Received: 29 June 2022 | Revised: 31 January 2024 | Accepted: 24 April 2024

DOI: 10.1002/sim.10103

RESEARCH ARTICLE

Statistics in Medicine WILEY

Propensity score matching for estimating a marginal hazard ratio

Tongrong Wang[1] | Honghe Zhao[2] | Shu Yang[2] | Shuhan Tang[1] | Zhanglin Cui[1] | Li Li[1] | Douglas E. Faries[1]

[1]Eli Lilly and Company, Indianapolis, Indiana, USA

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

Correspondence
Shu Yang, Department of Statistics, North Carolina State University, Raleigh, NC, USA.
Email: syang24@ncsu.edu

Funding information
National Science Foundation, Grant/Award Number: DMS 1811245; National Institutes of Health, Grant/Award Numbers: 1R01AG066883, 1R01ES031651

Propensity score matching is commonly used to draw causal inference from observational survival data. However, its asymptotic properties have yet to be established, and variance estimation is still open to debate. We derive the statistical properties of the propensity score matching estimator of the marginal causal hazard ratio based on matching with replacement and a fixed number of matches. We also propose a double-resampling technique for variance estimation that takes into account the uncertainty due to propensity score estimation prior to matching.

KEYWORDS

causal survival analysis, double resampling, martingale, propensity score matching, variance estimation

1 | INTRODUCTION

Survival analysis plays an increasingly important role in treatment effect estimation due to the frequent occurrence of time-to-event outcomes in biomedical studies. By comparing the hazard functions of survival times between treated and untreated individuals, the marginal hazard ratio is commonly used to measure the effect of treatment on a time-to-event outcome for a particular population of interest. The log of marginal hazard ratio corresponds to the coefficient indexing a univariate Cox proportional hazard model,[1] where the hazard of the outcome is regressed on an indicator denoting treatment status.

In observational studies, propensity score (PS) methods are standard approaches for reducing the effect of confounding. However, depending on the specific type or implementation of the propensity score methods, the population parameters or treatment effects being estimated may not be the same. In general, a measure of treatment effect can be classified as conditional or marginal. Conditional effects correspond to an average effect at the individual level, whereas marginal effects denote an average effect at the population level. Hazard ratios are non-collapsible measures, which means that the marginal and conditional hazard ratios generally do not coincide even in the absence of confounding.[2] Therefore, the non-collapsibility of the hazard ratios renders methods that estimate the conditional treatment effects biased for estimating the marginal hazard ratios.[3,4] For instance, stratification on the propensity score and covariate adjustment using the propensity score estimate conditional treatment effects, so they generally yield biased estimates of the marginal

Tongrong Wang and Honghe Zhao are joint first authors.

hazard ratio.[5,6] The inverse probability weighted (IPW) estimator fits a marginal Cox model with each observation weighted by the reciprocal of its estimated probability of receiving the observed treatment. This estimator is consistent for the marginal hazard ratio if the propensity score model is correctly specified and the overlap assumption is satisfied. However, it is sensitive to slight misspecification of the PS model and can yield large variance when the estimated PS is close to 0.[7-9] To protect against misspecification of the propensity score model, the doubly robust augmented inverse probability weighted (AIPW) estimator is proposed, which combines IPW with an outcome regression model.[10] It is consistent as long as either the outcome model or the propensity score model is correctly specified. However, non-collapsibility of the hazard ratios makes specifying a correct outcome model increasingly difficult since it needs to marginalize to a survival curve that satisfies the marginal proportional hazard assumption. Literature also suggests the performance of AIPW in finite sample scenarios can be unstable when both models of the PS and the outcome are misspecified and are also sensitive to extreme values of the estimated PS.[7,11]

Alternatively, matching methods enjoy multiple desirable features. First, matching is transparent and has great intuitive appeal as it seeks to emulate a completely randomized experiment using observational data.[12-14] Second, empirical evidence suggests that while matching on the PS and IPW do not uniformly outperform one another in all situations, PS matching tends to be more robust to misspecification of the PS model and to extreme values of the estimated PS (practical violation of the overlap assumption).[8,11,15-17] Moreover, matching does not rely on an outcome model and thus avoids the aforementioned congeniality issue between the outcome model and the proportional hazard assumption when applying AIPW. Simulation results have shown that greedy nearest neighbor matching on the propensity score without replacement results in unbiased estimation of the marginal hazard ratio over the subpopulation of treated individuals.[4] However, when the amount of treated and control units are comparable, bias could arise due to incomplete matches. PS matching with replacement not only circumvents this issue but also permits estimation of the marginal treatment effect on the overall population containing both treated and untreated subjects.[5,16,18]

While PS matching with replacement is potentially attractive for estimating the marginal hazard ratio, no formal asymptotic results have been established. For variance estimation, most existing approaches do not take into account the uncertainty of parametrically estimating the propensity score prior to matching and restrict inference to the matched sample.[4,16,19] Moreover, although Austin[20] demonstrated that bootstrap is valid for matching without replacement, it is inappropriate for matching with replacement because the bootstrap sample fails to preserve the distribution for the number of times each individual is used as a match.[21] Thus, it is important to develop a variance estimation procedure for the matching with replacement estimator of the marginal hazard ratio.

This article concerns propensity score matching with replacement and with a fixed number of matches for estimating the marginal hazard ratio for a point binary treatment given pre-treatment covariates. We will simply refer to this procedure as PSM or the PSM estimator. We note that PSM involves imputing the missing potential outcome processes,[18] and is distinct from paired (one-to-one) matching without replacement, which could result in dropping units from the analysis.

In this article, we first derive the large sample distribution of the PSM estimator based on known propensity scores. Secondly, because the propensity score function is often estimated prior to matching, we also derive the large sample distribution of PSM accounting for the uncertainty due to nuisance parameter estimation. Since PSM is a nonsmooth functional of the distribution of the data, our derivation is based on the technique developed by Andreou and Werker,[22] which offers a general recipe for deriving the limiting distribution of nonsmooth statistics that involve estimated nuisance parameters. This technique has been successfully applied by Abadie and Imbens[23] for estimating the average treatment effects for a continuous outcome. Our derivation extends their results to the survival context. This extension is not trivial because the survival outcome is often right-censored. We utilize the martingale theory of the counting process to establish asymptotic distributions of the PSM estimator of the marginal hazard ratio. In addition, we propose a replication-based variance estimator for PSM that accounts for the uncertainty of nuisance parameter estimation. For a continuous outcome, Adusumilli[24] proposed a bootstrap inference procedure, which improves upon the asymptotic variance estimator proposed by Abadie and Imbens.[23] Our proposed method extends Adusumilli's technique to the survival context. Simulation results suggest that it generally outperforms Wald-type inference based on the asymptotic theory in finite samples.

The rest of this paper proceeds as follows. Section 2 introduces the notation, model, and assumptions for identification. Section 3 presents the PSM estimator, including the matching procedure and estimating equations. In Section 4, we show the main asymptotic properties of the PSM estimator. Section 5 proposes a resampling-based inference procedure. Section 6 uses simulation to evaluate the finite-sample properties of the estimators. Section 7 applies the new estimator to the IMS Health Oncology electronic medical record data to evaluate the effects of two treatments on treating non-small cell lung cancer. Section 8 concludes with potential extensions. The supporting information contains the technical details, proofs and additional simulation results.

2 | NOTATION, MODEL, AND ASSUMPTIONS

2.1 | Potential outcomes and the causal PH model

We use the potential outcomes framework under the stable unit treatment value assumption, and let $T^{(\omega)}$ and $C^{(\omega)}$ be the potential values of the survival outcome and censoring indicator had the individual received treatment ω ($\omega = 0, 1$). We assume non-informative censoring under which $T^{(\omega)} \perp\!\!\!\perp C^{(\omega)}$, where $\perp\!\!\!\perp$ means "independent of". This assumption is reasonable if the censored times occur at the end of study follow-up τ, the so-called administrative censoring.

We define $\lambda_\omega(t) = \lim_{\delta_t \to 0} \delta_t^{-1} P\left(t \leq T^{(\omega)} < t + \delta_t | T^{(\omega)} \geq t\right)$ as the causal hazard rate of failing at time t for a population of patients had they received treatment ω. Adopting notation used by Cox[1] in the potential outcomes framework, we define $U^{(\omega)} = \min(T^{(\omega)}, C^{(\omega)})$ as the time to a clinical event or censoring, $\Delta^{(\omega)} = I(T^{(\omega)} \leq C^{(\omega)})$ the clinical event indicator, $N^{(\omega)}(t) = I(U^{(\omega)} \leq t, \Delta^{(\omega)} = 1)$ the counting process of observed event and $Y^{(\omega)}(t) = I(U^{(\omega)} \geq t)$ the at-risk process for a population of patients had they received treatment ω. Following Tchetgen Tchetgen and Robins,[10] we assume a causal PH model.

Definition (Causal PH model). The structural model for comparing treatment ω and treatment 0 is

$$\lambda_\omega(t) = \lambda_0(t) \exp(\beta_0 \omega), \tag{1}$$

where $\lambda_0(t)$ is the population hazard rate if all individuals had treatment $\omega = 0$.

The parameter β_0 describes log of the relative hazard of having a clinical event if all individuals received treatment $\omega = 1$ compared to if all individuals received treatment $\omega = 0$. Although focusing on the marginal hazard ratio is controversial in the literature,[25] it is still a useful estimand that summarizes the overall average of the hazard ratios over a certain time period. This explains its wide acceptance among clinical trial statisticians and regularity agencies such as the U.S. Food and Drug Administration.

2.2 | Observed data and identification assumptions

Let W_i be the binary treatment, and let T_i and C_i be the times to a clinical event and censoring, respectively, for individual $i = 1, \ldots, n$. We define $U_i = \min(T_i, C_i)$ as the time to a clinical event or censoring, $\Delta_i = I(T_i \leq C_i)$ the clinical event indicator, $N_i(t) = I(U_i \leq t, \Delta_i = 1)$ the observed data counting process, and $Y_i(t) = I(U_i \geq t)$ the at-risk process. Suppose we observe a set of pre-treatment baseline covariates $X_i \in \mathbb{R}^d$. Let $e(X_i) = P(W_i = 1|X_i)$ be the propensity score. The observed data are summarized as $\{O_i = (X_i, W_i, U_i, \Delta_i) : i = 1, \ldots, n\}$. We assume that $\{O_i : i = 1, \ldots, n\}$ are independent and identically distributed.

We make the consistency assumption that links the observed data processes with the potential outcome processes. In order to use the observed data to estimate the parameters in Model (1), we require the assumptions of unconfoundedness and positivity.[26]

Assumption 1 (Consistency). $T_i = T_i^{(W_i)}$ and $C_i = C_i^{(W_i)}$, or equivalently $N_i(t) = N_i^{(W_i)}(t)$ and $Y_i(t) = Y_i^{(W_i)}(t)$ for all t.

Assumption 2 (Unconfoundedness). $W_i \perp\!\!\!\perp \{T_i^{(0)}, T_i^{(1)}\} | X_i$.

Assumption 3 (Positivity). With probability 1, $0 < \underline{c} < e(X_i) < \overline{c} < 1$.

3 | THE PSM ESTIMATOR OF THE MARGINAL HAZARD RATIO

Although it is evident that paired matching (propensity score matching without replacement) is unbiased for the marginal hazard ratio over the treated population, this approach does not target estimation of β_0, the marginal hazard ratio over the population containing both treated and untreated individuals. Unlike paired matching, the PSM (with replacement) estimator permits estimation of β_0 and will be the focus of this article.[18] Previous simulation studies have provided evidence for the unbiasedness of PSM for estimating β_0.[16] In this section, we describe the PSM estimator to set the stage for our main results.

For the units in the sample, only one of the potential counting processes, $N_i^{(0)}(t)$ and $N_i^{(1)}(t)$, is observed and the other is unobserved or missing. The same is true for the potential at-risk processes. The construction of the PSM estimator first involves imputing all the missing potential outcome processes. To do so, for each unit i, it finds the first M closest units in the opposite treatment based on the Euclidean distance between the propensity scores. Note that M is fixed and does not vary between units. Define $\mathcal{J}_{M,i}$ as the set of indices for the first M matches for unit i. For all i and ω, the PSM estimator imputes the missing counting processes and at-risk processes respectively as

$$\overline{N}_i^{*(\omega)}(t) = \begin{cases} N_i(t) & \text{if } W_i = \omega, \\ M^{-1} \sum\limits_{j \in \mathcal{J}_{M,i}} N_j(t) & \text{if } W_i = 1 - \omega, \end{cases} \tag{2}$$

and

$$\overline{Y}_i^{*(\omega)}(t) = \begin{cases} Y_i(t) & \text{if } W_i = \omega, \\ M^{-1} \sum\limits_{j \in \mathcal{J}_{M,i}} Y_j(t) & \text{if } W_i = 1 - \omega. \end{cases} \tag{3}$$

Let $k_i = \sum_{l=1}^{n} I\left(i \in \mathcal{J}_{M,l}\right)$ denote the number of times unit i is used as a match. Note that each unit can be used as a match more than once so k_i can be larger than 1. In practice, the true propensity score is unknown. Following most of the empirical literature,[9,27,28] we assume that the propensity score follows a generalized linear model, $e(X_i^{\mathrm{T}}\theta_0)$. The matching procedure can be carried out with the estimated propensity score $e(X_i^{\mathrm{T}}\widehat{\theta})$, where $\widehat{\theta}$ is a consistent estimator of θ_0. We now denote k_i to be $k_{\widehat{\theta},i}$ to reflect its dependence on $\widehat{\theta}$.

Once the missing potential outcome processes are imputed, PSM then fits a marginal structural model to the imputed dataset. Define $\Lambda_0(t) = \int_0^t \lambda_0(v)\mathrm{d}v$ as the cumulative hazard function for $\omega = 0$ at time t. The estimating functions for $\Lambda_0(t)$, $t \geq 0$ and β_0 are

$$\sum_{i=1}^{n} \sum_{\omega=0}^{1} \left\{ \mathrm{d}\overline{N}_i^{*(\omega)}(t) - \mathrm{d}\Lambda_0(t) \exp(\beta\omega) \overline{Y}_i^{*(\omega)}(t) \right\}, \tag{4}$$

$$\sum_{i=1}^{n} \sum_{\omega=0}^{1} \int_0^{\tau} \omega \left\{ \mathrm{d}\overline{N}_i^{*(\omega)}(t) - \mathrm{d}\Lambda_0(t) \exp(\beta\omega) \overline{Y}_i^{*(\omega)}(t) \right\}, \tag{5}$$

where $\overline{N}_i^{*(\omega)}(t)$ and $\overline{Y}_i^{*(\omega)}(t)$ are defined in (2) and (3).

We can equivalently write (4) and (5) using observed outcome processes respectively as

$$\sum_{i=1}^{n} \left\{ 1 + \frac{k_{\widehat{\theta},i}}{M} \right\} \{ \mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t) \exp(\beta W_i) Y_i(t) \}, \tag{6}$$

$$\sum_{i=1}^{n} \left\{ 1 + \frac{k_{\widehat{\theta},i}}{M} \right\} \int_0^{\tau} W_i \{ \mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t) \exp(\beta) Y_i(t) \}. \tag{7}$$

Setting (6) equal to zero, we can obtain the estimator for $\mathrm{d}\Lambda_0(t)$ for fixed β as

$$\mathrm{d}\widehat{\Lambda}_0(t) = \frac{\sum_{i=1}^{n} \{1 + k_{\widehat{\theta},i}/M\} \mathrm{d}N_i(t)}{\sum_{i=1}^{n} \{1 + k_{\widehat{\theta},i}(W_i)/M\} \exp(\beta W_i) Y_i(t)}. \tag{8}$$

Substituting (8) into (7), we obtain the estimating equation to solve for β,

$$G_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ 1 + \frac{k_{\hat{\theta},i}}{M} \right\} \left\{ W_i - \hat{Q}(\beta, t) \right\} \mathrm{d}N_i(t) = 0, \tag{9}$$

where

$$\hat{Q}(\beta, t) = \frac{\sum_{j=1}^{n} \{1 + k_{\hat{\theta},i}/M\} W_j \exp(\beta W_j) Y_j(t)}{\sum_{j=1}^{n} \{1 + k_{\hat{\theta},i}/M\} \exp(\beta W_j) Y_j(t)}. \tag{10}$$

Equation (9) is the partial score equation for a Cox PH model with W_i as the covariate and weights $1 + k_{\hat{\theta},i}/M$. Thus, the PSM estimator $\hat{\beta}$ can be calculated by standard software.

