groups have good covariate overlap, OW and IPW correspond to an almost identical target population and estimands. The difference in the target population and estimands between OW and IPW emerges with increasing regions of poor overlap. Specifically, as $e_j(\mathbf{X})$ approaches zero, $w_j^h(\mathbf{X})$ under IPW increases to infinity, whereas $w_j^h(\mathbf{X})$ under OW approaches zero. Owing to such intrinsic differences in the construction of the weights, OW is expected to improve on the efficiency of IPW, and should be less susceptible to bias caused by extreme propensity scores. In the case of a complete outcome, OW has been proved to give the smallest total variance for pairwise comparisons among all balancing weights. However, the theory and optimality of OW, has not been explored with survival outcomes, and thus is investigated below.

3. Theoretical Properties

We present two main results on the theoretical properties of the proposed weighting estimator (2.3). The first result develops a new asymptotic variance expression for the weighted pairwise comparisons of the pseudo-observations, and the second result establishes the efficiency optimality of OW within the family of balancing weights based on the pseudo-observations.

Below, we first outline the main steps of deriving the asymptotic variance. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $(\mathbf{D}, \|\bullet\|)$ be a Banach space for distribution functions. Specifically, we choose the Banach space to the space of functions of bounded p-variation, and the corresponding norm $\|\bullet\|$ is the p-variation norm; see example 3.2 in Overgaard, Parner and Pedersen (2017) for the regularity details. We assume each tuple $\mathcal{O}_i = (Z_i, \mathbf{X}_i, \widetilde{T}_i, \Delta_i)$ is an independent and identically distributed (i.i.d.) draw from the sample space $\mathcal{S}$ in the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Defining the Dirac measure $\delta_{(\bullet)} : \mathcal{S} \to \mathbf{D}$, we write the *empirical distribution function* as $F_n = N^{-1}\sum_{i=1}^{N} \delta_{\mathcal{O}_i}$ and its limit as F. Following Overgaard, Parner and Pedersen (2017), we use functionals to represent differ-

ent estimators for the transformed survival outcomes with pseudo-observations. Suppose $\phi_k(\bullet;t) : \mathbf{D} \to \mathcal{R}$ is the functional mapping of a distribution to a real value, such as the Kaplan–Meier estimator, $\phi_1(F_N;t) = \widehat{S}(t)$. Then each pseudo-observation is represented as $\widehat{\theta}_i^k(t) = N\phi_k(F_N;t) - (N-1)\phi_k(F_N^{-i};t)$, where F_N^{-i} is the empirical distribution omitting $\mathcal{O}_i$.

To derive the asymptotic variance of estimator (2.3), we need to accommodate two sources of uncertainty. The first source stems from the calculation of the pseudo-observations. We consider the functional derivative of $\phi_k(\bullet;t)$ at $f \in \mathbf{D}$ along direction $s \in \mathbf{D}$ as $\phi'_{k,f}(s)$, which is a linear and continuous functional, $\{\phi_k(f+s;t) - \phi_k(f;t) - \phi'_{k,f}(s;t)\}^2 = o(||s||_{\mathbf{D}})$. Assuming $\phi_k(\bullet;t)$ is differentiable at the true distribution function F, we express the first-order influence function of $\mathcal{O}_i$ for the pseudo-observation estimator $\hat{\theta}^k(t)$ as the first-order derivative along the direction $\delta_{\mathcal{O}_i} - F$, denoted by $\phi'_{k,i}(t) \triangleq \phi'_{k,F}(\delta_{\mathcal{O}_i} - F;t)$. Similarly, the second-order derivative for the functional $\phi_k(\bullet;t)$ at f along direction (s,w) can be defined as $\phi''_{k,F}(s,w;t)$, and the second-order influence function for $(\mathcal{O}_i,\mathcal{O}_j)$ is given as $\phi''_{k,(l,i)}(t) \triangleq \phi''_{k,F}(\delta_{\mathcal{O}_l} - F, \delta_{\mathcal{O}_i} - F;t)$. To characterize the variability associated with jackknifing, we follow Graw, Gerds and Schumacher (2009) and Jacobsen and Martinussen (2016) to write the second-order von Mises expansion of the pseudo-observations as:

$$\widehat{\theta}_i^k(t) = \theta^k(t) + \phi'_{k,i}(t) + \frac{1}{N-1}\sum_{l\neq i}\phi''_{k,(l,i)}(t) + R^k_{N,i}, \tag{3.1}$$

where the first three terms dominate the asymptotic behavior of $\widehat{\theta}_i^k(t)$, and the remainder $R^k_{N,i}$ vanishes asymptotically because $\lim_{N\to 0}\sqrt{N}\max_i|R^k_{N,i}| = 0$, for any k. The second source of uncertainty in estimator (2.3) comes from estimating the unknown propensity scores, and hence the weights; such uncertainty is well studied in the causal inference literature, and is usually quantified using M-estimation (e.g., see Lunceford and Davidian (2004)). Typically, the unknown propensity score model is parameterized as $e_j(\mathbf{X}_i;\boldsymbol{\gamma})$, where the parameter $\boldsymbol{\gamma}$ is estimated by maximizing the multinomial likelihood.

Theorem 1. *Under suitable regularity conditions, specified in Appendix A in the Supplementary Materials, for $k = 1,2$, $j,j' \in \mathcal{J}$ and all continuously differentiable tilting functions $h(\mathbf{X})$, (a) $\widehat{\tau}^{k,h}_{j,j'}(t)$ is a consistent estimator for $\tau^{k,h}_{j,j'}(t)$; (b) $\sqrt{N}\{\widehat{\tau}^{k,h}_{j,j'}(t) - \tau^{k,h}_{j,j'}(t)\}$ converges in distribution to a mean-zero normal random variate with variance $\mathbb{E}\{\Psi_j(\mathcal{O}_i;t) - \Psi_{j'}(\mathcal{O}_i;t)\}^2/\{\mathbb{E}(h(\mathbf{X}_i))\}^2$, where*

$$\Psi_j(\mathcal{O}_i;t) = \mathbf{1}\{Z_i = j\}w_j^h(\mathbf{X}_i)\left\{\left(\theta^k(t) + \phi'_{k,i}(t) - m_j^{k,h}(t)\right) + Q_i\right\} \tag{3.2}$$

$$+\mathbb{E}\left\{\mathbf{1}\{Z_i=j\}\left(\theta^k(t)+\phi'_{k,i}(t)-m_j^{k,h}(t)\right)\frac{\partial}{\partial\gamma^T}w_j^h(\mathbf{X}_i)\right\}\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{S}_{\boldsymbol{\gamma},i},$$

