Causal Inference Methods for Comparing Multiple Treatments using Data from Large Insurance Claims Databases

by

Youfei Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2022

Doctoral Committee:

Professor Bhramar Mukherjee, Co-Chair
Professor Min Zhang, Co-Chair
Professor Andrew M. Ryan
Assistant Professor Zhenke Wu

Youfei Yu
youfeiyu@umich.edu
ORCID iD: 0000-0002-4986-848X

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my co-advisors, Prof. Bhramar Mukherjee and Prof. Min Zhang for their indispensable advice, selfless sharing of knowledge, continuous support, and patience during my PhD study. I would also like to thank Prof. Andrew Ryan and Prof. Zhenke Wu for being willing to serve on my dissertation committee and provide me with constructive feedback on my research work. I also appreciate all the support I received from the faculty and staff members at Department of Biostatistics, University of Michigan. Finally, I would like to thank my family and friends, who are always being encouraging and supportive.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Large healthcare databases used primarily for billing and payments, such as electronic health records and insurance claims data, have been increasingly used to conduct comparative effectiveness research (CER) that characterize multiple treatment/intervention strategies for a particular clinical condition. Estimation of causal treatment effects using such observational data is prone to bias due to confounders related to both treatment and outcome. Another potential source of bias is censoring, which occurs when patients drop out of the system or the reporting period ends (starts) before (after) the occurrence of the event of interest.

Our study is motivated by analysis embedded within the OptumInsight Clinformatics Data Mart, a database that consists of medical and pharmacy claims from a large national private health insurance network with over 83 million insured unique individuals. Our analytic cohort consists of around 700,000 patients with a claim for prostate cancer diagnosis recorded between 2001-2019. Interest is in assessing the effects of four common therapies for castration-resistant advanced-stage prostate cancer, with the adverse outcome being hospitalization and/or admission to the emergency room within a short time window of treatment initiation.

In Chapter 1, we consider CER from observational data with two or more treatments. Methods based on propensity scores are routinely used to correct for confounding biases. A large fraction of propensity score methods in the current literature consider the case of either two treatments or continuous outcome. There has been extensive literature with multiple treatment or binary outcome, but interest often lies in the intersection, for which the literature is still evolving. The contribution of this Chapter is to focus on this intersection and compare across existing methods, some of which are fairly recent. We assess the relative performance of these methods through a set of simulation studies and provide recommendations for the practitioners.

In Chapter 2, we take censoring into account and propose a method that directly models the binary outcome using logistic regression, with confounding and censoring properly accounted for by weighting. We call the method inverse probability weighted regression-based estimator that accounts for censoring, or CIPWR. The risk of event occurrence (and therefore the average treatment effect) is estimated based on standardization, which averages the outcome predictions obtained from the logistic regression model across all subjects. CIPWR estimates the average treatment effects by averaging the predicted outcomes obtained from a logistic regression model that is fitted

using a weighted score function. The CIPWR estimator has a double robustness property such that estimation consistency can be achieved when either the model for the outcome or the models for both treatment and censoring are correctly specified. We establish the asymptotic properties of the CIPWR estimator for conducting inference, and compare its finite sample performance with that of several alternatives through simulation studies. The methods under comparison are applied to a cohort of prostate cancer patients from an insurance claims database for comparing the adverse effects of four candidate drugs for advanced stage prostate cancer.

In Chapter 3, we consider a setting where a massive collection of candidate covariates are available in the data and are potentially related to both treatment and outcome. In addition, the treatment generating model possibly involves nonlinearity and nonadditivity. In this setting, a key challenge is to identify variables to be included in the propensity score model from a high-dimensional set of measured covariates to remove the bias. As in Chapter 2, we also aim to account for censoring at the same time, where the bias due to censoring is controlled for by applying the inverse probability of remaining uncensored as weights to the outcome. We focus on estimating the treatment effects on a binary outcome (that is possibly censored) among multiple treatment groups. We examine an ensemble of data-driven methods that select the variables to be adjusted for in the treatment model, including penalized regression and modern machine learning tools based on classification and regression trees (CART). We estimate the causal effects of treatment using the inverse probability weighting (IPW) estimator. We allow the associations between the outcome and the covariates to contribute to the variable selection process, and show through simulation studies that leveraging the information about the outcome-covariate relationship when modeling the propensity scores can improve statistical efficiency and robustness against model misspecification of propensity score-based methods, such as IPW. The improvement of precision in the estimates of treatment effects is also observed in our application to the prostate cancer data.