We conclude this section by summarizing the steps for calculating the PSM estimator for β_0 as follows.

Step 1. Fit a propensity score model $e(X_i^{\mathrm{T}}\theta)$, and obtain an estimate $\hat{\theta}$.

Step 2. Based on the estimated propensity scores $e(X_i^{\mathrm{T}}\hat{\theta})$, carry out the matching procedure described above. Record the number of times each unit is used as a match $k_{\hat{\theta},i}$.

Step 3. Obtain the PSM estimator $\hat{\beta}$ for β_0 by solving (9) using standard software, that is, by fitting a Cox PH model to the observed data with covariate W_i and weights $1 + k_{\hat{\theta},i}/M$.

4 | MAIN RESULTS

4.1 | Asymptotic results with known propensity scores

In this section, we establish the asymptotic normality of the PSM estimator of the marginal HR assuming the propensity scores are known. The results in this subsection are applicable to rare situations when the propensity scores are known, and are useful for understanding the effect of estimating the propensity scores on the PSM estimator when compared to the results in the next subsection.

Under regularity conditions in Assumption S1, we show that there exists a neighborhood $\mathcal{B}$ of β_0 and a function $Q(\beta_0, t)$ such that for all $(\beta, t) \in \mathcal{B} \times [0, \tau]$, $\hat{Q}(\beta, t) \rightarrow_p Q(\beta, t)$, as $n \rightarrow \infty$. We then define

$$H_i(\omega) = \int_0^{\tau} \{\omega - Q(\beta_0, t)\} \left\{ \mathrm{d}N_i^{(\omega)}(t) - \mathrm{d}\Lambda_0(t) \exp(\beta_0 \omega) Y_i^{(\omega)}(t) \right\}, \tag{11}$$

$\mu_H(\omega, X) = E\{H(\omega)|X\}$, and $\sigma_H^2(\omega, X) = \mathrm{var}\{H(\omega)|X\}$.

Because $\hat{\beta}$ is the solution to the estimating equation $G_n(\beta) = 0$ in (9), the key step is to characterize the asymptotic properties of $G_n(\beta_0)$. With a known θ_0, we can show that

$$n^{-1/2} G_n(\beta_0) = n^{-1/2} \sum_{i=1}^{n} \left\{ 1 + \frac{k_{\theta_0,i}}{M} \right\} H_i(W_i) + o_p(1). \tag{12}$$

Based on (12) and the M estimation theory, we derive the asymptotic results for $\hat{\beta}$ as follows.

Theorem 1. *Under Assumptions 1–3 and the regularity conditions in Assumption S1 presented in the supplementary material, with the known propensity score,*

$$n^{1/2}(\hat{\beta} - \beta_0) \rightarrow \mathcal{N}(0, V_1),$$

as $n \rightarrow \infty$, where $V_1 = \{A(\beta_0)\}^{-1} V_G \{A(\beta_0)\}^{-1}$,

$$A(\beta_0) = E\left(\int_0^{\tau} \left[\frac{E\left\{\exp(\beta_0) Y^{(1)}(t)\right\}}{E\left\{Y^{(0)}(t)\right\} + E\left\{\exp(\beta_0) Y^{(1)}(t)\right\}} - 1 \right] \times \frac{E\left\{\exp(\beta_0) Y^{(1)}(t)\right\}}{E\left\{Y^{(0)}(t)\right\} + E\left\{\exp(\beta_0) Y^{(1)}(t)\right\}} \sum_{\omega=0}^{1} \mathrm{d}N^{(\omega)}(t) \right). \tag{13}$$

and,

$$V_G = \sum_{\omega=0}^{1} E\left[\sigma_H^2\{\omega, e(X)\}\left\{\frac{2M+1}{2Mp(\omega|X)} - \frac{p(\omega|X)}{2M}\right\}\right] + E\left([\mu_H\{0, e(X)\} + \mu_H\{1, e(X)\}]^2\right), \tag{14}$$

$p(\omega|X) = pr(W = \omega|X)$.

Recall that the estimating function of the PSM estimator is $n^{-1/2}G_n(\beta_0)$ as in (12), which targets estimating $E\{\sum_{\omega=0}^{1} H_i(\omega)\}$. Theorem 1 shows that this estimating function has the asymptotic variance V_G as in (14). From the standard semiparametric estimation literature,[29] the semiparametric efficiency bound for the target parameter $E\{\sum_{\omega=0}^{1} H_i(\omega)\}$ is $\sum_{\omega=0}^{1} E\{\sigma_H^2(\omega, X)/p(\omega|X)\} + E[\{\mu_H(0, X) + \mu_H(1, X)\}^2]$. Thus, the PSM estimator does not attain the semiparametric efficiency bound. The asymptotic variance V_G in (14) becomes closer to the efficiency bound as the number of matches M gets larger and $e(X)$ can explain all the variation of $H_i(\omega)$ given X, that is, $\mu_H(\omega, X) = \mu_H\{\omega, e(X)\}$.

4.2 | Asymptotic results with estimated propensity scores

We now study the asymptotic properties of the PSM estimator with the estimated propensity score. The technique we will use is based on Andreou and Werker.[22] The main idea is to apply Le Cam's third lemma[30] to locally asymptotically normal models. Let P^{θ} be the distribution of $\{(A_i, X_i, U_i, \Delta_i) : i = 1, \dots, n\}$ indexed by the parameter θ from the propensity score model. Let θ_n be contiguous to θ_0, and P^{θ_n} be the distribution indexed by the local parameter θ_n. Under P^{θ_n}, denote the true parameter value as $\beta_0(\theta_n)$, the PSM estimator based on the true parameter θ_n as $\widehat{\beta}(\theta_n)$, and the log likelihood of P^{θ_0} with respect to P^{θ_n} as $\Lambda_n(\theta_0|\theta_n)$. Assume that *under* P^{θ_n},

$$\left(n^{1/2}\{\widehat{\beta}(\theta_n) - \beta_0(\theta_n)\}, n^{1/2}(\widehat{\theta} - \theta_n), \Lambda_n(\theta_0|\theta_n)\right)^{\mathrm{T}} \tag{15}$$

has a limiting normal distribution. Le Cam's third lemma states that *under* P^{θ_0}, $n^{1/2}\{\widehat{\beta}(\theta_n) - \beta_0(\theta_n)\}$ has a limiting normal distribution. Then heuristically by replacing θ_n with $\widehat{\theta}$, one can then approximate the asymptotic distribution of $n^{1/2}(\widehat{\beta} - \beta_0)$ as shown in Theorem 2.

Theorem 2. *Under Assumptions 1–3 and the regularity conditions in Assumption S1 presented in the supplementary material, with a correctly specified propensity score model,*

$$n^{1/2}(\widehat{\beta} - \beta_0) \to \mathcal{N}(0, V_2),$$

as $n \to \infty$, $V_2 = \{A(\beta_0)\}^{-1}\widetilde{V}_G\{A(\beta_0)\}^{-1}$, $\widetilde{V}_G = V_G - c^{\mathrm{T}}\mathcal{I}_{\theta_0}^{-1}c$, *where* V_G *is defined in Theorem1,* $\mathcal{I}_{\theta_0}$ *is the Fisher information of* θ_0, $\dot{e}(X^{\mathrm{T}}\theta_0) = \partial e(X^{\mathrm{T}}\theta_0)/\partial\theta$, *and*

$$c = E\left\{\left[\frac{\mathrm{cov}\{X, \mu_H(1, X)|e(X)\}}{e(X)} + \frac{\mathrm{cov}\{X, \mu_H(0, X)|e(X)\}}{1 - e(X)}\right]\dot{e}(X^{\mathrm{T}}\theta_0)\right\}. \tag{16}$$

By comparing V_2 and V_1, the change in asymptotic variance after adjusting for estimating propensity score function is $-\{A(\beta_0)\}^{-1}c^{\mathrm{T}}\mathcal{I}_{\theta_0}^{-1}c\{A(\beta_0)\}^{-1}$, which can either be negative or zero. This reduction is a result of utilizing the available treatment assignment information in the data, which improves the efficiency of the propensity score matching estimator. Therefore, Theorem 2 shows that matching based on the estimated propensity score generally improves the efficiency of the matching estimator compared to matching based on the true propensity score even if it is known. This phenomenon is in line with that in the setting with a continuous outcome.[23] Theorem 2 motivates a variance estimator based on an approximation of the asymptotic variance formula (see supplementary information).

5 | RESAMPLING-BASED INFERENCE

We propose resampling variance estimation that has a better finite-sample performance. Abadie and Imbens[21] demonstrated that the nonparametric bootstrap is invalid for the matching estimator based on matching with replacement and with a fixed number of matches. Otsu and Rai[31] suggested resampling the matching estimator directly based on its martingale residual terms, which only works for matching based on the full vector of covariates. In order to reflect the uncertainty in the estimation of the propensity score, Adusumilli[24] proposed a hybrid bootstrap procedure by re-assigning new values for the treatments and resampling the martingale residuals under both treatment conditions. This necessitates imputation of the unobserved martingale residuals under the opposite treatment. Extending this procedure to survival outcomes, we define the martingale residuals as

$$\widehat{r}_{1i}(\theta) = \widehat{\mu}_H\left\{0, e(X^{\mathrm{T}}\theta)\right\} + \widehat{\mu}_H\left\{1, e(X^{\mathrm{T}}\theta)\right\} - 0,$$
$$\widehat{r}_{2i}(\omega, \theta) = H_i(\omega) - \widehat{\mu}_H\left\{\omega, e(X^{\mathrm{T}}\theta)\right\},$$

where $\widehat{\mu}_H(\omega, e(X_i^{\mathrm{T}}\theta))$ is obtained by a nonparametric regression estimation of H_i on $e(X_i^{\mathrm{T}}\theta)$ among individuals with $W_i = \omega$. For individual i, define the secondary nearest neighbor matching function as $m(\omega, X_i)$; if $W_i = \omega$, $m(\omega, X_i) = i$, otherwise, $m(\omega, X_i)$ returns the index of the nearest neighbor in the opposite treatment group, where nearest neighbor is determined based on the full X rather than on the propensity score. The pair of potential residuals for individual i are defined as

$$\widehat{r}^*_{2i}(0, \theta) = \begin{cases} \widehat{r}_{2i}(0, \theta) & \text{if } W_i = 0, \\ \widehat{r}_{2m(0,X_i)}(0, \theta) & \text{if } W_i = 1, \end{cases} \quad \widehat{r}^*_{2i}(1, \theta) = \begin{cases} \widehat{r}_{2m(0,X_i)}(1, \theta) & \text{if } W_i = 0, \\ \widehat{r}_{2i}(1, \theta) & \text{if } W_i = 1. \end{cases} \tag{17}$$

We propose a double-resampling procedure as follows.

Step 0. Complete Steps 1–3 for obtaining the PSM estimator. For each individual i, compute the secondary nearest neighbor matching function $m(\omega, X_i)$ for $\omega = 0, 1$. Estimate the matching function $k_{\theta_0,i}(\omega)$ for $\omega = 0, 1$. For $\omega = W_i$, let $\widehat{k}_i(\omega) = k_{\widehat{\theta},i}$; for $\omega = 1 - W_i$, we use the following imputation strategy: create q_N quantile partitions based on $e(X_i^{\mathrm{T}}\widehat{\theta})$, identify the quantile partition individual i falls, randomly sample one, say l, from that partition with $W_j = \omega$, and let $\widehat{k}_i(\omega) = k_{\widehat{\theta},l}$.

Step 1. Generate the bootstrap treatment, $W_i^*, i = 1, \dots, n$, from 0 with probability $1 - e(X_i^{\mathrm{T}}\widehat{\theta})$, and 1 with probability $e(X_i^{\mathrm{T}}\widehat{\theta})$.

Step 2. Based on $(W_i^*, X_i)_{i=1}^n$, re-fit the propensity score model and obtain the replicate $\widehat{\theta}^*$.

Step 3. Generate a sequence of independent and identically distributed random variables $(u_1^*, \dots, u_n^*)$ from $u_i^* \sim \mathcal{N}(0, 1)$.

Step 4. Define the new bootstrap residuals as $\widehat{r}^*_{H,i}(\theta) = \widehat{r}_{1i}(\theta) + \{1 + k_{\widehat{\theta},i}(W_i^*)/M\}\widehat{r}_{2i}(W_i^*, \theta)$, whose expectation over the probability distribution implied by Step 1, conditional on the original data, is $\widehat{R}^*_H(\theta) = n^{-1}\sum_{j=1}^n[\widehat{r}_{1j}(\theta) + e(X^{\mathrm{T}}\theta)\{1 + k_{\widehat{\theta},i}(1)/M\}\widehat{r}_{2j}(1, \theta) + \{1 - e(X^{\mathrm{T}}\theta)\}\{1 + k_{\widehat{\theta},i}(0)/M\}\widehat{r}_{2j}(0, \theta)]$, where $k_{\widehat{\theta},i}(\omega)$ is imputed at Step 0. Re-center the bootstrap residuals and compute the replicate of $G_n(\beta_0)$ as $G_n^* = \sum_{i=1}^n\left\{\widehat{r}^*_{H,i}(\widehat{\theta}^*) - \widehat{R}^*_H(\widehat{\theta}^*)\right\}u_i^*$.

Step 5. Repeat Steps 1–4 for a large number B times and denote the bth replicate of $G_n(\beta_0)$ as $G_n^{*(b)}$. Construct the $100(1 - \alpha)\%$ percentile bootstrap confidence interval for $G_n(\beta_0)$ as $\left(G^*_{n(\alpha/2)}, G^*_{n(1-\alpha/2)}\right)$, where G^*_{nq} is the qth percentile of $\{G_n^{*(1)}, \dots, G_n^{*(B)}\}$.

We construct the $100(1 - \alpha)\%$ confidence interval for β_0 as $\left\{\mathrm{d}G_n(\widehat{\beta})/\mathrm{d}\beta\right\}^{-1}(G^*_{n(\alpha/2)}, G^*_{n(1-\alpha/2)})$, which has the nominal coverage level asymptotically. The double-resampling procedure is parallel to the weighted bootstrap procedure of Adusumilli.[24] Thus, the validity of the double-resampling procedure is a straightforward adaptation of the proofs in the work of Adusumilli[24] to our setting under regularity conditions given in Assumption S2. The interested reader is encouraged to consult the original article by Adusumilli[24] for details.

Remark 1. We provide some discussions to Step 0. Note that the secondary matching procedure matches on the full set of covariates rather than on the propensity scores. Doing so preserves the conditional covariance in (16) between X and the error terms $\hat{r}_{2i}$, given the propensity scores. This term reflects the improvement when using the estimated propensity score. See Adusumilli.[24] We impute $k_i(\omega)$ by drawing a value from the empirical distribution of $k_i(W_i)$ in the propensity score strata for $W_i = \omega$ to re-create the distribution of $k_i(W_i)$. The number of the propensity score strata q_N is required to go to infinity as the sample size increases. In finite samples, we suggest using quintiles ($q_N = 5$) and recommend conducting sensitive analysis with different choices of q_N. Our simulation study shows that the performance of the proposed double-sampling procedure is not sensitive to this choice.

6 | SIMULATION STUDY

6.1 | Overview

We conduct a simulation study to compare the performance of PSM estimator to existing approaches for estimating the log-marginal hazard ratio and to assess the performance of the proposed resampling-based variance estimator. Motivated by previous findings, we consider data generating mechanisms that simulate a varying level of covariate overlap under the marginal proportional hazard assumption. This simulation design attempts to highlight the differences between the PSM estimator and other existing approaches, by investigating the following hypotheses:

(1) When overlap is poor or when the propensity score model is misspecified, the PSM estimator is anticipated to outperform the IPW and AIPW estimators because PSM is more robust to extreme values of the propensity scores.[7,8,11] See supplementary information for a discussion comparing PSM and AIPW from a theoretical and practical point of view.
(2) Matching on high-dimensional covariates will result in noticeable bias because the search for close matches becomes increasingly difficult in higher dimensions.[18,32-34]
(3) The proposed resampling-based variance estimator is expected to outperform other existing approaches.[24]

Bias, empirical variance, the average of variance estimates, and the coverage rate for nominal 95% confidence intervals are obtained. The performance of the point estimators are evaluated using bias and empirical variance. Any deviation of the averaged variance estimates from the empirical variance reflects bias in variance estimation. The coverage rate for nominal 95% confidence intervals of the estimated log-marginal hazard ratios is used to measure the overall performance of the estimation procedures.