$Q_i=(N-1)^{-1}\sum_{l\neq i}\phi''_{k,(l,i)}(t)\mathbf{1}\{Z_l=j\}w_j^h(\mathbf{X}_l)$, *and* $\mathbf{S}_{\boldsymbol{\gamma},i}$ *and* $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ *are the score function and information matrix of* $\boldsymbol{\gamma}$*, respectively.*

Theorem 1 establishes the consistency and asymptotic normality of the proposed weighting estimator (2.3). In particular, the influence function $\Psi_j(\mathcal{O}_i;t)$ delineates the two aforementioned sources of variability, with the first and second terms characterizing the uncertainty due to estimating the pseudo-observations and the propensity scores, respectively. The jackknife pseudo-observation estimator for $\widehat{\theta}_i^k(t)$ includes information from the remaining $N-1$ observations, and thus is no longer independent across units. Therefore, the derivation of (3.2) requires invoking the central limit theorem for U-statistics (e.g., van der Vaart (1998, Chap. 12), and leads to a second-order term, Q_i, that properly accommodates the correlation between the estimated pseudo-observations of different units. Theorem 1 immediately suggests the following consistent variance estimator for pairwise comparisons, $\widehat{\mathbb{V}}\{\widehat{\tau}_{j,j'}^{k,h}(t)\}=\sum_{i=1}^N\{\widehat{\Psi}_j(\mathcal{O}_i;t)-\widehat{\Psi}_{j'}(\mathcal{O}_i;t)\}/\sum_{i=1}^N\widehat{h}(\mathbf{X}_i)^2$, where $\widehat{\Psi}_j(\mathcal{O}_i;t)$ is defined explicitly in Appendix A. In Appendix A, we also give explicit derivations of the functional derivatives for each transformation ν_k when the Kaplan–Meier estimator is used to construct the pseudo-observations, as in Section 2.2. This new closed-form estimator enables a fast computation of the variance of the weighting estimator (2.3) without resampling, a crucial advantage when the sample size is large.

Several important remarks related to Theorem 1 are in order.

Remark 1. Without censoring, each pseudo-observation degenerates to the observed outcome, which implies $\widehat{\theta}_i^k(t)=\theta^k(t)+\phi'_{k,i}(t)=\nu_k(T_i;t)$, and therefore $Q_i=0$. In this case, formula (3.2) coincides with the influence function derived in Li and Li (2019) for complete outcomes.

Remark 2. In the presence of censoring, we show in Appendix A that, some what counterintuitively, ignoring the uncertainty from estimating the pseudo-observations *overestimates* the variance of $\widehat{\tau}_{j,j'}^{k,h}(t)$. This insight for the weighting estimator supports the findings of Jacobsen and Martinussen (2016), who suggest that ignoring the uncertainty from estimating the pseudo-observations leads to a conservative inference for the outcome regression coefficients.

Remark 3. For $h(\mathbf{X})=1$ (and equivalently the IPW scheme), we show in Appendix A that treating the inverse probability weights as known, also counterintuitively, *overestimates* the variance for pairwise comparisons; this extends the

classic results of Hirano, Imbens and Ridder (2003) to multiple treatments. However, the implications of ignoring the uncertainty when estimating the propensity scores are, in general, uncertain for other choices of $h(\mathbf{X})$, which can lead to either conservative or anti-conservative inferences, as mentioned in Haneuse and Rotnitzky (2013). An exception is the randomized controlled trial (RCT), where the propensity score to any treatment group is a constant, and that is, any tilting function based on the propensity scores reduces to a constant, i.e. $h(\mathbf{X}) = \widetilde{h}(e_1(\mathbf{X}), \ldots, e_j(\mathbf{X})) \propto 1$. In this case, one can still estimate a "working" propensity score model, and use the subsequent weighting estimator (2.3) to adjust for a chance imbalance in the covariates. Equation (3.2) shows that such a covariate adjustment approach in an RCT leads to variance reduction for pairwise comparisons, extending the results developed in Zeng et al. (2020) to multiple treatments and censored survival outcomes.

Remark 4. Estimator (2.3) and Theorem 1 can be extended to accommodate covariate-dependent censoring: $T_i(j) \perp\!\!\!\perp C_i(j)|\mathbf{X}_i, Z_i$. In this case, one can consider the inverse probability of a censoring-weighted pseudo-observation (Robins and Finkelstein (2000); Binder, Gerds and Andersen (2014)):

$$\widehat{\theta}_i^k(t) = \frac{\nu_k(\widetilde{T}_i; t)\mathbf{1}\{C_i \geq \widetilde{T}_i \wedge t\}}{\widehat{G}(\widetilde{T}_i \wedge t|\mathbf{X}_i, Z_i)}, \tag{3.3}$$

where $\widehat{G}(u|\mathbf{X}_i, Z_i)$ is a consistent estimator of the censoring survival function $G(u|\mathbf{X}_i, Z_i) = \Pr(C_i \geq u|\mathbf{X}_i, Z_i)$, for example, given by the Cox proportional hazards regression. Other possible types of pseudo-observations exist to adjust for dependent censoring (Binder, Gerds and Andersen (2014)). We select (3.3) to simplify the computation, especially when deriving the consistent variance estimator. To show the consistency and asymptotic normality of the modified weighting estimator, we can similarly view (3.3) as a functional mapping from the empirical distribution of the data to a real value (Overgaard, Parner and Pedersen (2019)), and find the corresponding functional derivatives for the asymptotic expansion (see Appendix A).

Theorem 2 shows that the overlap weights, similarly to the case of noncensored outcomes, lead to the smallest total asymptotic variance for all pairwise comparisons based on the pseudo-observations among the family of balancing weights.