CHAPTER 1

A Comparison of Parametric Propensity Score-Based Methods for Causal Inference with Multiple Treatments and a Binary Outcome

1.1 Introduction

Comparative effectiveness research (CER) assesses alternative interventions for a particular clinical condition [1]. Randomized clinical trials are the gold standard for CER, but real-world evidence when drugs are released into the market is increasingly being used to make health care decisions [2]. CER for such observational data requires statistical methods for causal inference that control for confounding variables. The current literature on these methods largely focuses on two treatments and continuous outcomes [3, 4], but often interest lies in comparing more than two treatments and outcomes are binary, for example, the occurrence of an event [5]. We compare here causal inference methods when the outcome is binary and there are more than two treatments.

Our motivating study concerns men who used at least one of four commonly prescribed drugs (docetaxel, abiraterone, enzalutamide, sipuleucel-T) as a first-line therapy for metastatic castration-resistant prostate cancer (mCRPC). These four drugs have increased survival for mCRPC patients in individual studies [6, 7, 8, 9]. We are interested in evaluating the possible adverse effects of these drugs, by comparing patients' risk of experiencing at least one emergency room visit shortly after treatment initiation. Data are from the Optum Clinformatics Data Mart, a national private health insurance network.

In observational studies, the estimation of causal effects is prone to bias due to confounders related to both treatment and outcome. Methods to correct for this bias can be classified into two broad categories. The traditional approach is to model the multiple regression of the outcome on the treatment and measured potential confounders. This approach is vulnerable to misspecification of the regression model. An alternative approach is to model the propensity score, defined as the probability of being assigned to the treatment given a set of potential confounders. The treatment

effect is then estimated by matching [10, 11], weighting [12, 13, 14], stratification [3, 10, 15], or regression [16, 17] on the estimated propensity scores. This method was introduced by Rosenbaum and Rubin [10], who showed that propensity scores have a balancing property, such that the conditional distribution of the potential confounders given the balancing scores are the same for treated and control. This property implies that propensity score methods provide some protection against misspecification of the outcome models. However, propensity score models are still required to be correctly specified.

Methods based on the propensity score were initially developed for comparing two treatments [e.g., 10, 11, 15, 18, 19, 20], and then extended to the case of more than two treatment groups using generalized propensity scores (GPS) [21, 22], which consist of the vector of conditional probabilities of being assigned to each treatment. However, propensity score methods become more complex as the number of compared treatments increases, and the relative performance of propensity score methods is much less studied than the two-treatment group case [e.g., 13, 23, 24, 25, 26].

Matching is the most common propensity score method for two treatments [27]. There are a variety of matching algorithms (e.g., nearest neighbor matching, full matching) corresponding to different causal estimands [28]. With more than two treatments, the number of subjects that can be matched goes down as the number of treatment groups increases, and the complexity of the matching algorithm increases. Propensity score matching methods for multiple treatment comparison built upon the framework of conventional matching methods include common-referent matching [29] and "within-trio" matching [30]. In general, the study population of these methods consists of those receiving the reference treatment. In contrast, the method of matching with replacement [24, 25, 31] yields inferences for the overall population (i.e., population of those receiving any of the treatment under comparison).

Abadie and Imbens [31] proposed a matching procedure that uses a fixed number of matches and allows each unit to be matched more than once, a method we label AI-type matching. They derived the large sample properties of the AI-type matching estimators and proposed an estimator for the asymptotic variance. Yang et al. [24] extended AI-type matching procedures to the multiple treatment case by matching on a scalar function of the GPS. Applications of these methods to real studies appear limited [32]. More common applied approaches include combining therapies with similar features as a single group and then applying propensity score matching developed for binary treatment [33, 34, 35], or conducting pairwise analysis, ignoring individuals not assigned to one of the treatment pair being compared [36].