6.2 | Data generating mechanism

We consider a sample size of $n = 1000$ throughout. For individuals $i = 1, \dots, n$, we simulate the following data:

Step 1: vector of baseline covariates $X = (X_1, \dots, X_6)^{\mathrm{T}}$, each independently generated from an exponential distribution with mean 1.

Step 2: indicator of assignment to treatment W generated from the propensity score model: logit $P(W = 1|X) = \theta_0 + X^{\mathrm{T}}\theta_1$, where we let $(\theta_0, \theta_1^{\mathrm{T}})$ be $(-3.0, 0.5 \times \mathbf{1}^{\mathrm{T}}_{6\times1})$, $(-4.5, 0.75 \times \mathbf{1}^{\mathrm{T}}_{6\times1})$, and $(-5.0, \mathbf{1}^{\mathrm{T}}_{6\times1})$ to represent weak, moderate, and strong levels of confounding, leading to strong, moderate, and weak covariate overlap. The distribution of the true propensity scores is shown in Figure 1. We also examine a scenario where there is no confounding, that is, each subject has a true propensity score of 0.5 (see supplementary information).

Step 3: counterfactual outcome under control $T^{(0)}$ generated from an exponential distribution with survival function $S^{(0)}(t) = \exp(-\lambda_0 t)$; counterfactual outcome under treatment $T^{(1)}$ generated by first drawing u from a uniform distribution with support $[0, 1]$ and then solving

$$\left\{\prod_{k=1}^{6}(1 - \eta_k t)\right\} \exp\left[\left\{X^{\mathrm{T}}\eta - \lambda_0 \exp(\beta_0)\right\} t\right] - 1 + u = 0$$

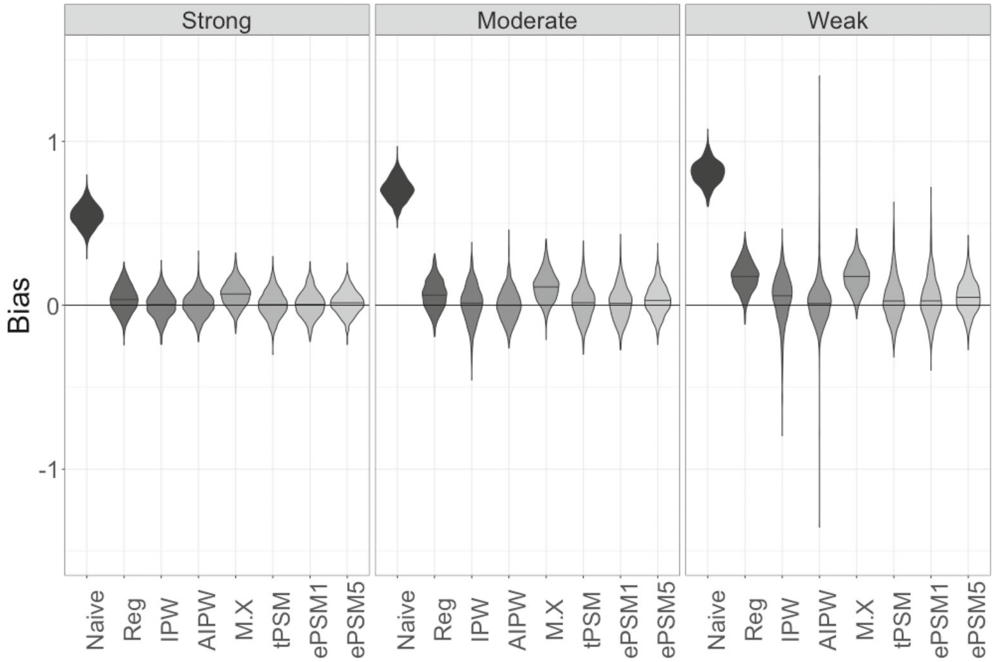

FIGURE 1 Simulation results of various point estimators of β_0 from 1000 Monte Carlo simulated datasets. `Naive` is the $\widehat{\beta}_{\text{nai}}$ estimator; `IPW` is the $\widehat{\beta}_{\text{ipw}}$ estimator; `AIPW` is the $\widehat{\beta}_{\text{aipw}}$ estimator; `M.X` is the $\widehat{\beta}_{\text{m.x}}$ estimator; `tPSM` is the $\widehat{\beta}_{\text{psm},0}$ estimator; `ePSM1` is the $\widehat{\beta}_{\text{psm}}$ estimator with $M = 1$; `ePSM5` is the $\widehat{\beta}_{\text{psm}}$ estimator with $M = 5$; some AIPW point estimations with absolute value greater than 1.5 are excluded from the plot in the scenario where the level of covariate overlap is strong.

for t, where $\eta_1 = \cdots = \eta_6 = -2$ and $\eta = (\eta_1, \ldots, \eta_6)^{\text{T}}$. We consider three choices for the estimand β_0 and let $\lambda_0 = 6$ for $\beta_0 = 0$ and $\beta_0 = 0.5$, and $\lambda_0 = 15$ for $\beta_0 = -0.5$. This method for simulating the counterfactual outcomes has been used to guarantee the PH assumption.[35] See supplementary information for justification.

Step 4: actual true event time is calculated as $T = (1 - W)T^{(0)} + WT^{(1)}$.

Step 5: event time $\Delta = I(T < C)$ and event indicator $U = \min(C, T)$, where C is generated from a uniform distribution with support $[0, a]$ and a is set so that between 20% and 30% individuals are censored.

In sum, we allow the following two factors to vary in our simulation designs: covariate overlap (weak, moderate, strong, and very strong) and the true log of marginal hazard ratio $(0, -0.5, 0.5)$. This gives rise to a total of 12 scenarios. For each scenario, we simulate 1000 datasets.

6.3 | Methods

We compare the following estimators of the log-marginal hazard ratio:

(i) $\widehat{\beta}_{\text{nai}}$, the unadjusted estimator obtained by fitting the Cox PH model with the treatment status as the only covariate, which serves as a benchmark for demonstrating the level of confounding bias;

(ii) $\widehat{\beta}_{\text{ipw}}$, the IPW estimator obtained by fitting a weighted Cox PH model with the treatment status as the only covariate, where each observation is weighted by the inverse of the probability of receiving the actual treatment;

(iii) $\widehat{\beta}_{\text{aipw}}$, the AIPW estimator,[10] where the working outcome model is the Cox PH model with all baseline covariates and treatment status as covariates;

(iv) $\widehat{\beta}_{\text{m.x}}$, the estimator based on matching on all covariates, which works the same way as the PSM estimator, except during the matching stage, nearest neighbors are determined based on the Euclidean distance between vectors of covariates rather than between propensity scores; the number of matches is set to $M = 1$;

(v) $\widehat{\beta}_{\text{psm},0}$, the PSM estimator based on the true propensity score; the number of matches is set to $M = 1$;

(vi) $\widehat{\beta}_{\rm psm}$, the PSM estimator based on the estimated propensity score; two versions of the PSM estimator corresponding to two different choices for the number of matches ($M = 1$ and $M = 5$) are considered.

For variance estimation of $\widehat{\beta}_{\rm nai}$, $\widehat{\beta}_{\rm ipw}$, and $\widehat{\beta}_{\rm m.x}$, we use the robust output from the standard software. The nonparametric bootstrap is used for estimating the variance of $\widehat{\beta}_{\rm aipw}$. For $\widehat{\beta}_{\rm psm,0}$, we use the proposed asymptotic variance estimator without the adjustment term. We compare four variance estimators for $\widehat{\beta}_{\rm psm}$: (i) the robust output from the standard software, (ii) the proposed asymptotic variance estimator (iii) the nonparametric bootstrap and (iv) the proposed double-resampling procedure with $q_N = 5$.

6.4 | Sensitivity analysis

For each simulation scenario above, we examine the effect of misspecifying the propensity score model on the three propensity score based approaches ($\widehat{\beta}_{\rm ipw}$, $\widehat{\beta}_{\rm aipw}$, and $\widehat{\beta}_{\rm psm}$). That is, we compare them under the correct PS model specification $\text{logit}P(W = 1|X) = \theta_0 + \theta_1^{\rm T} X$ and under a misspecified PS model $\text{logit}P(W = 1|X) = \theta_0 + \theta_1^{\rm T} X^{1/2}$. We also investigate the sensitivity of the double-resampling approach to a different choice for the number of strata ($q_N = 10$).

6.5 | Simulation study results

Figure 1 and Table 1 show the results when true $\beta_0 = 0$ and the propensity score model is correctly specified. The results for the misspecified propensity score model and results for $\beta_0 = -0.5$ and $\beta_0 = 0.5$ are presented in the supporting information. The unadjusted estimator $\widehat{\beta}_{\rm nai}$ has a severe bias and thus barely captures β_0, which becomes worse as the covariate overlap between the treatment groups becomes weaker. The matching estimator based on all covariates $\widehat{\beta}_{\rm x.m}$ has the second-largest bias following the naive estimator, $\widehat{\beta}_{\rm nai}$, confirming the theoretical result in Abadie and Imbens[18] that matching on more than one covariate may lead to a biased matching estimator. The IPW estimator $\widehat{\beta}_{\rm ipw}$ and the AIPW estimator $\widehat{\beta}_{\rm aipw}$ greatly reduce the bias; however, as shown in Figure 1, they are unstable when the two treatment groups have weak overlap in propensity scores. Moreover, $\widehat{\beta}_{\rm aipw}$ proposed by Tchetgen Tchetgen and Robins[10] is supposed to be doubly robust in the sense that its consistency relies only on the correct specification of either the propensity score model or a working outcome mean model given the covariates. However, in practice, specifying a correct outcome mean model that is compatible with the marginal structural model is difficult. The posited Cox regression model in the AIPW estimator is a convenient choice but is misspecified. Thus, the AIPW estimator is no longer doubly robust. For all investigated simulation scenarios, $\widehat{\beta}_{\rm psm}$ can significantly reduce the bias and are more stable when the two treatment groups have weak overlap in propensity scores compared to $\widehat{\beta}_{\rm ipw}$ and $\widehat{\beta}_{\rm aipw}$. The PSM estimator with $M = 5$ provides a smaller variance than the counterpart with $M = 1$. This is because increasing the number of matching can improve the efficiency of the PSM estimator. As a trade-off, when the sample size is small, a large number of matching could lead to a slight increase in bias. Moreover, the PSM estimators possess coverage rates around 95% using the proposed asymptotic variance estimation and double-resampling. When the two treatment groups have weak overlap, it is noticed that the asymptotic variance estimator tends to underestimate the variance and may result in negative values. In contrast, the double-resampling method recovers the 95% confidence level better than the asymptotic method for such cases. Comparing to the proposed inference approach, using the standard software's robust method and naive bootstrap method always overestimate the variances of $\widehat{\beta}_{\rm psm,0}$ and $\widehat{\beta}_{\rm psm}$. Thus, our proposed variance estimation approach is apparently beneficial for making a reliable inference.

In the extended simulations (see the supporting information), we conduct a sensitivity analysis on a different choice of the number of strata, q_N, for the double-resampling approach. The results show that the estimated variances are similar with a difference less than 10^{-3} across all scenarios, indicating that the variance estimator is insensitive to the choice of q_N. Moreover, when the propensity score model is misspecified, $\widehat{\beta}_{\rm ipw}$, $\widehat{\beta}_{\rm aipw}$ and $\widehat{\beta}_{\rm psm}$ become biased. Nonetheless, $\widehat{\beta}_{\rm psm}$ is still more robust than $\widehat{\beta}_{\rm ipw}$ and $\widehat{\beta}_{\rm aipw}$. The AIPW estimator performs the worst among all estimators when the level of covariate overlap is medium and weak, consistent with the results in Kang and Schafer[7] that the bias and variance of AIPW estimator can increase dramatically when both the propensity score and outcome models are misspecified. For the scenario where there is no confounding, while $\widehat{\beta}_{\rm psm}$ with $M = 5$ results in similar bias than $\widehat{\beta}_{\rm ipw}$ and $\widehat{\beta}_{\rm aipw}$, the PSM estimators are generally less efficient than

TABLE 1 Simulation results: Bias ($\times 10^2$) and variance ($\times 10^3$) of the point estimator of β_0, coverage (%) of 95% confidence intervals based on 1,000 Monte Carlo samples with true $\beta_0 = 0$.

		Bias	Var	VE	CR	Bias	Var	VE	CR	Bias	Var	VE	CR
Level of covariate overlap		**Strong**				**Medium**				**Weak**			
$\widehat{\beta}_{\text{nai}}$		54.6	4.9	5.1	0.0%	70.6	5.2	5.3	0.0%	80.9	5.2	5.7	0.0%
$\widehat{\beta}_{\text{ipw}}$		0.3	6.1	8.3	97.8%	1.0	12.5	14.6	95.6%	4.1	22.2	18.9	91.2%
$\widehat{\beta}_{\text{aipw}}$		0.1	5.3	6.0	95.9%	0.2	9.1	12.0	96.1%	2.3	31.3	40.6	96.5%
$\widehat{\beta}_{\text{x.m}}$		6.9	5.9	8.8	92.7%	11.4	8.2	11.8	84.0%	17.7	8.2	11.7	64.4%
$\widehat{\beta}_{\text{psm},0}$		0.5	6.6	10.9	98.5%	1.6	10.6	17.9	97.4%	3.7	16.1	25.4	96.1%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	0.7	6.3	11.1	98.9%	1.3	10.5	18.3	97.9%	3.9	17.2	26.0	95.8%
	asymp	0.7	6.3	6.5	95.1%	1.3	10.5	10.6	94.2%	3.9	17.2	16.4	91.3%
	naiveboot	0.7	6.3	9.8	98.8%	1.3	10.5	17.0	98.7%	3.9	17.2	25.1	96.3%
	double-rsp	0.7	6.3	8.2	97.7%	1.3	10.5	13.6	97.5%	3.9	17.2	24.1	95.4%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	1.5	5.0	8.1	99.0%	3.2	8.0	12.3	96.6%	5.1	10.1	16.3	97.0%
	asymp	1.5	5.0	5.0	94.2%	3.2	8.0	7.3	91.9%	5.1	10.1	9.2	90.6%
	naiveboot	1.5	5.0	7.6	99.0%	3.2	8.0	12.2	96.8%	5.1	10.1	16.6	97.5%
	double-rsp	1.5	5.0	6.1	96.1%	3.2	8.0	9.5	95.2%	5.1	10.1	13.9	95.8%

Note: "Var" is the variance of point estimates of β_0 across 1000 simulated datasets; "VE" is the average variance estimation for the point estimators over simulations, thus VE minus Var reflects the bias in estimated variance; "CR" is the empirical coverage rate of 95% confidence intervals. The proposed PSM estimator $\widehat{\beta}_{\text{psm}}$ is calculated with number of matches $M = 1$ and 5. Four types of variance estimates for $\widehat{\beta}_{\text{psm}}$ were compared: "software", output from the standard software; "asymp", the proposed asymptotic variance estimation; "naiveboot", the naive nonparametric bootstrap; "double-rsp", the proposed double-resampling method.

the weighting approaches. Among all the variance estimators of $\widehat{\beta}_{\text{psm}}$, the asymptotic variance estimator shows the smallest bias.

7 | AN APPLICATION

Non-small cell lung cancer (NSCLC) is the most commonly diagnosed lung cancer; typically, around half the NSCLC patients receiving chemotherapy will receive additional treatment in the post-progression setting, that is, second-line treatment setting, where "carboplatin + paclitaxel" and "erlotinib" are two historically commonly used treatments; see Cui et al.[36] In this section, we use the IMS Health Oncology electronic medical record (EMR) data to conduct a comparative effectiveness analysis of the two treatments in the second-line setting with the PSM estimator and other existing estimators mentioned in the previous section.