Theorem 2. *Under the regularity conditions in Appendix A, and assuming generalized homoscedasticity such that* $\lim_{N\to\infty} \mathbb{V}\{\hat{\theta}_i^k(t)|Z_i, \mathbf{X}_i\} = \mathbb{V}\{\phi'_{k,i}(t)|Z_i, \mathbf{X}_i\}$

is a constant across different levels of $(Z_i, \mathbf{X}_i)$*, the harmonic mean function* $h(\mathbf{X}) = \left\{\sum_{l\in\mathcal{J}} e_l^{-1}(\mathbf{X})\right\}^{-1}$ *leads to the smallest total asymptotic variance for pairwise comparisons among all tilting functions.*

Theorem 2 generalizes the findings of Crump et al. (2006), Li, Morgan and Zaslavsky (2018), and Li and Li (2019) to provide new theoretical justification for the efficiency optimality of the overlap weights, $w_j^h(\mathbf{X}) = e_j(\mathbf{X})\left\{\sum_{l\in\mathcal{J}} e_l^{-1}(\mathbf{X})\right\}^{-1}$, when applied to censored survival outcomes. Technically, this result relies on a generalized homoscedasticity assumption that requires the limiting variance of the estimated pseudo-observations to be constant within the strata defined by $(Z_i, \mathbf{X}_i)$. This condition includes the usual homoscedasticity for conditional outcome variance as a special case in the absence of censoring. Note that, the homoscedasticity condition may not hold in practice. However, this has been shown empirically not to be crucial for the efficiency property of OW, as exemplified in the simulations by Li, Morgan and Zaslavsky (2018) and numerous applications. Furthermore, in Section 4, we carry out extensive simulations to verify that OW leads to improved efficiency over that of IPW when generalized homoscedasticity is violated.

We can further augment estimator (2.3) using an outcome regression model of the pseudo-observations. Specifically, for any time t, we can posit treatment-specific outcome models $m_j^k(\mathbf{X}_i;\boldsymbol{\alpha}_j) = \mathbb{E}\{\widehat{\theta}_i^k(t)|\mathbf{X}_i, Z_i = j\}$, and define an augmented weighting estimator

$$\begin{aligned}\widehat{\tau}_{j,j',\mathrm{AUG}}^{k,h}(t) =& \frac{\sum_{i=1}^N \widehat{h}(\mathbf{X}_i)\{m_j(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}}_j) - m_{j'}(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}}_{j'})\}}{\sum_{i=1}^N \widehat{h}(\mathbf{X}_i)} + \\ & \frac{\sum_{i=1}^N \mathbf{1}\{Z_i = j\}\{\widehat{\theta}_i^k(t) - m_j(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}}_j)\}w_j^h(\mathbf{X}_i)}{\sum_{i=1}^N \mathbf{1}\{Z_i = j\}w_j^h(\mathbf{X}_i)} - \\ & \frac{\sum_{i=1}^N \mathbf{1}\{Z_i = j'\}\{\widehat{\theta}_i^k(t) - m_{j'}(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}}_{j'})\}w_{j'}^h(\mathbf{X}_i)}{\sum_{i=1}^N \mathbf{1}\{Z_i = j'\}w_{j'}^h(\mathbf{X}_i)},\end{aligned} \tag{3.4}$$

where $\widehat{\boldsymbol{\alpha}}_j$ denotes the estimated regression parameters in the jth outcome model. Such an augmented estimator generalizes those developed in Mao, Li and Greene (2019) to multiple treatments and survival outcomes. When $h(\mathbf{X}) = 1$, that is, with the IPW scheme, the augmented estimator becomes the doubly robust estimator for pairwise comparisons. When only $J = 2$ treatments are compared, (3.4) reduces to the estimator of Wang (2018), and provides an alternative to other doubly robust estimators studied in, for example, Zhang and Schaubel (2012). For other choices of $h(\mathbf{X})$, the augmented estimator is not necessarily

doubly robust, but may be more efficient than weighting alone when the outcome model is correctly specified (Mao, Li and Greene (2019)). For specifying an outcome regression model, Andersen and Pohar Perme (2010) reviewed a set of generalized linear models appropriate for pseudo-observations, and discussed residual-based diagnostic tools for checking model adequacy. We follow their strategies, and assume the outcome model is $m_j(\mathbf{X}_i;\boldsymbol{\alpha}_j) = g^{-1}(\mathbf{X}_i^T\boldsymbol{\alpha}_j)$, where g is a link function. Estimating $\boldsymbol{\alpha}_j$ can proceed using standard algorithms for fitting generalized linear models. For our estimands of interest, we can choose the identity or log link to estimate the ASCE and RACE, and the complementary log-log link (resembling a proportional hazards model) for the SPCE (Andersen, Hansen and Klein (2004)). Compared with Theorem 1 for the weighting estimator (2.3), deriving the asymptotic variance of (3.4) requires considering a third source of uncertainty from estimating $\boldsymbol{\alpha}_j$ in the outcome model. We outline the key derivation steps in Appendix A.

4. Simulation Studies

Simulation design. We conduct simulation studies to evaluate the finite-sample performance of the weighting estimator (2.3), and to illustrate the efficiency property of the OW estimator. We generate four pre-treatment covariates: $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})^T$, where $(X_{1i}, X_{2i})^T$ are drawn from a mean-zero bivariate normal distribution with variance two and correlation 0.25, $X_{3i} \sim \text{Bern}(0.5)$, and $X_{4i} \sim \text{Bern}(0.4 + 0.2X_{3i})$. We consider $J = 3$ treatments, with the true propensity score model given by $\log\{e_j(\mathbf{X}_i)/e_1(\mathbf{X}_i)\} = \widetilde{\mathbf{X}}_i^T\boldsymbol{\beta}_j$, for $j = 1, 2, 3$, where $\widetilde{\mathbf{X}}_i = (1, \mathbf{X}_i^T)^T$. We set $\boldsymbol{\beta}_1 = (0, 0, 0, 0, 0)^T$ and $\boldsymbol{\beta}_2 = 0.2\boldsymbol{\beta}_3$. Two sets of values for $\boldsymbol{\beta}_3$ are considered: (i) $\boldsymbol{\beta}_3 = (-0.4, 0.85, 0.9, 0.45, -0.25)^T$, and (ii) $\boldsymbol{\beta}_3 = (1.2, 1.5, 1, -1.5, -1)^T$, which represent good and poor covariate overlap across groups, respectively. The distributions of the true generalized propensity scores under each specification are presented in Figure 1 in the Supplementary Material.