Propensity score weighting methods are more easily extended to the multiple treatment setting. The asymptotic distributions of the weighting-based estimators can be characterized using the theory of M-estimation [37], which yields estimated standard errors that incorporate the uncertainty

associated with the estimation of propensity scores. A common weighting scheme is to weight units in one group by their inverse probability of being in that group (IPW). Evaluations of IPW are mainly confined to the two treatment setting, and suggest that the estimator is sensitive to extreme weights and can have high variability [3, 24, 38].

An important extension of IPW is the augmented inverse probability weighting (AIPW), where the IPW estimator is augmented using predictions from an outcome regression model. To implement AIPW method in a multiple treatment setting, one can first obtain the estimated GPS, possibly from a multinomial logistic regression model, and then the predicted outcomes for each treatment group from outcome models that describe the conditional expectation of the outcome variable given measured covariates and treatment status. The resulting estimator is known as having a double robustness (DR) property such that the estimator remains consistent as long as either the propensity score model or the outcome model is correctly specified. AIPW estimator is asymptotically efficient within a broad class of estimators that includes the IPW estimator [39]. Lunceford and Davidian [3] reviewed the theoretical properties of IPW, AIPW, and several other propensity score weighting estimators in the context of two treatments and continuous outcome. Simulation studies indicated that weighting-based methods with correct propensity score modeling produced approximately unbiased point estimates, and AIPW was more precise than IPW for sample sizes as small as 1000.

Other hybrid methods include outcome regression models weighted by inverse probability [40] and post-matching sample adjusted using overlap weights [14]. A multiple imputation-based approach called penalized spline of propensity methods for treatment comparison (PENCOMP), proposed by Zhou et al. [17], estimates causal effects by imputing the missing potential outcomes from a regression model for the outcome that incorporates splines of propensity scores as predictors. PENCOMP was developed and evaluated in the context of two treatments and a continuous outcome, but is extended here to the case with multiple treatments and binary outcome.

Studies of comparative effectiveness with continuous outcomes typically report an estimate of the Average Treatment Effect (ATE), which is the difference in average outcome if individuals were all assigned the treatment and the average outcome if all the individuals were assigned the comparator treatment [41]. In this paper, we measure treatment effectiveness by the risk difference, a measure of the ATE for a binary outcome, where the average outcome is the proportion of successes.

In Sections 1.2, 1.3, and 1.4, we provide more detail on several of these methods. In Section 1.5, we describe simulation studies that compare the finite sample performance of these methods. In Section 1.6, we apply the methods to estimate comparative effectiveness of four common therapies for mCRPC patients, using claims data from the Optum Clinformatics Data Mart, with the outcome being admission to the emergency room within a short time window of treatment initiation.

Conclusions and topics for future research are given in Section 1.7.

1.2 Notation and Setup

1.2.1 Estimands of Interest

Suppose an observational study of J treatments is carried out on a sample of n individuals from a target population. For individual i, let $Y_i(z)$, $z = 1, \cdots, J$, denote the potential outcome if assigned treatment z, Z_i denote the treatment actually assigned, and $\boldsymbol{X}_i$ denote a set of baseline covariates. The hypothetical complete data consist of $\{\boldsymbol{X}_j, Z_i, Y_i(1), \cdots, Y_i(J), i = 1, \cdots, n\}$, the observed data consist of $\{\boldsymbol{X}_i, Z_i, Y_i(Z_i), i = 1, \cdots, n\}$, and the outcomes $\{Y_i(z), z \neq Z_i\}$ are missing, as in the potential outcome framework [41]. For each pair (z, z') of treatments, we seek to estimate the average treatment effect (ATE),

$$\tau_{ATE}(z, z') = E[Y(z') - Y(z)],$$

where the expectation is over the population of interest. When Y is binary, the ATE is the risk difference $\tau_{ATE}(z, z') = pr\{Y(z') = 1\} - pr\{Y(z) = 1\}$.

In addition to risk difference, one can also consider estimands on multiplicative scale for treatment group z, such as causal odds ratio $pr\{Y(z) = 1\}pr\{Y(J) = 0\}/pr\{Y(z) = 0\}pr\{Y(J) = 1\}$ and relative risk $pr\{Y(z) = 1\}/pr\{Y(z) = 0\}$, where J is the reference group. We focus on the additive scale primarily for two reasons. The first is that the ratio-scale estimands can be derived using the counterfactual probabilities we estimate in each treatment group. The second is that the additive scale is more relevant to evaluating interventions as it directly yields the number of cases/deaths prevented by using one treatment as opposed to another.