The EMR data is deidentified observational patient-level clinical data with demographic and baseline clinical characteristics collected from medium and large community-based oncology practices across 50 states of the USA. The dataset used contains a retrospective cohort of 10,634 eligible patients at least 18 years old who received at least two lines of therapy, from 1 January 2007 to 31 December 2014; see Cui et al[36] for details.

Overall survival was defined as the time from the start date of second-line therapy to the death date. The death date of patients was assumed to be the last visit if a *sufficient period* had elapsed between the last visit and the end of the EMR data; patients with the time between last visit and the end of dataset shorter than a *sufficient period* were censored at the date of the last visit. Here we define a *sufficient period* as at least twice the average visit interval in the 3 months prior to last visit, given that a patients normally have multiple visits to the clinic. Patients alive at the end of the time period were censored at the end date of the dataset. Missing data were classified into its own category for each categorical variable; see de Rooij.[37] Among the eligible patients in the dataset, 1241 patients were treated with "carboplatin + paclitaxel", and 895 patients received single-agent "erlotinib" as second-line therapy, while the remaining patients received one of three other therapies. For illustration of our approach, we subsetted the data to contain only patients receiving "carboplatin + paclitaxel" or "erlotinib," to compare the treatment effects of the two therapies.

TABLE 2 Estimated β_0 and hazard ratio of comparing "carboplatin + paclitaxel" to "erlotinib".

		$\widehat{\beta}$	Estimated hazard ratio	95% Confidence interval
Naive		−0.085	0.918	(0.851, 0.991)
IPW		−0.065	0.937	(0.850, 1.033)
AIPW		−0.075	0.928	(0.842, 1.023)
X.M		−0.076	0.927	(0.836, 1.028)
PSM	asymp	−0.058	0.944	(0.839, 1.062)
($M = 1$)	double-rsp	−0.058	0.944	(0.830, 1.074)
PSM	asymp	−0.009	0.991	(0.868, 1.131)
($M = 5$)	double-rsp	−0.009	0.991	(0.903, 1.087)

Note: The proposed PSM estimator is calculated under the number of matches $M = 1$ and 5. For each PSM estimators, its variance is estimated by the proposed asymptotic variance, "asymp" and the proposed double-resampling method, "double-rsp".

The propensity scores were estimated using a logistic regression model with predictors: age at the initiation of second-line therapy, gender, race, region, disease stage at initial diagnosis, Eastern Cooperative Oncology Group performance status score at initiation of second-line therapy, facility types of academic or community cancer center, year of index diagnosis, and days from index diagnosis to initiation of second-line therapy; see Cui et al[36] for details.

Table 2 shows estimated log hazard ratio $\widehat{\beta}$, point estimates and 95% confidence intervals for the hazard ratio of "carboplatin + paclitaxel" to "erlotinib" based on the unadjusted estimator $\widehat{\beta}_{\text{nai}}$, the IPW estimator $\widehat{\beta}_{\text{ipw}}$, and matching estimator based on the full set of covariates $\widehat{\beta}_{\text{x.m}}$, with the robust variance estimation from the standard software for constructing Wald confidence intervals. We also include the AIPW estimator proposed by Tchetgen Tchetgen and Robins[10] where the working outcome model is the Cox PH model based on the same set of covariates as the propensity score model and treatment indicator. The variance of $\widehat{\beta}_{\text{aipw}}$ is estimated by a naive bootstrap procedure. For the proposed inference based on the PSM estimator $\widehat{\beta}_{\text{psm}}$ with the number of matches $M = 1$ and 5, we constructed both Wald confidence intervals based on the empirical asymptotic variance estimator and bootstrap percentile confidence intervals with the proposed double-resampling method.

All adjusted methods give larger hazard ratio estimates than the unadjusted naive method. Although the point estimates of the PSM approach with $M = 1$ and $M = 5$ indicate that "carboplatin + paclitaxel" might be slightly more advantageous, the 95% confidence intervals using both inference approaches include 1, implying insufficient evidence for a statistically significant difference at the 0.05 level in the effectiveness comparison. Similar to the simulation results, the PSM estimator with $M = 5$ provides narrower interval than the one with $M = 1$ for the double-resampling approach. The 95% confidence intervals of the other adjusted approaches presented in Table 2, including IPW, AIPW, and matching on covariates, also contain 1, reaching the same conclusion as the PSM estimators.

8 | DISCUSSION AND FUTURE STUDIES

PSM is prevailing in practice to handle confounding in observational studies. We establish the statistical properties of the PSM estimator of the marginal causal hazard ratio based on matching with replacement and a fixed number of matches. We also propose a double-resampling technique for variance estimation that takes into account the uncertainty due to propensity score estimation prior to matching. Existing simulation studies have indicated that for estimation of the average treatment effect, IPW or AIPW can perform unstably when extreme values of the estimated propensity scores are present, and instead PSM or alternative approaches such as the overlap weights should be considered and leveraged. Our simulation results echo the previous findings in the survival context for estimation of the marginal hazard ratio.

The theoretical results established in this article hold for any fixed number of matches M. We illustrated two different choices for M in the simulation study and real data analysis in order to show the validity of the resampling-based variance estimator. Intuitively, different choices of M affect the PSM estimator through a bias-variance trade-off. In practice, often a small M is recommended, since the increase in bias is often more significant than the reduction in the variance.[18,38]

Following the cross-validation idea for the matching estimator with continuous outcomes proposed in Xu,[39] we leave the development of an optimal way of choosing M for the PSM estimator as a topic for future research.

In this work, we derived the asymptotic distribution of the PSM estimator of the marginal hazard ratio assuming a generalized linear model for the propensity score. Notably, the same proof technique can potentially be extended to cases when the propensity scores are estimated assuming certain semiparametric or nonparametric models (eg, single- or multiple-index models; Huang and Chan).[40] Take the semiparametric single-index model for example. In this case, the key insight is that PSM does not rely on the exact functional form of the propensity score model but a sufficient dimension reduction of the mean space of A_i given X_i. We have $e(X_i) = e(X_i^T\theta_0)$, where $\theta_0 \in \mathbb{R}^p$ is a vector of unknown parameters and $e(\cdot)$ is left unspecified and does not require a restrictive parametric model assumption. Although the link function does not permit a root-n consistent estimator, the index $X^{\mathrm{T}}\theta$ enables a root-n consistent estimator,[40] denoted as $X^{\mathrm{T}}\widehat{\theta}$. Then it suffices to implement the PSM based on $X^{\mathrm{T}}\widehat{\theta}$. More specifically, in the proof of Theorem 2, under a parametric propensity score model, P^{θ_0} for (15) is naturally the probability measure governed by the likelihood function of θ_0. Under the single- or multiple index propensity score model, following Andreou and Werker,[22] we can formulate a working model for P^{θ_0} based on the asymptotic distribution of $X^{\mathrm{T}}\widehat{\theta}$. The remaining steps for the proof would remain the same and the inferential framework can be carried over. When assuming other nonparametric machine learning models for the propensity score, the same procedure for implementing the PSM estimator can still be applied, and we leave the development of inferential frameworks in those cases as future endeavors.

It is important to draw connections between clinical trials and observational studies from both design and analysis perspectives, which highlights the advantages of PSM and also motivates several future research directions. Similar to clinical trial designs, the matching step uses only the covariate and treatment information and does not touch the outcome data. Therefore, it mitigates the possibility of data snooping and dredging. As discussed in the introduction, PSM alone emulates a completely randomized trial, which may not be the most efficient. Stratified block randomization is often used to improve the complete randomization in clinical trials. Thus, we can imagine that combining stratification and PSM can also improve PSM alone in observational studies. Moreover, in trial data analysis, instead of a simple analysis of the outcome data, ANCOVA can be used to borrow information from auxiliary information. Thus, it is important to continue the development of more efficient analysis methods, such as general M and Z estimators,[41] for the matched observational data. We leave these topics to future research.

This work focuses on estimating the causal PH ratio, a scalar estimand, that summarizes the treatment effect over a certain period of time. There are many other treatment effect estimands for survival outcomes, such as the restrictive mean survival times, restrictive mean lost times, the difference in survival medians, and so on.[42] In the context of survival analyses, Chen and Tsiatis[43] proposed a regression model approach to estimate the average causal effect of restricted mean survival times. Xie and Liu[44] developed the adjusted Kaplan-Merier estimators of treatment-specific survival functions using IPW. Zhang and Schaubel[45] combined Chen and Tsiatis's regression method[43] and the inverse probability weighted Nelson-Aalen estimator, resulting in a doubly robust estimator of the average causal effect of restricted mean survival times. Diaz[46] proposed data-adaptive doubly robust estimators of treatment-specific survival functions. In our future work, we will develop matching estimators of general causal estimands for survival outcomes and compare them with existing approaches.

ACKNOWLEDGEMENTS

Yang is supported in part by the National Science Foundation grant DMS 1811245, National Institute of Aging grant 1R01AG066883, and National Institute of Environmental Health Sciences 1R01ES031651.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Releasing the data are subject to the third party approval.

ORCID

Shu Yang https://orcid.org/0000-0001-7703-707X
Shuhan Tang https://orcid.org/0000-0002-4978-2360
Douglas E. Faries https://orcid.org/0000-0001-8952-7738

REFERENCES

1. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*. 1972;34:187-220.
2. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Anal*. 2013;19(3):279-296.
3. Austin PC, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*. 2007;26(4):754-768.
4. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32:2837-2849.
5. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res*. 2012;21(3):273-293.
6. Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. *Stat Med*. 2014;33(23):4053-4072.
7. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22:523-539.
8. Busso M, DiNardo J, McCrary J. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Rev Econ Stat*. 2014;96(5):885-897.
9. Austin PC, Stuart EA. 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. *Stat Methods Med Res*. 2017;26:1654-1670.
10. Tchetgen Tchetgen EJ, Robins J. On parametrization, robustness and sensitivity analysis in a marginal structural Cox proportional hazards model for point exposure. *Stat Probab Lett*. 2012;82:907-915.
11. Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med*. 2012;31(15):1572-1581.
12. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat*. 2002;84(1):151-161.
13. Rubin DB. *Matched Sampling for Causal Effects*. Cambridge, England: Cambridge University Press; 2006.
14. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25:1-21.
15. Frölich M. Finite-sample properties of propensity-score matching and weighting estimators. *Rev Econ Stat*. 2004;86:77-90.
16. Austin PC, Cafri G. Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Stat Med*. 2020;39(11):1623-1640.
17. Greifer N, Stuart EA. Matching methods for confounder adjustment: an addition to the epidemiologist's toolbox. *Epidemiol Rev*. 2021;43(1):118-129.
18. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006;74: 235-267.
19. Gayat E, Resche-Rigon M, Mary JY, Porcher R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm Stat*. 2012;11:222-229.
20. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Stat Med*. 2014;33:4306-4319.
21. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. *Econometrica*. 2008;76:1537-1557.
22. Andreou E, Werker BJ. An alternative asymptotic analysis of residual-based statistics. *Rev Econ Stat*. 2012;94:88-99.
23. Abadie A, Imbens GW. Matching on the estimated propensity score. *Econometrica*. 2016;84:781-807.
24. Adusumilli K. Bootstrap inference for propensity score matching. Working Paper. 2022.
25. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11:561-570.
26. Robins JM. Optimal structural nested models for optimal sequential decisions. Proceedings of the Second Seattle Symposium in Biostatistics, Springer, New York, 189-326. 2004.
27. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge UK: Cambridge University Press; 2015.
28. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2017;113:390-400.
29. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61:962-973.
30. Le Cam L, Yang GL. *Asymptotics in Statistics: some Basic Concepts*. Berlin: Springer; 1990.
31. Otsu T, Rai Y. Bootstrap inference of matching estimators for average treatment effects. *J Am Stat Assoc*. 2017;112(520): 1720-1732.
32. Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*. 2016;72:1055-1065.
33. Zhao H, Zhang X, Yang S. Double score matching in observational studies with multi-level treatments. *Commun Stat Simul Comput*. 2022;1-17.
34. Zhao H. *Advances in Matching Methods for Causal Inference with Multiple Treatments*. PhD thesis. Raleigh, NC: North Carolina State University; 2023.
35. Wang L, Tchetgen Tchetgen E, Martinussen T, Vansteelandt S. Instrumental variable estimation of the causal hazard ratio. *Biometrics*. 2023;79(2):539-550.
36. Cui ZL, Hess LM, Goodloe R, Faries D. Application and comparison of generalized propensity score matching versus pairwise propensity score matching. *J Comp Eff Res*. 2018;7:923-934.
37. Rooij M. Transitional modeling of experimental longitudinal data with missing values. *Adv Data Anal Classif*. 2018;12:107-130.

38. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *Am J Epidemiol*. 2010;172(9):1092-1097.
39. Xu T. *Advances in Causal Inference and the Study of Interlocus Gene Conversion*. PhD thesis. Raleigh, NC: North Carolina State University; 2023.
40. Huang MY, Chan KCG. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*. 2017;104:583-596.
41. Vaart AW. *Asymptotic Statistics*. Cambridge: Cambridge University Press; 1998.
42. Yang S, Zhang Y, Liu GF, Guan Q. SMIM: a unified framework of survival sensitivity analysis using multiple imputation and martingale. arXiv Preprint arXiv:200702339. 2020.
43. Chen PY, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*. 2001;57: 1030-1038.
44. Xie J, Liu C. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med*. 2005;24:3089-3110.
45. Zhang M, Schaubel DE. Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics*. 2012;68:999-1009.
46. Díaz I. Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Stat Med*. 2019;38:2735-2748.
47. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96:187-199.
48. Yang S, Ding P. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*. 2018;105:487-493.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang T, Zhao H, Yang S, et al. Propensity score matching for estimating a marginal hazard ratio. *Statistics in Medicine*. 2024;1-28. doi: 10.1002/sim.10103

APPENDIX A. COMPARISON BETWEEN AIPW AND PSM FOR ESTIMATING THE MARGINAL HAZARD RATIO

The AIPW estimator is also a popular approach for estimating the average treatment effect for a continuous outcome and has been adapted to estimate the marginal hazard ratio in the survival context.[10] The AIPW estimator is consistent as long as either the propensity score model or the outcome model is correctly specified. In comparison, PSM does not rely on the outcome model and is consistent if the propensity score model is correctly specified. The AIPW estimator is semiparametrically efficient when both models are correctly specified. In comparison, based on Theorem 1, the PSM estimator based on a correctly specified propensity score model will be less efficient than AIPW asymptotically as it does not in general attain the semiparametric efficiency bound. However, as shown in our simulation study, in cases where the overlap of covariate distributions between the two treatment groups is poor, PSM can yield more stable marginal hazard ratio estimates than AIPW in finite samples. This phenomenon is in line with findings from earlier numerical studies involving survival or continuous outcomes.[8,9] This is in part because AIPW weights the outcomes by the inverse of the estimated propensity scores, whereas PSM reuses the original value of the outcomes regardless of how extreme the propensity scores are.

Trimming offers a general solution to violation of the overlap assumption by excluding units with extreme propensity scores from the analysis.[47,48] However, trimming methods necessarily alter the target estimand and restrict inference to regions of the covariate space with sufficient overlap. PSM is more resistant to the violation of the overlap assumption because even for units with extreme propensity scores, their missing potential outcomes can still be imputed with minimal bias as long as they have close matches from the opposite treatment group. Therefore, trimming is not needed for PSM except maybe when the matching discrepancies become noticeably large.