Two outcome models are used to generate potential survival times. Model A is a Weibull proportional hazards model, with the hazard rate for $T_i(j)$ given as $\lambda_j(t|\mathbf{X}_i) = \eta\nu t^{\nu-1}\exp\{L_i(j)\}$, and $L_i(j) = \mathbf{1}\{Z_i = 2\}\gamma_2 + \mathbf{1}\{Z_i = 3\}\gamma_3 + \mathbf{X}_i^T\boldsymbol{\alpha}$. We specify $\eta = 0.0001$, $\nu = 3$, $\boldsymbol{\alpha} = (0, 2, 1.5, -1, 1)^T$, and $\gamma_2 = \gamma_3 = 1$, implying a worse survival experience from treatments $j = 2$ and $j = 3$. The potential survival time is drawn using $T_i(j) = \{-\log(U_i)/(\eta\exp(L_i(j)))\}^{1/\nu}$, where $U_i \sim \text{Unif}(0, 1)$. Model B is an accelerated failure time model that violates the proportional hazards assumption. Specifically, $T_i(j)$ is drawn from a log-normal distribution

$\log\{T_i(j)\} \sim \mathcal{N}(\mu, \sigma^2 = 0.64)$, with $\mu = 3.5 - \gamma_2 \mathbf{1}\{Z_i = 2\} - \gamma_3 \mathbf{1}\{Z_i = 3\} - \mathbf{X}_i^T \boldsymbol{\alpha}$. For simplicity, we assume that the treatment has no causal effect on the censoring time, such that $C_i(j) = C_i$, for all $j \in \mathcal{J}$. Under completely independent censoring, $C_i \sim \text{Unif}(0, 115)$. Under covariate-dependent censoring, C_i is generated from a Weibull survival model with hazard rate $\lambda^c(t|\mathbf{X}_i) = \eta_c \nu_c t^{\nu_c - 1} \exp(\mathbf{X}_i^T \boldsymbol{\alpha}_c)$, where $\boldsymbol{\alpha}_c = (1, 0.5, -0.5, 0.5)^T$, $\eta_c = 0.0001$, and $\nu_c = 2.7$. These parameters are specified so that the marginal censoring rate is roughly 50%. Neither data-generating process assumes generalized homoscedasticity in Theorem 2. Thus, both provide an objective evaluation of the efficiency property of OW.

Under each data-generating process, we consider the OW and IPW estimators based on (2.3), and focus our comparison on two standard estimators: the g-formula estimator based on the confounder-adjusted Cox model, and the IPW-Cox model (Austin and Stuart (2017)). Details of these and other alternative estimators are included in Appendix B of the Supplementary Materials. Whereas the IPW estimator (2.3) and the Cox model-based estimators focus on the combined population with $h(\mathbf{X}) = 1$, the OW estimator focuses on the overlap population with the optimal tilting function suggested in Theorem 2. When comparing treatments $j = 2$ (or $j = 3$) with $j = 1$, the true values of the target estimands can differ between OW and the other estimators (albeit very similar under good overlap), and are computed using Monte Carlo integration. Nonetheless, when we compare treatments $j = 2$ and $j = 3$, the true conditional average effect $\tau_{2,3}^k(\mathbf{X}; t) = 0$ for all k, and thus the true estimand $\tau_{2,3}^{k,h}(t)$, has the same value (zero), regardless of $h(\mathbf{X})$. This represents a natural scenario for comparing the bias and efficiency between estimators, without differences in the true values of the estimands. We vary the study sample size $N \in \{150, 300, 450, 600, 750\}$, and fix the evaluation point $t = 60$ when estimating the SPCE ($k = 1$) and RACE ($k = 2$). We consider 1,000 simulations, and calculate the absolute bias, root mean squared error (RMSE), and empirical coverage corresponding to each estimator. To obtain the empirical coverage for OW and IPW, we construct 95% confidence intervals (CIs) based on the consistent variance estimators suggested by Theorem 1. Bootstrap CIs are used for the Cox g-formula and IPW-Cox estimators. Additional simulations comparing OW with alternative regression estimators and the augmented weighting estimators (3.4) can be found in Appendix C in the Supplementary Materials.

Simulation results. Under good overlap, Figure 2 in the Supplementary Material presents the absolute bias, RMSE, and coverage for the OW, IPW estimators based on (2.3), Cox g-formula, and IPW-Cox estimators when the survival out-

comes are generated from model A and censoring is completely independent. Here, we compare treatment $j = 2$ versus $j = 3$, and thus the true average causal effect in any target population is null. Across all three estimands (SPCE, RACE, and ASCE), OW consistently outperforms IPW, with a smaller absolute bias and RMSE, and is closer to the nominal coverage across all levels of N. Owing to the correctly specified outcome model, the Cox g-formula estimator is, as expected, more efficient than the weighting estimators. However, its empirical coverage is not always close to the nominal level, especially when estimating the ASCE. The IPW-Cox estimator has the largest bias, because the proportional hazards assumption does not hold for the target population. Figure 1 represents the counterpart of Figure 2 in the Supplementary Material but under poor overlap. The IPW estimator based on (2.3) is susceptible to a lack of overlap owing to extreme inverse probability weights, resulting in an extremely large bias and variance and low coverage. The bias and under-coverage remain for IPW, even after trimming units with extreme propensities, that is, with $\max_j\{e_j(\mathbf{X}_i)\} > 0.97$ and $\min_j\{e_j(\mathbf{X}_i)\} < 0.03$. (Figure 3 in the Supplementary Material). Under poor overlap, OW is more efficient than IPW, regardless of trimming, and is almost as efficient as the Cox g-formula estimator for estimating the RACE and ASCE. Furthermore, the proposed OW interval estimator carries close to nominal coverage for all estimands. The patterns when comparing treatments $j = 2$ and $j = 1$ with the non-null true average causal effect are similar and presented in Figure 7 in the Supplementary Material.

Table 1 summarizes the performance metrics for the different estimators when the proportional hazards assumption is violated and/or censoring depends on the covariates. Similarly to Figure 1, we compare treatment $j = 2$ versus $j = 3$ such that the true average causal effect is null in any target population. When the survival outcomes are generated from model B with non-proportional hazards, both the Cox g-formula and the IPW-Cox estimators have the largest bias, especially under poor overlap. In these scenarios, OW maintains the highest efficiency, and consistently outperforms IPW in terms of bias and variance. Whereas the coverage of the IPW estimator deteriorates under poor overlap, the coverage of the OW estimator is robust to a lack of overlap. When the censoring further depends on covariates, we modify the OW and IPW estimators using (3.3), where the censoring survival functions are estimated using a Cox model. With the addition of the inverse probability of the censoring weights, only OW maintains the smallest bias, highest efficiency, and closest-to-nominal coverage under poor overlap across all estimands. The results when comparing treatments $j = 2$ and $j = 1$ are similar, and are included in Table 1 in the Supplementary Material.