For a study with binary treatments, one quantity of possible interest is the average treatment effect on the treated (ATT), which refers to the treatment effect averaged across the group of individuals who received the treatment. When there are more than two treatment groups under comparison, one common way to define the ATT is to specify a reference group ($Z = z^*$), possibly the one with the smallest sample size or of the greatest clinical interest [25]. The ATT is defined as $\tau_{ATT}(z, z') = E[Y(z') - Y(z)|Z = z^*]$, where z^* is not necessarily the same as z or z'. This implies that one can compare any treatment pair (z, z') on any subpopulation, in this case, those who received treatment z^*.

A more general form of ATE is the weighted average treatment effect [14, 42]:

$$\tau^*_{ATE}(z, z') = \frac{\int w(\boldsymbol{x})E[Y(z') - Y(z)|\boldsymbol{X} = \boldsymbol{x}]f(\boldsymbol{x})d\boldsymbol{x}}{\int w(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x}},$$

where $f(\boldsymbol{x})$ is the density function of the covariates $\boldsymbol{X}$ and $w(\boldsymbol{x})$ is a prespecified function of $\boldsymbol{x}$. Different choices of $w(\cdot)$ yield the ATE for different target populations, as discussed further in Section 1.4.2.

Note that $\tau_{ATE}(z, z')$ is equivalent to $\tau^*_{ATE}(z, z')$ if $w(\boldsymbol{x}) = 1$, or if the treatment effect conditional on $\boldsymbol{x}$, $E[Y(z') - Y(z)|\boldsymbol{X} = \boldsymbol{x}]$, is the same for all $\boldsymbol{x}$ (i.e. homogeneous), an unlikely event. When the treatment effect is heterogeneous, the ATE should always be defined with respect to a clearly specified study population.

1.2.2 Assumptions

In an observational study where treatment is not randomly assigned, valid inferences for the ATE require some standard assumptions:

1. The individuals in the study are randomly sampled from the population.

2. *(stable unit treatment value assumption, or SUTVA)*. For any individual i, $i = 1, \cdots, n$, if $Z_i = z$, then $Y_i = Y_i(z)$, for all $z \in \{1, \cdots, J\}$.

3. *(strong unconfoundedness)*. Assignment to treatment Z is strongly unconfounded if $Z_i \perp\!\!\!\perp \{Y_i(1), \cdots, Y_i(J)\}|\boldsymbol{X}_i$, for all $z \in \{1, \cdots, J\}$.

4. *(overlap)*. For all values of z and $\boldsymbol{x}$, $0 < e_z(x) < 1$, where $e_z(x) \equiv pr(Z_i = z|\boldsymbol{x})$ is the generalized propensity score [21].

SUTVA states that the potential outcomes of one unit are not affected by the treatments received by other units, and there are no hidden treatment versions [43]. Strong unconfoundedness and overlap are an extension of the strong ignorability assumption in Rosenbaum and Rubin [10] to the case of multiple treatments. In some cases, a weaker version of unconfoundedness is sufficient for identifying the causal effect [21, 24], namely

2*. Three *(weak unconfoundedness)* Assignment to treatment Z is weakly unconfounded if $D_i(z) \perp\!\!\!\perp Y_i(z)|\boldsymbol{X}_i$, for all $z \in \{1, \cdots, J\}$.

Weak unconfoundedness only requires pairwise independence for each treatment rather than the independence between treatment assignment and the whole vector of potential outcomes. As commented by Imbens [21], though Assumption 2* is more relaxed in its form than Assumption 2, their difference has limited practical implications. Under these assumptions, the differences in outcomes among the treatment groups has a causal interpretation with respect to the target population.