APPENDIX B. ASYMPTOTIC VARIANCE ESTIMATION

In this section, we discuss estimation of the large sample variances of $\widehat{\beta}$ adjusting for first step estimation of the propensity score. Recall that we have that

$$V_2 = \{A(\beta_0)\}^{-1}\{V_G - c^{\mathrm{T}}\mathcal{I}_{\theta_0}^{-1}c\}\{A(\beta_0)\}^{-1}.$$

We will estimate each component on the right hand side of the equation separately. We first estimate the Fisher information $\mathcal{I}_{\theta_0}$ using

$$\widehat{\mathcal{I}}_{\theta_0} = \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{e}^2(X_i^{\mathrm{T}}\widehat{\theta})}{e(X_i^{\mathrm{T}}\widehat{\theta})(1-e(X_i^{\mathrm{T}}\widehat{\theta}))} X_i X_i^{\mathrm{T}}.$$

Let $m_k\{\omega, e(X_i^{\mathrm{T}}\widehat{\theta})\}$ denote the index of the k-th nearest neighbor matched to unit i based on the estimated propensity scores. For estimation of V_G, we first create an imputed dataset $\left\{H_{i1}^*(\omega), H_{i2}^*(\omega)\right\}_{i=1}^{n}$ for $\omega = 0, 1$, where

$$H_{i1}^*(\omega) = \begin{cases} H_i(\omega) & \text{if } W_i = \omega \\ H_{m_1\left\{\omega, e(X_i^{\mathrm{T}}\widehat{\theta})\right\}}(\omega) & \text{if } W_i \neq \omega \end{cases}$$

and

$$H_{i2}^*(\omega) = \begin{cases} H_{m_1\left\{\omega, e(X_i^{\mathrm{T}}\widehat{\theta})\right\}}(\omega) & \text{if } W_i = \omega \\ H_{m_2\left\{\omega, e(X_i^{\mathrm{T}}\widehat{\theta})\right\}}(\omega) & \text{if } W_i \neq \omega. \end{cases}$$

Then, $\widetilde{V}_G$ can be estimated by

$$\widehat{V}_G = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\omega=0}^{1} H_{i1}^*(\omega)\right\}^2 + \frac{1}{n}\sum_{i=1}^{n}\left\{k_{\widehat{\theta},i}(W_i) + k_{\widehat{\theta},i}(W_i)^2\right\}\widehat{\sigma}_i^2,$$

where $\widehat{\sigma}_i^2 = \sum_{k=1}^{2}\left[H_{ik}^*(\omega) - \frac{1}{2}\left\{H_{i1}^*(\omega) + H_{i2}^*(\omega)\right\}\right]^2 = \frac{1}{2}\left\{H_{i1}^*(\omega) - H_{i2}^*(\omega)\right\}^2$.

Next we can construct an estimator of c by averaging over the sample:

$$\widehat{c} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\mathrm{cov}}\left\{X_i, \mu_H(1, X_i)|e(X_i^{\mathrm{T}}\theta_0)\right\}}{e(X_i^{\mathrm{T}}\widehat{\theta})} + \frac{\widehat{\mathrm{cov}}\left\{X_i, \mu_H(0, X_i)|e(X_i^{\mathrm{T}}\theta_0)\right\}}{1 - e(X_i^{\mathrm{T}}\widehat{\theta})}\right]\dot{e}(X_i^{\mathrm{T}}\widehat{\theta}).$$

For estimation of the conditional covariance, we follow the same matching procedure to create an imputed dataset $\left\{X_{i1}^*(\omega), X_{i2}^*(\omega)\right\}_{i=1}^{n}$. Then $\widehat{\mathrm{cov}}\left\{X_i, \mu_H(\omega, X_i)|e_\omega(X_i^{\mathrm{T}}\theta_0)\right\}$ can be estimated by $\frac{1}{2}\left\{X_{i1}^*(\omega) - X_{i2}^*(\omega)\right\}\left\{H_{i1}^*(\omega) - H_{i2}^*(\omega)\right\}$.

Finally, to estimate $A(\beta_0)$, we use

$$\widehat{A}(\beta_0) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\left\{\widehat{Q}(\widehat{\beta}, t_k) - \widehat{Q}(\widehat{\beta}, t_k)^2\right\}\left\{\sum_{\omega=0}^{1} \mathrm{d}\overline{N}_i^{*(\omega)}(t_k)\right\}\right],$$

where $\{t_1, \ldots t_K\}$ are distinct observed time points. Putting everything together, our final estimator of the asymptotic variance is

$$\widehat{V}_2 = \{\widehat{A}(\beta_0)\}^{-1}\{\widehat{V}_G - \widehat{c}^{\mathrm{T}}\widehat{\mathcal{I}}_{\theta_0}^{-1}\widehat{c}\}\{\widehat{A}(\beta_0)\}^{-1}.$$

APPENDIX C. REGULARITY CONDITIONS AND LEMMAS

In this section, we provide the regularity conditions and lemmas. For simplicity, we introduce more notations. Let $\mathcal{N} \equiv \{\theta : ||\theta - \theta_0|| < \epsilon\}$ be a neighborhood of θ_0 given an $\epsilon > 0$. We use a generalized linear specification for the propensity score, $e(x) = e(X^{\mathrm{T}}\theta)$ where $e(\cdot)$ is a link function. Moreover, denote $S_n(\beta, t) = n^{-1}\sum_{j=1}^{n}\{1 + k_{\hat{\theta},i}/M\}I(W_j = \omega)\exp(\beta W_i)Y_j(t)$ and $S_n'(\beta, t) = \partial S_n(\beta, t)/\partial\beta = n^{-1}\sum_{j=1}^{n}\{1 + k_{\hat{\theta},i}/M\}I(W_j = \omega)\exp(\beta W_i)Y_j(t)W_i$, then $\widehat{Q}(\beta, t) = S_n(\beta, t)/S_n'(\beta, t)$.

Assumption A1. The following regularity conditions hold:

(i) $\theta_0 \in int(\Theta)$ with Θ compact, X has a bounded support and $\mathrm{E}[X^{\mathrm{T}}X]$ is non-singular;
(ii) $e(\cdot) : \mathbb{R} \to (0, 1)$ is twice continuously differentiable with strictly bounded first and second derivatives, $\dot{e}(\cdot)$ and $\ddot{e}(\cdot)$ where $\dot{e}(\cdot)$ is strictly positive;
(iii) $\forall\theta \in \mathcal{N}$, the random variable $e(X^{\mathrm{T}}\theta)$ is continuously distributed with interval support, and its pdf $g_\theta(\cdot)$ is uniformly Lipschitz continuous over θ;
(iv) there exists a component of X that is continuously distributed, has nonzero coefficient in θ_0, and has a continuous density function conditional on the rest of X;
(v) $\forall\theta \in \mathcal{N}$ and $\omega = 0, 1$, $\mu_H(\omega, X)$ and $\sigma_H^2(\omega, X)$ is Lipschitz-continuous in p with the Lipschitz constants independent of θ;
(vi) $E\left(|H_i(\omega)|^{4+\delta}|W_i = \omega, X_i = x\right)$ is uniformly bounded over the support of X, for some $\delta > 0$;
(vii) $\mathrm{E}\left\{\mathrm{d}M^{(1)}(t)|p\right\}$ is Lipschiz continuous in p;
(viii) There exists $\tau > 0$ is such that $\int_0^\tau \lambda_0(t)d(t) < \infty$;
(ix) there exists a neighborhood $\mathcal{B}$ of β_0 and functions $s_0(\beta, t)$ and $s_1(\beta, t)$such that $sup_{t\in[0,\tau],\beta\in\mathcal{B}}||S_n(\beta, t) - s_0(\beta, t)|| \to_p 0$ and $sup_{t\in[0,\tau],\beta\in\mathcal{B}}||S_n'(\beta, t) - s_1(\beta, t)|| \to_p 0$ as $n \to \infty$;
(x) the function $s_0(\beta, t)$ and $s_1(\beta, t)$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$;
(xi) $\forall\epsilon > 0$, there exists $\delta > 0$ such that $||\beta - \beta_0|| < \delta$ implies $|s_l(\beta, t) - s_l(\beta_0, t)| < \epsilon, \forall t \in [0, \tau], l = 0, 1$;
(xii) $n^{-1}\mathrm{d}G_n(\tilde{\beta})/\mathrm{d}\beta > 0$ for $\forall\tilde{\beta} \in \mathcal{B}$.

The regularity conditions (i)–(vi) are adopted from Abadie and Imbens[23] and Adusumilli,[24] modified for survival outcomes. The regularity conditions (vii)–(xi) are standard in the survival analysis literature and are often assumed for technical convenience. We note here that since β is a scale and ω only takes values in $\{0, 1\}$, the first and second derivatives of $S_n(\beta, t)$ with respect to β are the same. Therefore, it suffices to impose the regulation conditions on the first derivative in (vii)–(xi).

Assumption A2. Additional regularity conditions for variance estimation hold:

(i) Uniformly over all $\theta \in \mathcal{N}$, it holds under $\mathbb{P}_0$, $N^{-1}\sum_{i=1}^{N}\left(\hat{e}_{1i}(\theta) - e_{1i}(\theta)\right)^2 = o_p(N^{-\xi})$ and $N^{-1}\sum_{i=1}^{N}\left(\hat{e}_{2i}(W_i; \theta) - e_{2i}(W_i; \theta)\right)^2 = o_p(N^{-\xi})$ for some $\xi > 0$;
(ii) the number of quantile partitions satisfies $q_N \to \infty$ and $q_N^{2+\eta} \to 0$ as $N \to \infty$ for some $\eta > 0$.

The Assumption A2 are adopted from Adusumilli[24] to ensure the validity of the double-resampling bootstrap estimator.

Lemma A1 below is Lemma S.11 in Abadie and Imbens,[23] which is useful in the proofs of our results.

Lemma A1. *Suppose that $(W_1, X_1), \ldots, (W_n, X_n)$ are independent and identically distributed, where X_i is a scalar continuous variable with a bounded support. Suppose also that $\sigma_H^2(\omega, x)$ is uniformly bounded over the support for X. Let $n_\omega = \sum_{i=1}^{n}I(W_i = \omega)$ be the number of individuals received treatment ω, and $p^* = pr(W = 1) > 0$ and $f_\omega(X)$ be the density function of the scalar continuous variable X when $W = \omega$. Then, for a given ω,*

$$\frac{1}{n_\omega}\sum_{i=1}^{n}I(W_i = \omega)\sigma_H^2(\omega, X_i)k_i \to M\mathrm{E}\left\{\sigma_H^2(\omega, X)\left(\frac{p^*}{1 - p^*}\right)^{1-2\omega}\frac{f_{1-\omega}(X)}{f_\omega(X)}|W_i = \omega\right\},$$

and

$$\frac{1}{n_\omega}\sum_{i=1}^{n} I(W_i=\omega)\sigma_H^2(\omega,X_i)k_i^2 \to M\mathrm{E}\left\{\sigma_H^2(\omega,X)\left(\frac{p^*}{1-p^*}\right)^{1-2\omega}\frac{f_{1-\omega}(X)}{f_\omega(X)}|W_i=\omega\right\} + \frac{M(2M+1)}{2}\mathrm{E}\left\{\sigma_H^2(\omega,X)\left[\left(\frac{p^*}{1-p^*}\right)^{1-2\omega}\frac{f_{1-\omega}(X)}{f_\omega(X)}\right]^2|W_i=\omega\right\},$$

in probability, as $n \to \infty$.

APPENDIX D. PROOF OF THE ASYMPTOTIC UNBIASEDNESS OF $N^{-1}G_N(\beta_0)$

This section includes three parts that follow the similar logic of proof. The first and the second parts provide some results useful for later sections. The proof for the asymptotic unbiasedness of $n^{-1}G_n(\beta_0)$ is located in the third part.

For $\omega = 0, 1$, define $\mathrm{d}M^{(\omega)}(t) = \mathrm{d}N^{(\omega)}(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_0\omega)Y^{(\omega)}(t)$. From the standard theory for the counting process, $\mathrm{d}M^{(\omega)}(t)$ is a martingale process with respect to the population and its baseline hazard is $\Lambda_0(t)$. Next we will prove that

$$I(W_i=\omega)\{1+k_i/M\}\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_0\omega)Y_i(t)\}$$

is a martingale for the imputed pseudo-population which means that the imputed pseudo-population has similar covariates distribution with the target population. First, we show that for $\omega = 0, 1$,

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n} I(W_i=\omega)\{1+k_i/M\}\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_0\omega)Y_i(t)\} \\ &\to \mathrm{E}\left\{\mathrm{d}N^{(\omega)}(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_0\omega)Y^{(\omega)}(t)\right\} = \mathrm{E}\left\{\mathrm{d}M^{(\omega)}(t)\right\}, \end{aligned} \tag{D1}$$

as $n \to \infty$. We show (D1) for $\omega = 1$. The proof for $\omega = 0$ is similar and therefore omitted. We express (D1) for $\omega = 1$ as

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n} W_i\{1+k_i/M\}\{\mathrm{d}N_i(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_0)Y_i(t)\} - \mathrm{E}\left\{\mathrm{d}M^{(1)}(t)\right\} \\ &= n^{-1}\sum_{i=1}^{n} W_i\{1+k_i/M\}\left\{\mathrm{d}N_i^{(1)}(t) - \mathrm{d}\Lambda_0(t)\exp(\beta_0)Y_i^{(1)}(t)\right\} - \mathrm{E}\left\{\mathrm{d}M^{(1)}(t)\right\} \\ &= n^{-1}\sum_{i=1}^{n} W_i\{1+k_i/M\}\mathrm{d}M_i^{(1)}(t) - \mathrm{E}\left\{\mathrm{d}M^{(1)}(t)\right\} \\ &= n^{-1}\sum_{i=1}^{n} W_i\{1+k_i/M\}\left[\mathrm{d}M_i^{(1)}(t) - \mathrm{E}\left\{\mathrm{d}M^{(1)}(t)|e(X_i)\right\}\right] \\ &+ n^{-1}\sum_{i=1}^{n}(1-W_i)\left[\mathrm{E}\left\{\mathrm{d}M^{(1)}(t)|e(X_{m\{1,e(X_i)\}})\right\} - \mathrm{E}\left\{\mathrm{d}M^{(1)}(t)|e(X_i)\right\}\right] \\ &+ n^{-1}\sum_{i=1}^{n}\mathrm{E}\left\{\mathrm{d}M^{(1)}(t)|e(X_i)\right\} - \mathrm{E}\left\{\mathrm{d}M^{(1)}(t)\right\} \\ &= T_1 + T_2 + T_3, \end{aligned}$$

where the second line follows by the consistent assumption, and

$$
\begin{aligned}
T_1 &= n^{-1}\sum_{i=1}^{n} W_i\{1+k_i/M\}\Big[\mathrm{d}M_i^{(1)}(t) - \mathrm{E}\big\{\mathrm{d}M^{(1)}(t)|e(X_i)\big\}\Big],\\
T_2 &= n^{-1}\sum_{i=1}^{n}(1-W_i)\big[\mathrm{E}\big\{\mathrm{d}M^{(1)}(t)|e(X_{m\{1,e(X_i)\}})\big\} - \mathrm{E}\big\{\mathrm{d}M^{(1)}(t)|e(X_i)\big\}\big],\\
T_3 &= n^{-1}\sum_{i=1}^{n}\mathrm{E}\big\{\mathrm{d}M^{(1)}(t)|e(X_i)\big\} - \mathrm{E}\big\{\mathrm{d}M^{(1)}(t)\big\}.
\end{aligned}
\tag{D2}
$$

Under Assumption A1 (i)–(vi), Abadie and Imbens[18] showed that k_i^{δ} is bounded almost surely for any $\delta > 0$, and the discrepancy due to matching is $|M^{-1}\sum_{j\in\mathcal{J}_M(1,e(X_i))} e(X_j) - e(X_i)| = O_p(n^{-1})$ for a scalar $e(X)$. It follows that T_1 is consistent for zero. Moreover, under Assumption A1 (vi), T_2 is consistent for zero. Lastly, by the strong law of large numbers, T_3 is consistent for zero. Therefore, (D1) follows. Since $n^{-1}\sum_{i=1}^{n} I(W_i=\omega)(1+k_i/M)\Big\{\omega - \widehat{Q}(\beta_0,t)\Big\}\mathrm{d}M_i^{(\omega)}(t)$ is bounded, by dominated convergence theorem,

$$
n^{-1}G_n(\beta_0) = n^{-1}\sum_{\omega=0}^{1}\sum_{i=1}^{n}\int_0^{\tau} I(W_i=\omega)\{1+k_i/M\}\Big\{\omega - \widehat{Q}(\beta_0,t)\Big\}\mathrm{d}M_i^{(\omega)}(t) \tag{D3}
$$

$$
\to \sum_{\omega=0}^{1}\int_0^{\tau}\Big\{\omega - \widehat{Q}(\beta_0,t)\Big\}\mathrm{E}\big\{\mathrm{d}M^{(\omega)}(t)\big\} = 0. \tag{D4}
$$

APPENDIX E. PROOF FOR THEOREM 1

Taylor expansion of $G_n(\widehat{\beta}) = 0$ around β_0 leads to

$$
0 = G_n(\widehat{\beta}) = G_n(\beta_0) + \frac{\mathrm{d}}{\mathrm{d}\beta}G_n(\tilde{\beta})(\widehat{\beta} - \beta_0),
$$

where $\tilde{\beta}$ is on the line segment between $\widehat{\beta}$ and β_0. Then,

$$
n^{1/2}(\widehat{\beta} - \beta_0) = \left\{ n^{-1}\frac{\mathrm{d}G_n(\tilde{\beta})}{\mathrm{d}\beta}\right\}^{-1} n^{-1/2}G_n(\beta_0). \tag{E1}
$$

Under Assumption 3 (ix), $n^{-1}\frac{\mathrm{d}G_n(\tilde{\beta})}{\mathrm{d}\beta} > 0$ for $\tilde{\beta} \in \mathcal{B}$. Then, the reminder is to show the asymptotic distribution of $n^{-1/2}G_n(\beta_0)$.