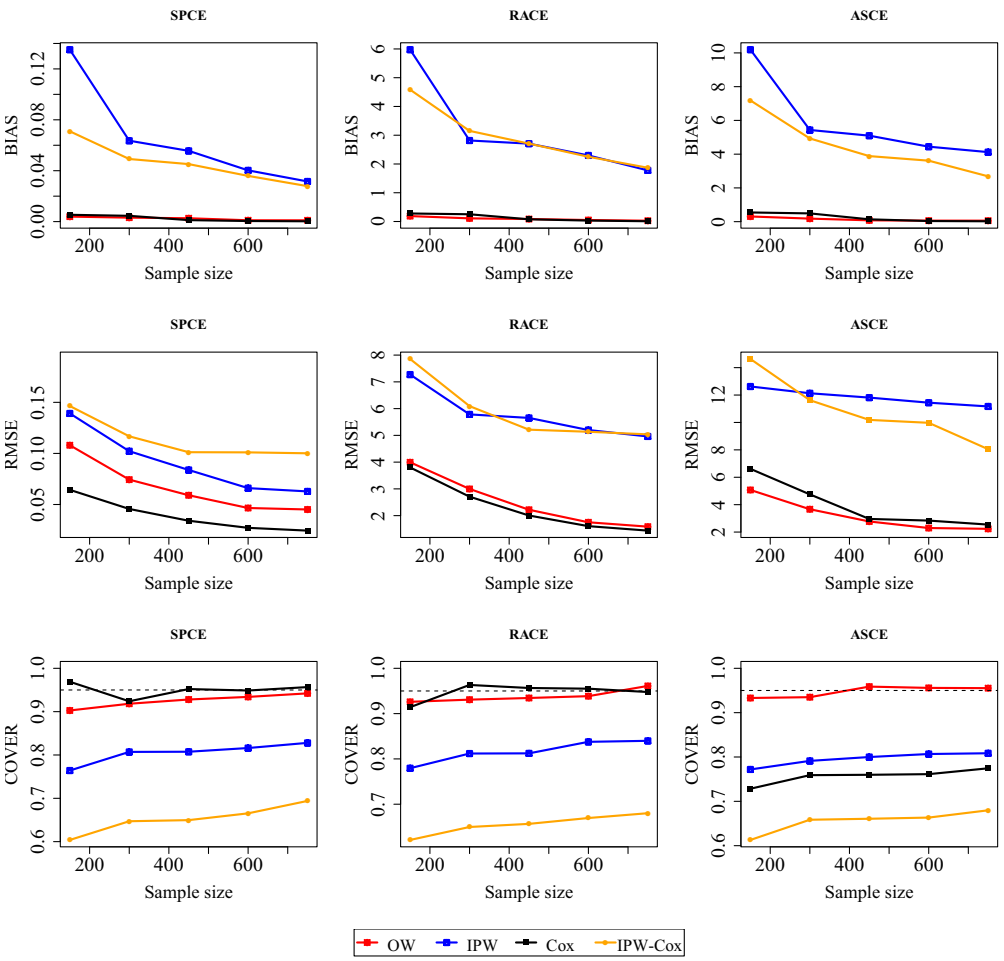


Figure 1. Absolute bias, root mean squared error (RMSE), and coverage when comparing treatment $j = 2$ versus $j = 3$ under poor overlap, and when the survival outcomes are generated from model A and censoring is completely independent.

We additionally compare OW with alternative outcome regression estimators similar to those of Mao et al. (2018) and the g-formula estimator based on the pseudo-observations. These estimators were originally developed for binary treatments, and we adapt them in Appendix C to multiple treatments. Compared with the proposed OW estimator (2.3), these regression estimators are frequently less efficient and have less than nominal coverage under poor overlap. An exception is the OW regression estimator that generalizes the work of Mao et al. (2018), which performs similarly to the OW estimator based on (2.3) when the outcome is generated from model A. When the outcome is generated from

Table 1. Absolute bias, root mean squared error (RMSE), and coverage when comparing treatment $j = 2$ versus $j = 3$ under different degrees of overlap. In the "proportional hazards" scenario, the survival outcomes are generated from a Cox model (model A), and in the "non-proportional hazards" scenario, the survival outcomes are generated from an accelerated failure time model (model B). The sample size is fixed at $N = 300$.

	Degree of	Absolute bias				RMSE				95% Coverage			
	overlap	OW	IPW	Cox	IPW-Cox	OW	IPW	Cox	IPW-Cox	OW	IPW	Cox	IPW-Cox
Model A, completely random censoring													
SPCE	Good	0.003	0.006	0.001	0.023	0.062	0.098	0.018	0.091	0.924	0.901	0.949	0.795
	Poor	0.003	0.007	0.005	0.049	0.074	0.102	0.046	0.117	0.917	0.879	0.922	0.647
RACE	Good	0.096	0.304	0.086	1.449	2.243	3.379	1.094	4.453	0.937	0.919	0.961	0.797
	Poor	0.109	0.391	0.252	3.151	2.998	3.496	2.709	6.093	0.930	0.901	0.967	0.644
ASCE	Good	0.181	0.354	0.153	2.336	2.916	4.974	1.911	8.959	0.941	0.903	0.849	0.790
	Poor	0.181	0.443	0.490	4.930	3.666	6.373	4.750	11.625	0.934	0.899	0.755	0.656
Model B, completely random censoring													
SPCE	Good	0.003	0.005	0.005	0.024	0.087	0.112	0.074	0.176	0.958	0.923	0.749	0.779
	Poor	0.005	0.008	0.016	0.081	0.097	0.118	0.150	0.222	0.941	0.921	0.770	0.712
RACE	Good	0.102	0.112	0.239	1.530	2.761	4.304	4.219	8.758	0.960	0.937	0.745	0.787
	Poor	0.105	0.299	0.947	4.646	3.627	4.669	8.653	11.275	0.936	0.929	0.742	0.709
ASCE	Good	0.129	0.443	0.468	2.382	4.238	7.174	7.354	16.583	0.958	0.959	0.846	0.777
	Poor	0.223	0.638	1.661	7.562	4.840	7.189	15.027	20.920	0.961	0.934	0.743	0.705
Model A, covariate-dependent censoring													
SPCE	Good	0.002	0.005	0.003	0.038	0.052	0.082	0.047	0.121	0.917	0.889	0.921	0.741
	Poor	0.005	0.007	0.009	0.089	0.060	0.084	0.056	0.149	0.908	0.882	0.881	0.642
RACE	Good	0.048	0.154	0.117	2.201	2.773	3.838	2.801	5.382	0.938	0.926	0.908	0.763
	Poor	0.168	0.223	0.532	4.603	3.534	4.207	3.334	7.159	0.935	0.926	0.900	0.634
ASCE	Good	0.055	0.425	0.183	1.161	5.562	8.722	6.005	36.021	0.940	0.909	0.885	0.804
	Poor	0.067	0.568	1.032	11.657	9.557	9.735	7.157	43.651	0.928	0.892	0.752	0.772
Model B, covariate-dependent censoring													
SPCE	Good	0.001	0.001	0.009	0.005	0.050	0.053	0.087	0.075	0.954	0.930	0.699	0.900
	Poor	0.002	0.005	0.012	0.025	0.052	0.082	0.164	0.082	0.925	0.925	0.723	0.896
RACE	Good	0.072	0.081	0.498	0.139	4.733	5.879	4.684	6.327	0.954	0.946	0.711	0.850
	Poor	0.109	0.146	0.712	1.594	6.250	7.115	9.092	7.515	0.956	0.955	0.705	0.839
ASCE	Good	0.072	0.258	0.794	0.340	4.436	5.738	7.337	7.756	0.954	0.946	0.835	0.847
	Poor	0.138	0.350	1.339	1.973	5.026	6.503	13.039	8.835	0.955	0.955	0.757	0.847