1.3 Generalized Propensity Score and its Estimation

An important tool in comparing causal treatment effects of J treatment groups is the vector of generalized propensity scores (GPS), denoted as $e(\boldsymbol{X}_i) \equiv \{e_1(\boldsymbol{X}_i), \cdots, e_{(J-1)}(\boldsymbol{X}_i)\}^T$, where $e_z(x) \equiv pr(Z_i = z|\boldsymbol{x})$. In an observational study, the treatment assignment mechanism is unknown, and therefore $e(\boldsymbol{X}_i)$ needs to be estimated from the observed data. A common approach is to fit a multinomial logistic regression model for the treatment received as a function of the covariates, that is, to assume that

$$\log \frac{pr(Z_i = z|\boldsymbol{X}_i)}{pr(Z_i = J|\boldsymbol{X}_i)} = \boldsymbol{X}_i^T \boldsymbol{\beta}_z, \tag{1.1}$$

where $z = 1, \cdots, J-1$, and $\boldsymbol{X}_i$ includes an intercept term. The corresponding estimated GPS, denoted as GLMPS, is then

$$e_{z,GLMPS}(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_z) = \frac{\exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}}_z)}{1 + \sum_{j=1}^{J-1} \exp(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_j)}$$

for $z = 1, \cdots, J-1$, where $\hat{\boldsymbol{\beta}}_z$ is the maximum likelihood estimate of $\boldsymbol{\beta}_z$. For $z = J$, the reference group, we replace the numerator by 1.

Even moderate misspecification of the functional form for 1.1 may result in substantial bias in the estimates of treatment effects [44]. Imai and Ratkovic [45] proposed the Covariate Balancing Propensity Score (CBPS) for the comparison of two groups and provided an extension to the multiple treatment case. CBPS exploits the covariate balancing property of the GPS (i.e., $\boldsymbol{X}_i \perp\!\!\!\perp D_i(z)|e_z(\boldsymbol{X}_i)$ for $z = 1, \cdots, J$) by computing generalized method of moments estimates based on the covariate balancing moment conditions,

$$E\left\{\frac{D_i(z+1)\boldsymbol{X}_i}{e_{z+1}(\boldsymbol{X}_i)} - \frac{D_i(z)\boldsymbol{X}_i}{e_z(\boldsymbol{X}_i)}\right\} = 0$$

and the moment conditions derived from the score functions of a multinomial logistic model under the likelihood framework,

$$E\left\{\frac{D_i(z)}{e_z(\boldsymbol{X}_i)} \cdot \frac{\partial e_z(\boldsymbol{X}_i)}{\partial \boldsymbol{\beta}_z^T}\right\} = 0$$

for $z = 1, \cdots, J$. The CBPS is called just-identified if the model only uses the covariate balancing conditions and overidentified if both conditions are used in the estimation step. These two types of CBPS have different asymptotic and finite sample properties, and the authors examined both types of scores in their simulation studies [45]. They showed that the use of CBPS, regardless of which

conditions were involved, can improve the precision and reduce bias of some common weighting estimators (e.g. IPW and AIPW) compared to using propensity score estimated by GLM when both propensity score and outcome models were misspecified. In our study, we only evaluate the just-identified CBPS, because of computational limitations. The CBPS method can be implemented through the R package CBPS [46].

1.4 Methods for Estimating the Average Treatment Effect

1.4.1 Matching methods based on the propensity scores

The AI-type matching methods [31] can be regarded as a group-by-group imputation procedure. The missing outcome $Y_i(z)$, $z \neq Z_i$, is imputed by the observed outcome $Y_{k(i,z)}$ for one of the units $k(i,z)$ in the set of units, say $S(z)$, assigned to treatment z. That is, the observed or imputed outcome for unit i is

$$\hat{Y}_i(z) = \begin{cases} Y_i, & \text{if } Z_i = z; \\ Y_{k(i,z)}, & \text{if } Z_i \neq z. \end{cases}$$

The matched unit $k(i,z)$ is chosen to be the closest to unit i in $S(z)$ with respect to a matching metric m based on the values of $\boldsymbol{X}$. That is, $m(\boldsymbol{X}_i, \boldsymbol{X}_{k(i,z)}) \leq m(\boldsymbol{X}_i, \boldsymbol{X}_l)$ for all $l \in S(z)$. The matches are with replacement, so units in the matching set $S(z)$ can be reused. The resulting estimate of the ATE comparing treatments z and z' is

$$\hat{\tau}_{ATE}(z, z') = n^{-1} \sum_{i=1}^{n} \{\hat{Y}_i(z) - \hat{Y}_i(z')\}$$

The standard error can be computed using the delta method.