Theorem A1. *Suppose Assumptions 1–3 and Assumption A1 hold and X is a continuous scalar variable. Then,*

$$
n^{-1/2}G_n(\beta_0) \to \mathcal{N}(0, \overline{V}_G), \tag{E2}
$$

in distribution, as n → ∞, where

$$
\overline{V}_G = \sum_{\omega=0}^{1} E\Big[\sigma_H^2(\omega,X)\Big\{\tfrac{2M+1}{2M}\tfrac{1}{p(\omega|X)} - \tfrac{1}{2M}p(\omega|X)\Big\}\Big] + \mathrm{E}\big[\{\mu_H(0,X) + \mu_H(1,X)\}^2\big]. \tag{E3}
$$

Proof. We will show that $n^{-1/2}G_n(\beta_0)$ can be expressed as a sum of n independent and identically distributed random vectors plus a term that converges in probability to a zero vector. By some algebra, we obtain

$$
\sum_{i=1}^{n}\{1+k_i/M\}\Big\{W_i - \widehat{Q}(\beta_0,t)\Big\}\mathrm{d}\Lambda_0(t)\exp(\beta_0 W_i)Y_i(t) = 0. \tag{E4}
$$

Therefore, continuing with (D3), we obtain

$$
\begin{aligned}
&n^{-1/2}G_n(\beta_0)\\
&= n^{-1/2}\sum_{i=1}^{n}\{1+k_i/M\}\int_0^{\tau}\left\{W_i-\widehat{Q}(\beta_0,t)\right\}\mathrm{d}M_i(t) && \text{(E5)}\\
&= n^{-1/2}\sum_{i=1}^{n}\{1+k_i/M\}\int_0^{\tau}\left\{W_i-\widehat{Q}(\beta_0,t)\right\}\{\mathrm{d}N_i(t)-\mathrm{d}\Lambda_0(t)\exp(\beta_0W_i)Y_i(t)\} && \text{(E6)}\\
&= n^{-1/2}\sum_{i=1}^{n}\{1+k_i/M\}\int_0^{\tau}\{W_i-Q(\beta_0,t)\}\left\{\mathrm{d}N_i(t)-\mathrm{d}\Lambda_0(t)e^{\beta_0W_i}Y_i(t)\right\}\\
&+ n^{-1/2}\sum_{i=1}^{n}\{1+k_i/M\}\int_0^{\tau}\left\{Q(\beta_0,t)-\widehat{Q}(\beta_0,t)\right\}\left\{\mathrm{d}N_i(t)-\mathrm{d}\Lambda_0(t)e^{\beta_0W_i}Y_i(t)\right\}, && \text{(E7)}
\end{aligned}
$$

where (E6) follows from (E4). Under Assumptions S1 (viii), (ix) and (xi), there exists a function $Q(\beta_0, t)$ such that,

$$
\int_0^{\tau}\left\{\widehat{Q}(\beta_0,t)-Q(\beta_0,t)\right\}\left[n^{-1/2}\sum_{i=1}^{n}(1+k_i/M)\left\{\mathrm{d}N_i(t)-\mathrm{d}\Lambda_0(t)e^{\beta_0W_i}Y_i(t)\right\}\right]\to 0,
$$

in probability, as $n \to \infty$. Therefore, (E7) becomes

$$
n^{-1/2}G_n(\beta_0)=n^{-1/2}\sum_{i=1}^{n}\{1+k_i/M\}H_i(W_i)+o_p(1), \tag{E8}
$$

where $H_i(\omega)$ is defined in (11). Moreover,

$$
\begin{aligned}
\mathrm{E}\{H_i(W_i)\} &= \int_0^{\tau}\mathrm{E}\big[\{W_i-Q(\beta_0,t)\}\{\mathrm{d}N_i(t)-\mathrm{d}\Lambda_0(t)\exp(\beta_0W_i)Y_i(t)\}\big]\\
&= \int_0^{\tau}\mathrm{E}\big[e(X_i)\{1-Q(\beta_0,t)\}\mathrm{d}M^{(1)}(t)\big]\\
&\quad - \int_0^{\tau}\mathrm{E}\big[\{1-e(X_i)\}Q(\beta_0,t)\mathrm{d}M^{(0)}(t)\big]\\
&= 0,
\end{aligned}
$$

where the last line follows by the martingale property for the potential outcome process.

We write

$$
\begin{aligned}
&n^{-1/2}G_n(\beta_0)\\
&= n^{-1/2}\left\{G_n(\beta_0)-\sum_{\omega=0}^{1}\sum_{i=1}^{n}\mu_H(\omega,X_i)\right\}+n^{-1/2}\sum_{\omega=0}^{1}\sum_{i=1}^{n}\mu_H(\omega,X_i)\\
&= n^{-1/2}\left[\sum_{i=1}^{n}\{1+k_i/M\}H_i(W_i)-\sum_{i=1}^{n}\sum_{\omega=0}^{1}\mu_H(\omega,X_i)\right]+n^{-1/2}\sum_{\omega=0}^{1}\sum_{i=1}^{n}\mu_H(\omega,X_i)\\
&= \sum_{\omega=0}^{1}n^{-1/2}\sum_{i=1}^{n}\Big[I(W_i=\omega)\{1+k_i/M\}\{H_i(\omega)-\mu_H(\omega,X_i)\}\\
&\qquad +\{1-I(W_i=\omega)\}\{\mu_H(\omega,X_{m(\omega,X_i)})-\mu_H(\omega,X_i)\}+\mu_H(\omega,X_i)\Big].
\end{aligned}
$$

Similar to (D2), we have $n^{-1/2}\sum_{i=1}^{n}\{1-I(W_i=\omega)\}\{\mu_H(\omega,X_{m(\omega,X_i)})-\mu_H(\omega,X_i)\}=O_p(n^{-1/2})=o_p(1)$, for $\omega=0,1$. Therefore, we can write

$$n^{-1/2}G_n(\beta_0)=\sum_{l=1}^{2n}\xi_{n,l}+o_p(1),$$

where

$$\xi_{n,l}=\begin{cases} n^{-1/2}\{\mu_H(0,X_l)+\mu_H(1,X_l)\}, & 1\le l\le n,\\ n^{-1/2}\{1+k_{l-n}\}\{H_{l-n}(W_{l-n})-\mu_H(W_{l-n},X_{l-n})\}, & n+1\le l\le 2n.\end{cases}$$

Consider the σ-fields

$$\mathcal{F}_{n,l}=\begin{cases}\{W_1,\ \dots\ ,W_l,X_1,\ \dots\ ,X_l\}, & 1\le l\le n,\\ \{W_1,\ \dots\ ,W_n,X_1,\ \dots\ ,X_n,H_1(W_1),\ \dots\ ,H_{l-n}(W_{l-n})\}, & n+1\le l\le 2n.\end{cases}$$

Then, for each $n\ge 1$,

$$\left\{\sum_{j=1}^{l}\xi_{n,j},\mathcal{F}_{n,l},1\le l\le 2n\right\}$$

is a martingale. Moreover, we evaluate

$$\sum_{l=1}^{n}E\left(\xi_{n,l}^2|\mathcal{F}_{n,l-1}\right)\to \mathrm{E}\left[\{\mu_H(0,X)+\mu_H(1,X)\}^2\right],$$

and based on Lemma A1

$$\begin{aligned}\sum_{l=n+1}^{2n}E\left(\xi_{n,l}^2|\mathcal{F}_{n,l-1}\right)&=n^{-1}\sum_{i=1}^{n}\sum_{\omega=0}^{1}I(W_i=\omega)\{1+k_i/M\}^2\sigma_H^2(\omega,X_i)\\ &\to\sum_{\omega=0}^{1}\mathrm{E}\left[\sigma_H^2(\omega,X)\left\{\frac{2M+1}{2M}\frac{1}{p(\omega|X)}-\frac{1}{2M}p(\omega|X)\right\}\right],\end{aligned}$$

as $n\to\infty$. Apply the Central Limit Theorem for martingale arrays, (E2) follows. ■

To establish the result in Theorem 1, we replace X by $e(X)$ as the matching variable; therefore, (E2) holds for

$$\begin{aligned}V_G=&\sum_{\omega=0}^{1}\mathrm{E}\left[\sigma_H^2\{\omega,e(X)\}\left\{\frac{2M+1}{2M}\frac{1}{p(\omega|X)}-\frac{1}{2M}p(\omega|X)\right\}\right]\\ &+E\left([\mu_H\{0,e(X)\}+\mu_H\{1,e(X)\}]^2\right).\end{aligned}\tag{E9}$$

Combining (E1) and (E9), $n^{1/2}(\hat{\beta}-\beta_0)$ is asymptotically normal with mean zero and covariance matrix $V_1=A(\beta_0)^{-1}V_GA(\beta_0)^{-1}$. This completes the proof for Theorem 1.

APPENDIX F. PROOF FOR THEOREM 2

Theorem A2. *Suppose that Assumptions 1–3 and Assumption A1 hold. Suppose that $e(X)$ follows a logistic regression model $e(X^{\mathrm{T}}\theta)$ with the true parameter value θ_0. Let $\hat{\theta}$ be the maximum likelihood estimator for θ, and $\mathcal{I}_{\theta_0}$ be the Fisher information matrix. Then, based on matching on the estimated propensity score $e(X^{\mathrm{T}}\hat{\theta})$,*

$$n^{-1/2}G_n(\beta_0) \to \mathcal{N}\left(0, V_G - c^{\mathrm{T}}\mathcal{I}_{\theta_0}^{-1}c\right),$$

where

$$c = \mathrm{E}\left\{\left[\frac{\operatorname{cov}\{X, \mu_H(1,X)|e(X)\}}{e(X)} + \frac{\operatorname{cov}\{X, \mu_H(0,X)|e(X)\}}{1-e(X)}\right]\dot{e}(X^{\mathrm{T}}\theta_0)\right\}.$$

Proof. Let $Z_i = \{X_i, W_i, H_i(W_i)\}$, and let $L(\theta|Z_1, \dots, Z_n)$ be the log likelihood function of θ, that is,

$$\begin{aligned} L(\theta|Z_1, \dots, Z_n) &= \log\left[\prod_{i=1}^{n} e(X_i^{\mathrm{T}}\theta)^{W_i}\{1-e(X_i^{\mathrm{T}}\theta)\}^{1-W_i}\right] \\ &= \sum_{i=1}^{n}\left[W_i \log e(X_i^{\mathrm{T}}\theta) + (1-W_i)\log\{1-e(X_i^{\mathrm{T}}\theta)\}\right]. \end{aligned}$$

Following Abadie and Imbens,[23] we use the local experiment argument. Let $\theta_n = \theta_0 + n^{-1/2}h$, and P^{θ_n} be the data distribution under $e(X^{\mathrm{T}}\theta_n)$. Also, we define

$$\Lambda_n(\theta_0|\theta_n) = L(\theta_0|Z_1, \dots, Z_n) - L(\theta_n|Z_1, \dots, Z_n). \tag{F1}$$

Under P^{θ_n}, we can express $n^{-1/2}G_n(\beta_0) = D_n(\theta_n) + o_P(1)$, where

$$\begin{aligned} D_n(\theta_n) = n^{-1/2}\sum_{i=1}^{n}\sum_{\omega=0}^{1}\Big[& I(W_i=\omega)\{1+k_i/M\}\left[H_i(\omega) - \mu_H\left\{\omega, e(X_i^{\mathrm{T}}\theta_n)\right\}\right] \\ &+ \mu_H\left\{\omega, e(X_i^{\mathrm{T}}\theta_n)\right\}\Big]. \end{aligned}$$

We shall show that under P^{θ_n} :

$$\begin{pmatrix} D_n(\theta_n) \\ n^{1/2}(\widehat{\theta}_n - \theta_n) \\ \Lambda_n(\theta_0|\theta_n) \end{pmatrix} \to \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \\ -h^{\mathrm{T}}\mathcal{I}_{\theta_0}h/2 \end{pmatrix}, \begin{pmatrix} V_G & c^{\mathrm{T}}\mathcal{I}_{\theta_0}^{-1} & -c^{\mathrm{T}}h \\ \mathcal{I}_{\theta_0}^{-1}c & \mathcal{I}_{\theta_0}^{-1} & h \\ -h^{\mathrm{T}}c & -h^{\mathrm{T}} & h^{\mathrm{T}}\mathcal{I}_{\theta_0}h \end{pmatrix}\right\}, \tag{F2}$$

in distribution, as $n \to \infty$. Then, by Le Cam's third lemma,[30] $n^{-1/2}G_n(\beta_0) \to \mathcal{N}\left(0, V_G - c^{\mathrm{T}}\mathcal{I}_{\theta_0}^{-1}c\right)$ in distribution, as $n \to \infty$.

To show (F2), denote

$$\Delta_n(\theta) = n^{-1/2}\frac{\partial}{\partial\theta}L(\theta|Z_1, \dots, Z_n) = n^{-1/2}\sum_{i=1}^{n}X_i\dot{e}(X_i^{\mathrm{T}}\theta)\frac{W_i - e(X_i^{\mathrm{T}}\theta)}{e(X_i^{\mathrm{T}}\theta)\{1-e(X_i^{\mathrm{T}}\theta)\}}.$$

Then, under P^{θ_n}:

$$\Delta_n(\theta_n) \to \mathcal{N}(0, \mathcal{I}_{\theta_0}),\ \ \mathcal{I}_{\theta_0} = \mathrm{E}\left[\frac{\dot{e}(X^{\mathrm{T}}\theta)^2XX^{\mathrm{T}}}{e(X^{\mathrm{T}}\theta)\{1-e(X^{\mathrm{T}}\theta)\}}\right],$$

in distribution, as $n \to \infty$.We also note that under P^{θ_n}:

$$\begin{aligned} n^{1/2}(\widehat{\theta}_n - \theta_n) &= \mathcal{I}_{\theta_0}^{-1}\Delta_n(\theta_n) + o_P(1), \\ \Lambda_n(\theta_0|\theta_n) &= -h^{\mathrm{T}}\Delta_n(\theta_n) - \frac{1}{2}h^{\mathrm{T}}\mathcal{I}_{\theta_0}h + o_P(1). \end{aligned} \tag{F3}$$