model B, the OW estimator in Mao et al. (2018) is subject to a larger bias and RMSE owing to an incorrect proportional hazards assumption. We carry out additional simulations in Appendix C to compare the performance of the augmented OW and IPW estimators (3.4) with that of the OW and IPW estimators (2.3). While including an outcome regression component can notably improve the efficiency of IPW, the efficiency gain for the OW estimator because of an additional outcome model is negligible. This speaks to the appeal of the simple (non-augmented) OW estimator, because outcome models are almost always misspecified in practice. Additionally, we replicate our simulations under a three-arm RCT, similarly to Zeng et al. (2020) (see Remark 3 and Appendix C). We

confirm that both the OW and the IPW estimators are valid for covariate adjustment in RCTs and lead to substantially improved efficiency over the unadjusted comparisons of pseudo-observations in the presence of chance imbalance. Finally, under covariate-dependent censoring, we further compare OW and IPW under a misspecified censoring model, finding that OW outperforms IPW in all scenarios. With a misspecified censoring model, OW also maintains nominal coverage, except when the failure times are generated from model B and the target estimand is SPCE or ASCE. The details are presented in Appendix D.

5. Application to National Cancer Database

We illustrate the proposed weighting estimators by comparing three treatment options for prostate cancer in an observational data set with 44,551 high-risk, localized prostate cancer patients drawn from the National Cancer Database (NCDB). These patients were diagnosed between 2004 and 2013, and either underwent a surgical procedure (radical prostatectomy, RP), or were treated using one of two therapeutic procedures namely external beam radiotherapy combined with androgen deprivation (EBRT+AD) or external beam radiotherapy plus brachytherapy, with or without androgen deprivation (EBRT+brachy±AD). We focus on time to death since treatment initiation as the primary outcome, and the pre-treatment covariates include age, clinical T stage, Charlson-Deyo score, biopsy Gleason score, prostate-specific antigen (PSA), year of diagnosis, insurance status, median income level, education, race, and ethnicity. A total of 2,434 patients died during the study period with their survival outcome observed, while other patients have right-censored outcomes. The median and maximum follow-up times are 21 and 115 months, respectively.

We used a multinomial logistic model to estimate the generalized propensity scores, and show the distribution of the estimated scores in Figure 9 in the Supplementary Material. The 11 pre-treatment covariates introduced earlier were considered as confounders that affect both the treatment assignment and mortality, and are included in the propensity score model. We model age and PSA by natural splines, following Ennis et al. (2018), and keep linear terms for all other covariates. We found good overlap across groups for the propensity of receiving EBRT+brachy±AD, but a slight lack of overlap for the propensity of receiving RP and EBRT+AD. To assess the adequacy of the propensity score model specification, we checked the weighted covariate balance under IPW and OW based on the maximum pairwise absolute standardized difference (MPASD) criterion, and present the balance statistics in Table 4 in the Supplementary Material. The

MPASD for the pth covariate is defined as $\max_{j<j'}\{|\bar{X}_{p,j} - \bar{X}_{p,j'}|/S_p\}$, where $\bar{X}_{p,j} = \sum_{i=1}^{N} \mathbf{1}\{Z_i = j\}X_{i,p}w_j^h(\mathbf{X}_i)/\sum_{i=1}^{N} \mathbf{1}\{Z_i = j\}w_j^h(\mathbf{X}_i)$ is the weighted covariate mean in group j, and $S_p^2 = J^{-1}\sum_{j=1}^{J} S_{p,j}^2$ is the unweighted sample variance averaged across all groups. Both IPW and OW improved the covariate balance compared with the option of no weighting. Note that, while OW with logistic propensity scores leads to an exact covariate balance for $J = 2$ groups (Li, Morgan and Zaslavsky (2018)), OW with multinomial logistic propensity scores does not guarantee an exact covariate balance among $J \geq 3$ groups (Li and Li (2019)). Nonetheless, Table 4 in the Supplementary Material shows that OW still leads to a consistently smaller MPASD compared with that of IPW, with values below the usual 0.1 threshold across all covariates.

Figure 10 in the Supplementary Material presents the estimated causal survival curves for each treatment, $\mathbb{E}\{h(\mathbf{X})\mathbf{1}\{T_i(j) \geq t\}\}/\mathbb{E}(h(\mathbf{X}))$, along with the 95% confidence bands in the combined population (corresponding to IPW) and the overlap population (corresponding to OW). We chose 220 grid points, equally spaced by half a month, for this evaluation. The estimated causal survival curves among the two target populations are similar, in general, which is expected, given that there is only a slight lack of overlap. The surgical treatment, RP, shows the largest survival benefit, followed by the radiotherapeutic treatment, EBRT+brachy±AD, while EBRT+AD results in the worst survival outcomes during the first 80 months or so. Importantly, the estimated causal survival curves for the RP and EBRT+brachy±AD treatments cross after month 80, suggesting potential violations to the proportional hazards assumption commonly assumed in survival analysis. Figures 2a and 2b further characterize the the SPCE and RACE as functions of time t, with the associated 95% confidence bands. The SPCE results confirm the largest causal survival benefit for RP, followed by EBRT+brachy±AD. The associated confidence band of the SPCE from OW is often narrower than that from IPW, and frequently excludes zero. While the analysis of the pairwise RACE yielded similar findings, the efficiency of OW over IPW became more relevant when comparing RP and EBRT+brachy±AD. Specifically, the confidence band of the RACE from OW excludes zero until month 80, while the confidence band of the RACE from IPW straddles zero across the entire follow-up period. This analysis sheds new light on the significant causal survival benefit of RP over EBRT+brachy±AD at the 0.05 level in terms of the restricted mean survival time, which has not been identified in previous analyses.

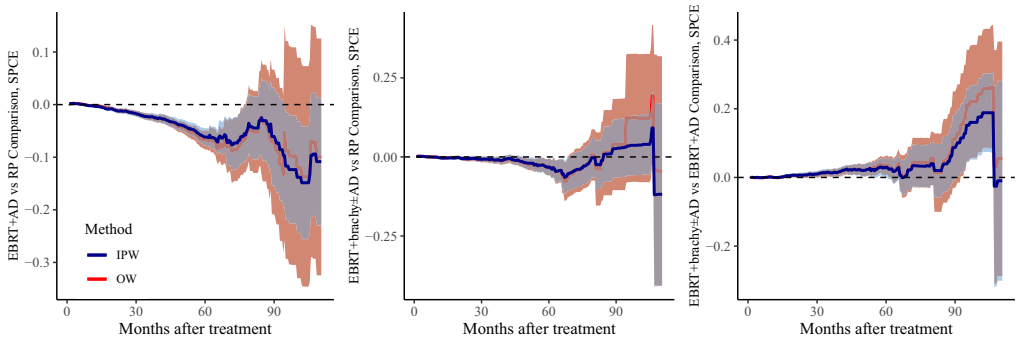
In Table 4 in the Supplementary Material, we report the SPCE and RACE using the IPW and OW estimators, as well as the Cox g-formula and IPW-Cox estimators at $t = 60$ months, that is, the 80th quantile of the follow-up time.

All methods conclude that RP leads to a significantly lower mortality rate at 60 months than does EBRT+AD. Compared with IPW, OW provides similar point estimates and no larger variance estimates. Consistent with Figure 2b, the smaller variance estimate due for OW (compared with IPW) leads to a change in conclusion when comparing EBRT+brachy±AD versus RP in terms of the RACE at the 0.05 level, and confirms the significant treatment benefit of RP. The Cox g-formula and IPW-Cox estimators sometimes provide considerably different results to those of weighting estimators based on (2.3), because they assume proportional hazards that may not hold (the estimated causal survival curves cross in Figure 10 in the Supplementary Material). Overall, we found that, compared with RP, the two radiotherapeutic treatments led to a shorter restricted mean survival time (1.2 months shorter with EBRT+AD, and 0.5 month shorter with EBRT+brachy±AD) up to five years after treatment. The five-year survival probability is also 6.7% lower under EBRT+AD and 3.1% lower under EBRT+brachy±AD compared with RP.

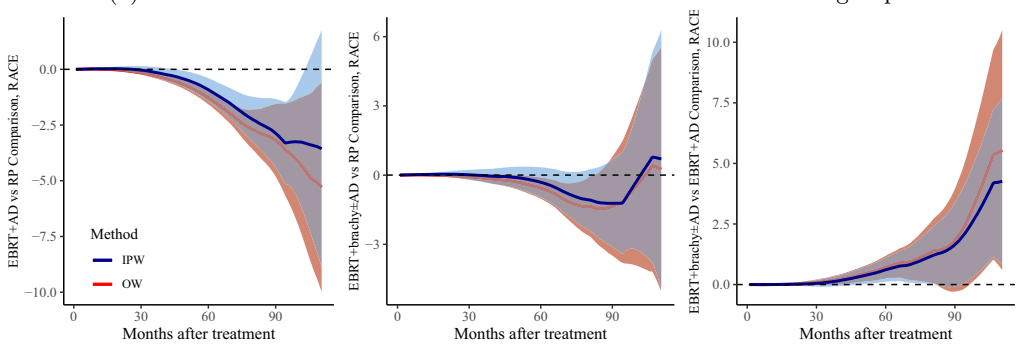
6. Discussion

We have proposed a class of propensity score weighting estimators for survival outcomes based on pseudo-observations. These estimators are applicable to several different target populations, survival causal estimands, and binary and multiple treatments. We also extended our estimators to accommodate covariate-dependent censoring and augmentation with outcome models. Previous studies rely on a bootstrap for the variance estimation of similar weighting estimators, which is computationally intensive when combined with jackknife pseudo-observations. We establish the asymptotic properties of our estimators to motivate a new closed-form variance estimator that takes into account the uncertainty due to both the calculation of the pseudo-observations and the propensity score estimation; this allows a valid and fast inference for large observational data. Within the family of balancing weights, we further established the optimal efficiency property of the overlap weights, expanding the theory of overlap weights to survival outcomes.

An important step in propensity score analysis is to specify the propensity score model. Because the goal of weighting is to balance confounders and remove bias, the weighted covariate balance is routinely used to check whether a propensity score model is adequately specified, and an iterative checking-fitting procedure is conventionally used to improve the model specification. Note that, with $J = 2$ treatments, the overlap weights obtained from the logistic propen-



(a) Estimated SPCE as a function of time t across three treatment groups.



(b) Estimated RACE as a function of time t across three treatment groups.

Figure 2. Point estimates and 95% confidence bands of SPCE and RACE as functions of time from the pseudo-observations-based IPW and OW estimators in the prostate cancer application in Section 5.

sity score model reduce the absolute standardized difference for each covariate to zero, which represents a challenge to operationalizing the iterative checking-fitting procedure (Mao, Li and Greene (2019)). As a potential remedy, one may consider alternative balance metrics, such as the weighted differences in empirical distribution function, as in McCaffrey et al. (2013). With $J \geq 3$ treatments, the overlap weights do not reduce the MPASD balance metric to zero, in general, which suggests that the iterative checking-fitting procedure based on a weighted mean balance remains feasible for improving the generalized propensity score model fit. However, because overlap weights often result in relatively satisfactory balance among treatment groups compared with IPW for almost any specification of the generalized propensity score model, a tailored rule of thumb for an adequate weighted balance would be of interest, and remains an important topic for future research.