To show (F2), it suffices to show that

$$\begin{pmatrix} D_n(\theta_n) \\ \Delta_n(\theta_n) \end{pmatrix} \to \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_G & c^{\mathrm{T}} \\ c & \mathcal{I}_{\theta_0} \end{pmatrix} \right\}, \tag{F4}$$

in distribution, as $n \to \infty$. Toward this end, we consider the linear combination $L_n = z_1 D_n(\theta_n) + z_2^{\mathrm{T}} \Delta_n(\theta_n)$, for any z_1 and z_2. We write $L_n = \sum_{l=1}^{3n} \xi_{n,l}$, where

$$\xi_{n,l} = \begin{cases} z_1 n^{-1/2} \sum\limits_{\omega=0}^{1} \mu_H\left\{\omega, e(X_l^{\mathrm{T}}\theta_n)\right\} + z_2^{\mathrm{T}} n^{-1/2} \mathrm{E}\left\{X_l | e(X_l^{\mathrm{T}}\theta_n)\right\} \dot{e}(X_i^{\mathrm{T}}\theta_n) \frac{W_i - e(X_i^{\mathrm{T}}\theta_n)}{e(X_i^{\mathrm{T}}\theta_n)\{1-e(X_i^{\mathrm{T}}\theta_n)\}}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } 1 \le l \le n; \\ z_1 n^{-1/2} \sum\limits_{\omega=0}^{1} I(W_{l-n} = \omega)\{1 + k_{l-n}/M\}\left[\mu_H(\omega, X_{l-n}) - \mu_H\left\{\omega, e(X_{l-n}^{\mathrm{T}}\theta_n)\right\}\right] \\ + z_2^{\mathrm{T}} n^{-1/2} \left(X_{l-n} - \mathrm{E}\left\{X_{l-n} | e(X_{l-n}^{\mathrm{T}}\theta_n)\right\}\right) \dot{e}(X_{l-n}^{\mathrm{T}}\theta_n) \frac{W_{l-n} - e(X_{l-n}^{\mathrm{T}}\theta_n)}{e(X_{l-n}^{\mathrm{T}}\theta_n)\{1-e(X_{l-n}^{\mathrm{T}}\theta_n)\}}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } n+1 \le l \le 2n; \\ z_1 n^{-1/2} \sum\limits_{\omega=0}^{1} I(W_{l-2n} = \omega)\{1 + k_{l-2n}/M\}\{H_{l-2n}(W_{l-2n}) - \mu_H(\omega, X_{l-2n})\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } 2n+1 \le l \le 3n. \end{cases}$$

Consider $\sigma-$fields

$$\mathcal{F}_{n,l} = \begin{cases} \left\{W_1, \ldots, W_l, X_1^{\mathrm{T}}\theta_n, \ldots, X_l^{\mathrm{T}}\theta_n\right\}, & 1 \le l \le n, \\ \left\{W_1, \ldots, W_n, X_1^{\mathrm{T}}\theta_n, \ldots, X_n^{\mathrm{T}}\theta_n, X_1, \ldots, X_{l-n}\right\}, & n+1 \le l \le 2n, \\ \{W_1, \ldots, W_n, X_1, \ldots, X_n, H_1(W_1), \ldots, H_{l-2n}(W_{l-2n})\}, & 2n+1 \le l \le 3n. \end{cases}$$

Then, $\left\{\sum_{j=1}^{l} \xi_{n,j}, F_{n,i}, 1 \le l \le 3n\right\}$ is a martingale for each $n \ge 1$. Under P^{θ_n},

$$L_n \to \mathcal{N}(0, \sigma_{L,1}^2 + \sigma_{L,2}^2 + \sigma_{L,3}^2),$$

in distribution, as $n \to \infty$, where

$$\sigma_{L,1}^2 = z_1^2 E\left([\mu_H\{0, e(X)\} + \mu_H\{1, e(X)\}]^2\right) + z_2^{\mathrm{T}} \mathrm{E}\left[\mathrm{E}\{X|e(X)\}\mathrm{E}\left\{X^{\mathrm{T}}|e(X)\right\} \frac{\dot{e}(X^{\mathrm{T}}\theta_0)^2}{e(X)\{1-e(X)\}}\right] z_2,$$

$$\begin{aligned} \sigma_{L,2}^2 &= z_1^2 \sum_{\omega=0}^{1} \mathrm{E}\left[\mathrm{var}\{\mu_H(\omega, X)|e(X)\}\left\{\frac{2M+1}{2M}\frac{1}{p(\omega|X^{\mathrm{T}}\theta_0)} - \frac{1}{2M}p(\omega|X^{\mathrm{T}}\theta_0)\right\}\right] \\ &\quad + 2z_2^{\mathrm{T}} \mathrm{E}\left[\frac{\mathrm{cov}\{X, \mu_H(1, X)|e(X)\}\dot{e}(X^{\mathrm{T}}\theta_0)}{e(X)} - \frac{\mathrm{cov}\{X, \mu_H(0, X)|e(X)\}\dot{e}(X^{\mathrm{T}}\theta_0)}{1-e(X)}\right] z_1 \\ &\quad + z_2^{\mathrm{T}} \mathrm{E}\left[\mathrm{var}\{X|e(X)\}\frac{\dot{e}(X^{\mathrm{T}}\theta_0)^2}{e(X)\{1-e(X)\}}\right] z_2, \end{aligned}$$

and

$$\sigma_{L,3}^2 = z_1^2 \sum_{\omega=0}^{1} \mathrm{E}\left[\sigma_H^2(\omega, X)|e(X)\right]\left\{\frac{2M+1}{2M}\frac{1}{p(\omega|X)} + \frac{1}{2M}p(\omega|X)\right\}.$$

Then, $\sigma_{L,1}^2 + \sigma_{L,2}^2 + \sigma_{L,3}^2 = z_1^2 V_G + z_2^{\mathrm{T}} \mathcal{I}_{\theta_0}^{-1} z_2 + 2z_2^{\mathrm{T}} c z_1$. Thus, under P^{θ_n}, (F4) follows. This completes the proof for Theorem 2. ■

Algorithm 1. For generating $T^{(0)}$ and $T^{(1)}$ that are congenial with Model (1)

```
Step 1. Generate $T^{(0)}$ from $S^{(0)}(t) = \exp(-\lambda_0 t)$, where $\lambda_0 = 6$ for $\beta_0 = 0$ and 0.5, and $\lambda_0 = 15$ for $\beta_0 = -0.5$.
Step 2. Generate $u$ from Unif[0, 1], solve
        $\left\{\prod_{k=1}^{6}\left(1-\frac{\eta_k t}{\lambda_k}\right)\right\}\exp\left[\left\{X^{\mathrm{T}}\eta-\lambda_0\exp(\beta_0)\right\}t\right]-1+u=0$
    for $t$, where $\eta_1 = \cdots = \eta_6 = -2$ and $\eta = (\eta_1, \ldots, \eta_6)^{\mathrm{T}}$. Let $T^{(1)}$ be the solution $t$.
```

APPENDIX G. SIMULATION

G.1 The data-generating algorithm

We describe the algorithm for generating $T^{(0)}$ and $T^{(1)}$ that are congenial with Model (1) as follows.

By Algorithm 1, $T^{(1)}$ given X follows

$$S_{T|W,X}(t|W=1,X)=\left\{\prod_{k=1}^{6}\left(1-\frac{\eta_k t}{\lambda_k}\right)\right\}\exp\left[\left\{X^{\mathrm{T}}\eta-\lambda_0\exp(\beta_0)\right\}t\right].$$

Under our parameter specification, we have (i) $S_{T|W,X}(t=0|W=1,X=x)=1$, (ii) $S_{T|W,X}(t=\tau|W=1,X=x)=0$, because of $X^{\mathrm{T}}\boldsymbol{\eta}-\lambda_0\exp(\beta_0)\le 0$, and (iii) $\mathrm{d}S_{T|W,X}(t|W=1,X=x)/\mathrm{d}t\le 0$, because of $\sum_{i=1}^{6}\lambda_i^{-1}\le\min(\lambda_0,\lambda_0 e^{\beta_0})$.

The marginal distribution of $T^{(1)}$ is

$$\begin{aligned}S^{(1)}(t) &= \int S_{T|W,X}(t|W=1,X=x)\mathrm{d}F(x)\\ &= \int\left\{\prod_{k=1}^{6}\left(1-\frac{\eta_k t}{\lambda_k}\right)\right\}\exp\left[\left\{X^{\mathrm{T}}\eta-\lambda_0\exp(\beta_0)\right\}t\right]\mathrm{d}F(x)\\ &= \left\{S^{(0)}(t)\right\}^{\exp(\beta_0)}.\end{aligned}$$

Therefore, the marginal distributions of $\{T^{(0)},T^{(1)}\}$ satisfy $S^{(1)}(t)=\left\{S^{(0)}(t)\right\}^{\exp(\beta_0)}$.

G.2 Illustration of propensity score distributions with weak, medium, and strong covariate overlap

This section demonstrates the mentioned simulation settings of weak, medium, strong covariate overlap between the two treatment groups. Figure G1 shows density curves for the true propensity scores of the two treatment groups are presented in dashed lines of $W=0$ and in solid lines of $W=1$.

G.3 Additional simulation results

We provide additional simulation results in Tables G1–G4.

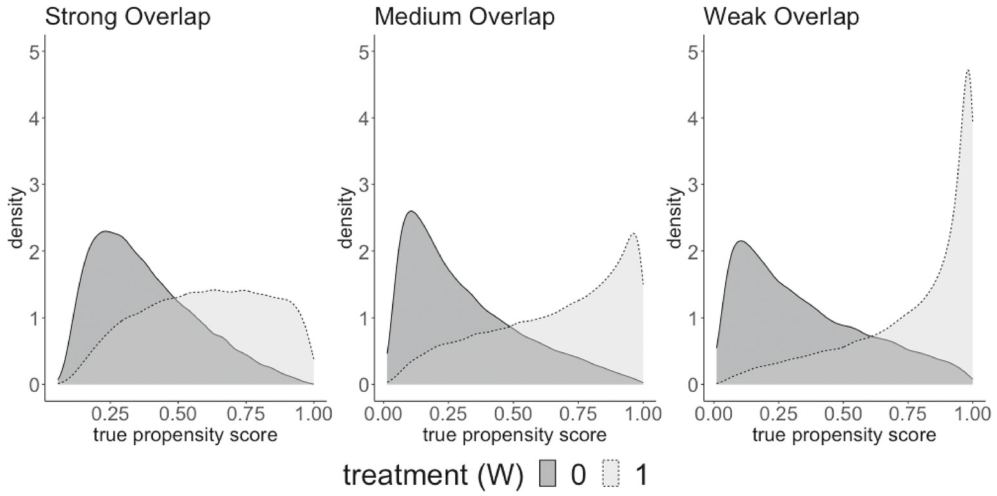

FIGURE G1 Plots of the true propensity score distribution between treatment groups $W=0$ and $W=1$ for strong, medium and weak covariate overlap.

TABLE G1 Simulation results: Bias ($\times 10^2$) and variance ($\times 10^3$) of the point estimator of β_0, coverage (%) of 95% confidence intervals based on 1,000 M onte Carlo samples with true $\beta_0 = 0$ when propensity score model is misspecified.

		Bias	Var	VE	CR	Bias	Var	VE	CR	Bias	Var	VE	CR
Level of covariate overlap		**Strong**				**Medium**				**Weak**			
Scenario (i) Correct specification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		54.6	4.9	5.1	0.0%	70.6	5.2	5.3	0.0%	80.9	5.2	5.7	0.0%
$\widehat{\beta}_{\text{ipw}}$		0.3	6.1	8.3	97.8%	1.0	12.5	14.6	95.6%	4.1	22.2	18.9	91.2%
$\widehat{\beta}_{\text{aipw}}$		0.1	5.3	6.0	95.9%	0.2	9.1	12.0	96.1%	2.3	31.3	40.6	96.5%
$\widehat{\beta}_{\text{x.m}}$		6.9	5.9	8.8	92.7%	11.4	8.2	11.8	84.0%	17.7	8.2	11.7	64.4%
$\widehat{\beta}_{\text{psm.0}}$		0.5	6.6	10.9	98.5%	1.6	10.6	17.9	97.4%	3.7	16.1	25.4	96.1%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	0.7	6.3	11.1	98.9%	1.3	10.5	18.3	97.9%	3.9	17.2	26.0	95.8%
	asymp	0.7	6.3	6.5	95.1%	1.3	10.5	10.6	94.2%	3.9	17.2	16.4	91.3%
	naiveboot	0.7	6.3	9.8	98.8%	1.3	10.5	17.0	98.7%	3.9	17.2	25.1	96.3%
	double-rsp(5)	0.7	6.3	8.2	97.7%	1.3	10.5	13.6	97.5%	3.9	17.2	24.1	95.4%
	double-rsp(10)	0.7	6.3	8.9	98.2%	1.3	10.5	14.7	97.4%	3.9	17.2	22.2	94.5%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	1.5	5.0	8.1	99.0%	3.2	8.0	12.3	96.6%	5.1	10.1	16.3	97.0%
	asymp	1.5	5.0	5.0	94.2%	3.2	8.0	7.3	91.9%	5.1	10.1	9.2	90.6%
	naiveboot	1.5	5.0	7.6	99.0%	3.2	8.0	12.2	96.8%	5.1	10.1	16.6	97.5%
	double-rsp(5)	1.5	5.0	6.1	96.1%	3.2	8.0	9.5	95.2%	5.1	10.1	13.9	95.8%
	double-rsp(10)	1.5	5.0	6.5	97.1%	3.2	8.0	10.0	96.0%	5.1	10.1	14.4	96.2%
Scenario (ii) Misspecification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		54.6	4.9	5.1	0.0%	70.4	5.0	5.3	0.0%	80.8	5.2	5.7	0.0%
$\widehat{\beta}_{\text{ipw}}$		4.5	19.4	12.4	84.6%	3.8	42.2	19.6	79.0%	3.4	68.5	26.6	74.7%
$\widehat{\beta}_{\text{aipw}}$		17.2	121.5	72.8	73.1%	30.1	614.2	308.9	79.7%	30.0	1054.2	603.7	89.4%
$\widehat{\beta}_{\text{x.m}}$		6.9	5.9	8.8	92.7%	11.4	8.3	11.8	84.1%	17.6	8.1	11.7	64.8%
$\widehat{\beta}_{\text{psm.0}}$		0.5	6.6	10.9	98.5%	1.3	10.7	18.0	97.4%	3.3	15.6	25.4	96.5%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	5.4	6.7	10.6	95.3%	7.7	10.4	16.8	93.7%	10.9	15.1	22.8	89.6%
	asymp	5.4	6.7	6.5	88.3%	7.7	10.4	9.4	86.1%	10.9	15.1	11.0	77.4%
	naiveboot	5.4	6.7	9.5	94.6%	7.7	10.4	15.8	93.9%	10.9	15.1	21.6	90.9%
	double-rsp(5)	5.4	6.7	9.8	95.5%	7.7	10.4	16.1	93.4%	10.9	15.1	17.7	88.2%
	double-rsp(10)	5.4	6.7	11.0	95.6%	7.7	10.4	21.2	94.5%	10.9	15.1	19.8	89.6%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	6.0	5.1	7.9	93.9%	9.5	7.9	11.5	88.5%	13.1	9.9	14.7	84.9%
	asymp	6.0	5.1	4.8	85.0%	9.5	7.9	6.6	76.1%	13.1	9.9	5.1	64.3%
	naiveboot	6.0	5.1	7.5	93.6%	9.5	7.9	11.6	88.9%	13.1	9.9	15.0	86.4%
	double-rsp(5)	6.0	5.1	7.3	93.6%	9.5	7.9	12.1	89.1%	13.1	9.9	11.8	79.1%
	double-rsp(10)	6.0	5.1	8.2	94.2%	9.5	7.9	14.4	91.4%	13.1	9.9	13.9	82.5%

Note: "Var" is the variance of point estimates of β_0 across simulated datasets; "VE" is the average variance estimation for the point estimators over simulations, thus VE minus Var reflects the bias in estimated variance; "CR" is the empirical coverage rate of 95% confidence intervals. Five types of variance estimates for $\widehat{\beta}_{\text{psm}}$ were compared: "software", output from the standard software; "asymp", the proposed asymptotic variance estimation; "naiveboot", the naive nonparametric bootstrap; "double-rsp(5)", the proposed double-resampling method with five quantile strata and "double-rsp(10)", the proposed double-resampling method with ten quantile strata.

TABLE G2 Simulation results: Bias ($\times 10^2$) and variance ($\times 10^3$) of the point estimator of β_0, coverage (%) of 95% confidence intervals based on 1,000 Monte Carlo samples with true $\beta_0 = 0.5$.