The proposed weighting estimators can be extended in several directions. First, while we have focused on estimands on the difference scale, it is straightforward to adapt our weighting estimators to accommodate ratio estimands, which are also of interest in practice. For example, we can write $m_j^{k,h}(t) = \int_{\mathcal{X}} m_j^k(\mathbf{X};t) f(\mathbf{X}) h(\mathbf{X}) \mu(d\mathbf{X}) / \int_{\mathcal{X}} f(\mathbf{X}) h(\mathbf{X}) \mu(d\mathbf{X})$, and define the pairwise ratio estimands as $\delta_{j,j'}^{k,h}(t) = m_j^{k,h}(t)/m_{j'}^{k,h}(t)$, $\forall\ j \neq j'$. Therefore, a point identification of $\delta_{j,j'}^{k,h}(t)$ boils down to estimating the average potential outcomes $m_j^{k,h}(t)$, for each j, using pseudo-observations, and the variance calculation can proceed by applying the Delta method to Theorem 1. In addition, one may further exploit the relationship between the survival function and the hazard function to define the hazard difference as $-d\tau_{j,j'}^{k=1,h}(t)/dt$. However, inferences with these types of estimands require additional research, because our estimator for $\tau_{j,j'}^{k=1,h}(t)$ is non-smooth in t and because the causal interpretation of hazard-based estimands can be controversial. Second, under covariate-dependent censoring, our proposed estimator requires computing pseudo-observations under the inverse probability of censoring weighting (IPCW), as in Remark 4, which may be inefficient, just as when using IPW for balancing weights. When the inverse probability of censoring weights is estimated using the Cox model, improvement is possible, for example, by smoothing the baseline hazard estimator to provide a potentially more efficient estimation of $\widehat{G}(\widetilde{T}_i \wedge t|\mathbf{X}_i, Z_i)$, and hence the weights (Anderson and Senthilselvan (1980)). Alternatively, it may be interesting to develop an augmented-IPCW (hence, doubly robust) pseudo-observation estimator along the lines of a doubly robust censoring unbiased transformation (Rubin and van der Laan (2007)), which tends to be more efficient than using IPCW alone. Adapting these techniques to construct pseudo-observations is beyond the scope of this work, and requires additional research. Finally, Wallace and Moodie (2015) studied OW in constructing the optimal dynamic treatment regimen (DTR) under an additive structural mean model, and demonstrated the efficiency gain over IPW using simulations. Their approach has recently been extended to an additive structural survival model (Simoneau et al. (2020)). We conjecture that the pseudo-observation approach combined with OW can be a useful alternative to that of Simoneau et al. (2020) for identifying survival DTR under a dynamic weighted ordinary least squares framework.

Supplementary Material

The online Supplementary Material includes the Appendices A—F with technical details and additional simulations, as well as several tables and figures

referenced in Sections 4 and 5. We provide reproducible R code at `https://github.com/zengshx777/OW_Survival_CodeBase`.

Acknowledgments

The authors thank the editor, associate editor, and two anonymous referees for their constructive comments and suggestions.

References

Andersen, P. K., Hansen, M. G. and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* **10**, 335–350.

Andersen, P. K., Klein, J. P. and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**, 15–27.

Andersen, P. K. and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* **19**, 71–99.

Andersen, P. K., Syriopoulou, E. and Parner, E. T. (2017). Causal inference in survival analysis using pseudo-observations. *Statistics in Medicine* **36**, 2669–2681.

Anderson, J. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *Journal of the Royal Statistical Society: Series B (Methodological)* **42**, 322–327.

Austin, P. C. and Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* **26**, 1654–1670.

Binder, N., Gerds, T. A. and Andersen, P. K. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis* **20**, 303–315.

Cheng, C., Li, F., Thomas, L. and Li, F. (2022). Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American Journal of Epidemiology* **191**, 1140–1151.

Cole, S. R. and Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* **75**, 45–49.

Crump, R., Hotz, V. J., Imbens, G. and Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report. National Bureau of Economic Research Cambridge, MA.

Ennis, R. D., Hu, L., Ryemon, S. N., Lin, J. and Mazumdar, M. (2018). Brachytherapy-based radiotherapy and radical prostatectomy are associated with similar survival in high-risk localized prostate cancer. *Journal of Clinical Oncology* **36**, 1192–1198.

Graw, F., Gerds, T. A. and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* **15**, 241–255.

Haneuse, S. and Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in Medicine* **32**, 5260–5277.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology* **21**, 13.

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

Hubbard, A. E., Van Der Laan, M. J. and Robins, J. M. (2000). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 135–177. Springer, Berlin.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–710.

Jacobsen, M. and Martinussen, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics* **43**, 845–862.

Li, F. and Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics* **13**, 2389–2415.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113**, 390–400.

Li, F., Thomas, L. E. and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology* **188**, 250–257.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23**, 2937–2960.

Mao, H., Li, L. and Greene, T. (2019). Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research* **28**, 2439–2454.

Mao, H., Li, L., Yang, W. and Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine* **37**, 3745–3763.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* **32**, 3388–3414.

Overgaard, M., Parner, E. T. and Pedersen, J. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics* **45**, 1988–2015.

Overgaard, M., Parner, E. T. and Pedersen, J. (2019). Pseudo-observations under covariate-dependent censoring. *Journal of Statistical Planning and Inference* **202**, 112–122.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56**, 779–788.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. and van der Laan, M. J. (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics* **3**, Article 4.

Simoneau, G., Moodie, E. E., Nijjar, J. S., Platt, R. W. and the Scottish Early Rheumatoid Arthritis Inception Cohort Investigators (2020). Estimating optimal dynamic treatment regimes with survival outcomes. *Journal of the American Statistical Association* **115**, 1531–1539.

Su, C.-L., Platt, R. W. and Plante, J.-F. (2020). Causal inference for recurrent event data using pseudo-observations. *Biostatistics*. DOI:10.1093/biostatistics/kxaa020.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wallace, M. P. and Moodie, E. E. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71**, 636–644.

Wang, J. (2018). A simple, doubly robust, efficient estimator for survival functions using pseudo observations. *Pharmaceutical Statistics* **17**, 38–48.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E. and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72**, 1055–1065.

Zeng, S., Li, F., Wang, R. and Li, F. (2020). Propensity score weighting for covariate adjustment in randomized clinical trials. *Statistics in Medicine* **40**, 842–858.

Zhang, M. and Schaubel, D. E. (2012). Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics* **68**, 999–1009.

Shuxi Zeng

Department of Statistical Science, Duke University, Durham, NC 27708, USA.

E-mail: zengshx777@gmail.com

Fan Li

Department of Statistical Science, Duke University, Durham, NC 27708, USA.

E-mail: fl35@duke.edu

Liangyuan Hu

Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Piscataway, NJ 08854, USA.

E-mail: liangyuan_hu@alumni.brown.edu

Fan Li

Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA.

E-mail: fan.f.li@yale.edu

(Received May 2021; accepted December 2021)