		Bias	**Var**	**VE**	**CR**	**Bias**	**Var**	**VE**	**CR**	**Bias**	**Var**	**VE**	**CR**
Level of covariate overlap		**Strong**				**Medium**				**Weak**			
Scenario (i) Correct specification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		46.4	5.5	5.6	0.0%	59.7	5.8	5.8	0.0%	59.7	6.1	6.5	0.0%
$\widehat{\beta}_{\text{ipw}}$		−0.5	7.1	8.7	97.0%	0.1	14.8	15.0	94.5%	0.1	28.4	21.5	89.8%
$\widehat{\beta}_{\text{aipw}}$		−0.7	6.2	7.4	95.2%	−0.6	11.6	17.3	94.4%	−0.6	35.7	42.6	97.2%
$\widehat{\beta}_{\text{x.m}}$		6.3	7.7	8.9	90.3%	10.0	10.5	11.7	83.8%	16.8	10.4	12.4	66.0%
$\widehat{\beta}_{\text{psm.0}}$		−0.2	8.6	11.3	96.9%	0.5	14.9	18.6	95.8%	0.5	23.5	29.9	94.0%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	−0.1	8.3	11.5	97.0%	0.2	14.6	18.8	96.3%	0.2	25.3	30.6	93.2%
	asymp	−0.1	8.3	8.4	94.5%	0.2	14.6	14.5	92.9%	0.2	25.3	23.1	90.6%
	naiveboot	−0.1	8.3	11.1	97.1%	0.2	14.6	18.6	96.8%	0.2	25.3	31.7	94.6%
	double-rsp(5)	−0.1	8.3	9.6	95.9%	0.2	14.6	16.0	95.5%	0.2	25.3	29.9	95.0%
	double-rsp(10)	−0.1	8.3	10.4	96.5%	0.2	14.6	16.9	96.1%	0.2	25.3	27.5	93.3%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	0.4	6.5	8.5	97.3%	1.6	10.6	12.9	95.5%	1.6	13.7	18.9	96.5%
	asymp	0.4	6.5	6.3	93.9%	1.6	10.6	9.4	92.6%	1.6	13.7	12.5	91.9%
	naiveboot	0.4	6.5	8.9	97.9%	1.6	10.6	13.8	96.8%	1.6	13.7	20.9	97.4%
	double-rsp(5)	0.4	6.5	7.0	95.9%	1.6	10.6	10.9	94.8%	1.6	13.7	16.8	96.2%
	double-rsp(10)	0.4	6.5	7.5	96.5%	1.6	10.6	11.4	95.0%	1.6	13.7	17.5	96.5%
Scenario (ii) Misspecification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		46.4	5.5	5.6	0.0%	59.7	5.6	5.8	0.0%	72.0	6.0	6.5	0.0%
$\widehat{\beta}_{\text{ipw}}$		−0.8	27.3	15.7	91.5%	−3.5	61.3	25.5	84.7%	−1.9	90.4	35.2	79.4%
$\widehat{\beta}_{\text{aipw}}$		11.6	111.1	73.3	90.4%	20.1	607.2	308.4	93.7%	10.5	778.5	538.8	94.4%
$\widehat{\beta}_{\text{x.m}}$		6.8	8.1	9.7	90.5%	10.4	11.7	13.4	82.9%	18.1	13.5	15.1	65.8%
$\widehat{\beta}_{\text{psm.0}}$		−0.3	8.8	11.2	96.8%	0.3	14.4	18.6	96.3%	2.2	23.2	30.1	94.5%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	4.3	8.7	10.9	93.9%	6.3	13.4	17.1	92.3%	9.7	20.7	26.2	89.2%
	asymp	4.3	8.7	8.3	90.9%	6.3	13.4	12.3	88.2%	9.7	20.7	14.5	81.1%
	naiveboot	4.3	8.7	10.7	93.9%	6.3	13.4	17.0	92.7%	9.7	20.7	26.9	91.6%
	double-rsp(5)	4.3	8.7	11.0	94.5%	6.3	13.4	17.7	93.9%	9.7	20.7	21.0	88.9%
	double-rsp(10)	4.3	8.7	12.1	95.2%	6.3	13.4	22.0	94.7%	9.7	20.7	23.1	90.1%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	4.4	6.5	8.3	94.0%	7.4	9.9	12.0	89.7%	11.6	12.8	16.9	89.3%
	asymp	4.4	6.5	6.0	89.0%	7.4	9.9	8.1	82.0%	11.6	12.8	6.3	74.7%
	naiveboot	4.4	6.5	8.7	95.0%	7.4	9.9	12.9	91.8%	11.6	12.8	18.6	91.1%
	double-rsp(5)	4.4	6.5	8.1	93.4%	7.4	9.9	13.0	90.5%	11.6	12.8	13.8	84.4%
	double-rsp(10)	4.4	6.5	8.9	94.1%	7.4	9.9	14.8	91.7%	11.6	12.8	16.1	86.0%

Note: "Var" is the variance of point estimates of β_0 across simulated datasets; "VE" is the average variance estimation for the point estimators over simulations, thus VE minus Var reflects the bias in estimated variance; "CR" is the empirical coverage rate of 95% confidence intervals. Five types of variance estimates for $\widehat{\beta}_{\text{psm}}$ were compared: "software", output from the standard software; "asymp", the proposed asymptotic variance estimation; "naiveboot", the naive nonparametric bootstrap; "double-rsp(5)", the proposed double-resampling method with five quantile strata and "double-rsp(10)", the proposed double-resampling method with ten quantile strata.

TABLE G3 Simulation results: Bias ($\times 10^2$) and variance ($\times 10^3$) of the point estimator of β_0, coverage (%) of 95% confidence intervals based on 1,000 Monte Carlo samples with true $\beta_0 = -0.5$.

		Bias	Var	VE	CR	Bias	Var	VE	CR	Bias	Var	VE	CR
Level of covariate overlap		**Strong**				**Medium**				**Weak**			
Scenario (i) Correct specification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		33.7	4.3	4.8	0.1%	42.8	4.6	4.8	0.0%	44.9	4.7	5.0	0.0%
$\widehat{\beta}_{\text{ipw}}$		0.8	6.1	7.3	97.1%	1.0	11.5	11.6	95.3%	2.7	17.8	15.0	92.9%
$\widehat{\beta}_{\text{aipw}}$		0.6	5.9	6.5	96.7%	0.6	10.7	11.5	94.5%	2.6	55.3	48.1	95.7%
$\widehat{\beta}_{\text{x.m}}$		4.5	6.7	8.2	94.0%	7.3	8.9	10.3	89.4%	10.4	8.6	10.2	83.3%
$\widehat{\beta}_{\text{psm.0}}$		0.9	8.2	9.8	96.5%	1.7	12.8	14.9	96.5%	3.2	20.2	20.7	95.1%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	1.0	8.0	10.0	97.5%	1.5	13.1	15.1	96.3%	3.3	21.5	21.1	94.9%
	asymp	1.0	8.0	8.4	95.0%	1.5	13.1	12.6	93.8%	3.3	21.5	19.2	91.9%
	naiveboot	1.0	8.0	9.3	96.6%	1.5	13.1	14.0	96.1%	3.3	21.5	19.4	94.0%
	double-rsp(5)	1.0	8.0	9.2	95.8%	1.5	13.1	15.0	95.5%	3.3	21.5	27.1	93.2%
	double-rsp(10)	1.0	8.0	9.8	96.9%	1.5	13.1	15.2	95.7%	3.3	21.5	25.0	92.4%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	1.5	6.1	7.5	96.9%	2.5	9.8	10.9	95.5%	3.4	12.0	14.1	95.4%
	asymp	1.5	6.1	6.3	95.0%	2.5	9.8	8.8	91.6%	3.4	12.0	10.8	90.9%
	naiveboot	1.5	6.1	7.8	97.4%	2.5	9.8	11.0	95.5%	3.4	12.0	14.2	95.3%
	double-rsp(5)	1.5	6.1	7.0	95.7%	2.5	9.8	10.4	94.4%	3.4	12.0	16.3	95.2%
	double-rsp(10)	1.5	6.1	7.2	95.8%	2.5	9.8	10.6	95.0%	3.4	12.0	16.2	95.6%
Scenario (ii) Misspecification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		33.7	4.4	4.8	0.2%	42.7	4.2	4.8	0.0%	44.8	4.6	5.0	0.0%
$\widehat{\beta}_{\text{ipw}}$		3.7	15.0	9.9	89.1%	4.5	34.7	16.3	84.2%	4.6	55.1	22.9	82.7%
$\widehat{\beta}_{\text{aipw}}$		14.5	184.1	97.1	87.9%	31.8	811.4	371.4	89.6%	37.6	1223.2	666.2	95.2%
$\widehat{\beta}_{\text{x.m}}$		4.3	7.0	8.2	93.9%	7.2	8.5	10.4	89.6%	10.3	8.5	10.2	84.1%
$\widehat{\beta}_{\text{psm.0}}$		0.6	8.4	9.9	96.5%	1.3	12.6	15.1	96.7%	2.4	20.4	20.7	95.4%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	4.2	8.2	9.6	93.7%	5.6	10.9	14.0	93.6%	6.9	17.4	18.7	92.2%
	asymp	4.2	8.2	7.9	90.9%	5.6	10.9	10.8	91.3%	6.9	17.4	12.8	84.8%
	naiveboot	4.2	8.2	9.0	92.7%	5.6	10.9	13.1	93.6%	6.9	17.4	17.4	92.0%
	double-rsp(5)	4.2	8.2	9.8	93.7%	5.6	10.9	15.2	94.6%	6.9	17.4	17.7	91.1%
	double-rsp(10)	4.2	8.2	10.3	94.4%	5.6	10.9	18.0	95.2%	6.9	17.4	18.5	91.1%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	4.2	6.1	7.4	94.4%	6.5	8.3	10.2	92.1%	8.1	11.3	13.0	91.7%
	asymp	4.2	6.1	5.9	91.0%	6.5	8.3	7.5	86.8%	8.1	11.3	5.0	82.8%
	naiveboot	4.2	6.1	7.6	95.2%	6.5	8.3	10.4	92.6%	8.1	11.3	13.2	92.1%
	double-rsp(5)	4.2	6.1	7.3	94.2%	6.5	8.3	11.2	93.2%	8.1	11.3	11.8	89.4%
	double-rsp(10)	4.2	6.1	7.7	94.9%	6.5	8.3	12.3	94.5%	8.1	11.3	12.9	90.1%

Note: "Var" is the variance of point estimates of β_0 across simulated datasets; "VE" is the average variance estimation for the point estimators over simulations, thus VE minus Var reflects the bias in estimated variance; "CR" is the empirical coverage rate of 95% confidence intervals. Five types of variance estimates for $\widehat{\beta}_{\text{psm}}$ were compared: "software", output from the standard software; "asymp", the proposed asymptotic variance estimation; "naiveboot", the naive nonparametric bootstrap; "double-rsp(5)", the proposed double-resampling method with five quantile strata and "double-rsp(10)", the proposed double-resampling method with ten quantile strata.

TABLE G4 Simulation results for perfect overlap: Bias ($\times 10^2$) and variance ($\times 10^3$) of the point estimator of β_0, coverage (%) of 95% confidence intervals based on 1,000 Monte Carlo samples where the true propensity score equals 0.5 for each subject, that is, perfect overlap.

		Bias	**Var**	**VE**	**CR**	**Bias**	**Var**	**VE**	**CR**	**Bias**	**Var**	**VE**	**CR**
		$\boldsymbol{\beta_0 = 0}$				$\boldsymbol{\beta_0 = 0.5}$				$\boldsymbol{\beta_0 = -0.5}$			
Scenario (i) Correct specification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		−0.1	4.6	4.8	95.2%	0.7	4.7	4.8	94.7%	0.1	4.5	4.6	95.7%
$\widehat{\beta}_{\text{ipw}}$		0.0	3.6	4.8	98.0%	0.8	3.9	4.8	96.9%	0.1	4.1	4.6	97.1%
$\widehat{\beta}_{\text{aipw}}$		0.0	3.3	3.5	95.7%	1.0	3.8	3.9	94.4%	0.1	3.9	4.2	96.2%
$\widehat{\beta}_{\text{x.m}}$		0.0	4.5	6.3	98.1%	−0.3	5.1	6.3	97.0%	0.9	5.4	6.0	95.5%
$\widehat{\beta}_{\text{psm.0}}$		−0.1	6.0	6.0	95.6%	0.6	6.0	6.0	95.1%	0.1	5.6	5.7	94.7%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	0.3	5.5	6.6	96.5%	1.1	6.0	6.6	95.5%	0.4	6.2	6.3	95.4%
	asymp	0.3	5.5	5.3	94.3%	1.1	6.0	5.7	94.0%	0.4	6.2	5.9	94.1%
	naiveboot	0.3	5.5	5.8	95.6%	1.1	6.0	6.1	95.0%	0.4	6.2	6.1	95.2%
	double-rsp(5)	0.3	5.5	7.0	96.9%	1.1	6.0	6.9	96.2%	0.4	6.2	6.6	95.7%
	double-rsp(10)	0.3	5.5	7.1	97.4%	1.1	6.0	7.0	96.6%	0.4	6.2	6.7	95.8%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	0.0	4.0	5.2	96.8%	0.8	4.4	5.2	96.1%	0.2	4.6	5.0	96.1%
	asymp	0.0	4.0	4.0	95.0%	0.8	4.4	4.3	94.3%	0.2	4.6	4.5	94.9%
	naiveboot	0.0	4.0	4.9	96.6%	0.8	4.4	5.1	96.1%	0.2	4.6	5.3	96.7%
	double-rsp(5)	0.0	4.0	5.4	97.4%	0.8	4.4	5.4	96.2%	0.2	4.6	5.2	96.1%
	double-rsp(10)	0.0	4.0	5.5	97.4%	0.8	4.4	5.5	96.5%	0.2	4.6	5.2	96.3%
Scenario (ii) Misspecification of the propensity score model													
$\widehat{\beta}_{\text{nai}}$		−0.1	4.6	4.8	95.2%	0.7	4.7	4.8	94.7%	0.1	4.5	4.6	95.7%
$\widehat{\beta}_{\text{ipw}}$		−0.1	3.9	4.8	97.4%	0.7	4.2	4.8	96.8%	0.1	4.2	4.6	96.5%
$\widehat{\beta}_{\text{aipw}}$		0.0	3.7	3.9	95.9%	0.9	4.0	4.2	94.9%	0.1	4.0	4.3	96.1%
$\widehat{\beta}_{\text{x.m}}$		0.0	4.8	6.5	97.9%	−0.6	5.4	6.5	97.5%	1.3	5.8	6.2	95.8%
$\widehat{\beta}_{\text{psm.0}}$		−0.1	6.0	6.0	95.6%	0.6	6.0	6.0	95.1%	0.1	5.6	5.7	94.7%
$\widehat{\beta}_{\text{psm}}$ (M=1)	software	0.2	5.5	6.6	97.5%	0.9	5.9	6.6	96.7%	0.3	6.0	6.3	95.4%
	asymp	0.2	5.5	5.5	95.8%	0.9	5.9	5.8	95.2%	0.3	6.0	6.0	94.7%
	naiveboot	0.2	5.5	5.9	96.7%	0.9	5.9	6.1	96.1%	0.3	6.0	6.1	95.1%
	double-rsp(5)	0.2	5.5	6.9	98.2%	0.9	5.9	6.9	96.8%	0.3	6.0	6.6	96.0%
	double-rsp(10)	0.2	5.5	7.0	98.0%	0.9	5.9	7.0	96.8%	0.3	6.0	6.7	96.2%
$\widehat{\beta}_{\text{psm}}$ (M=5)	software	0.0	4.1	5.2	96.8%	0.7	4.5	5.2	96.4%	0.2	4.5	5.0	96.3%
	asymp	0.0	4.1	4.3	95.3%	0.7	4.5	4.5	94.7%	0.2	4.5	4.6	95.9%
	naiveboot	0.0	4.1	5.2	96.6%	0.7	4.5	5.4	96.8%	0.2	4.5	5.4	97.2%
	double-rsp(5)	0.0	4.1	5.4	97.6%	0.7	4.5	5.4	96.5%	0.2	4.5	5.2	96.8%
	double-rsp(10)	0.0	4.1	5.5	97.5%	0.7	4.5	5.4	96.5%	0.2	4.5	5.2	96.8%

Note: "Var" is the variance of point estimates of β_0 across simulated datasets; "VE" is the average variance estimation for the point estimators over simulations, thus VE minus Var reflects the bias in estimated variance; "CR" is the empirical coverage rate of 95% confidence intervals. Five types of variance estimates for $\widehat{\beta}_{\text{psm}}$ were compared: "software", output from the standard software; "asymp", the proposed asymptotic variance estimation; "naiveboot", the naive nonparametric bootstrap; "double-rsp(5)", the proposed double-resampling method with five quantile strata and "double-rsp(10)", the proposed double-resampling method with ten quantile strata.