Ideally the matching units would be exact matches, that is, $\boldsymbol{X}_i = \boldsymbol{X}_{k(i,z)}$ for all i, z, which leads to unbiased estimates of ATEs under the strong unconfoundedness assumption. In practice, exact matching is rarely possible, especially with continuous covariates. With the Mahalanobis metric, $m(\boldsymbol{X}_i, \boldsymbol{X}_l) = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_l)^T C_X^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_l)}$ for $l \in S(z)$, where C_X is the covariance matrix of $\boldsymbol{X}_i$ and $\boldsymbol{X}_l$, we label this method as MCOV. This method may not work well for high-dimensional $\boldsymbol{X}_i$ [28]. An alternative is to match on closeness of the estimated GPS vector under a postulated model, $\hat{\boldsymbol{e}}(\boldsymbol{X}_i) = \{\hat{e}_1(\boldsymbol{X}_i), \cdots, \hat{e}_{J-1}(\boldsymbol{X}_i)\}^T$. The Mahalanobis distance $m(\boldsymbol{X}_i, \boldsymbol{X}_l) = \sqrt{\{\hat{\boldsymbol{e}}(\boldsymbol{X}_i) - \hat{\boldsymbol{e}}(\boldsymbol{X}_l)\}^T C_{GPS}^{-1} \{\hat{\boldsymbol{e}}(\boldsymbol{X}_i) - \hat{\boldsymbol{e}}(\boldsymbol{X}_l)\}}$, where C_{GPS} is the covariance matrix of $\hat{\boldsymbol{e}}(\boldsymbol{X}_i)$ and $\hat{\boldsymbol{e}}(\boldsymbol{X}_l)$, is one measure of closeness. We label this method MGPSV. The balancing score property of the propensity score implies that, under strong unconfoundedness, it yields approximately unbiased estimates of ATEs.

Yang et al. [24] proposed a method that matches units on the closeness of the corresponding estimated propensity score for each treatment group (MGPSS). The matching metric for imputing the missing outcomes for treatment z for units assigned to treatments other than z is then $m(\boldsymbol{X}_i, \boldsymbol{X}_l) = |\hat{e}_z(\boldsymbol{X}_i) - \hat{e}_z(\boldsymbol{X}_l)|$, where $l \in S(z)$. The resulting estimate of the ATE is approximately unbiased under the weak unconfoundedness assumption, because the definition of GPS implies that

$$\tau_{ATE}(z, z') = E\{E[Y_i|Z_i = z', e_{z'}(\boldsymbol{X}_i)]\} - E\{E[Y_i|Z_i = z, e_z(\boldsymbol{X}_i)]\}.$$

There are several differences between AI-type matching estimators and traditional matching estimators in applied research, such as nearest neighbor matching without replacement [28]. Traditional matching procedures address the issue of confounding by only including matches of high quality for the subsequent analysis. Normally each unit is only used once, as in a randomized control trial, and inferences on the matched data set do not account for matching error. On the other hand, AI-type matching allows reuse of each unit, and does not ensure overlap of covariates unless combined with methods for dealing with limited overlap, such as trimming [47]. An advantage of the AI-type matching estimators is that their large-sample distributions can be characterized [31, 48], permitting calculation of variance estimates that take into account the uncertainty in the propensity score estimation and matching procedure. MCOV, MGPSV, and MGPSS estimate τ_{ATE} while the estimand of traditional matching procedure may deviate from τ_{ATE}.

1.4.2 Propensity score weighting-based methods

For weighting-based estimators, the problem of estimating τ_{ATE} or τ^*_{ATE} can be generalized to the estimation of the (weighted) average potential outcome $\nu_z \equiv E[w(\boldsymbol{X})Y(z)]/E[w(\boldsymbol{X})]$ for each treatment separately. When $w(\boldsymbol{x}) = 1$, ν_z is equivalent to the average potential outcome μ_z. Solving the estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{w(\boldsymbol{X}_i)D_i(z)(Y_i - \nu_z)}{\hat{e}_z(\boldsymbol{X}_i)} \right\} = 0, \tag{1.2}$$

we are able to obtain a consistent estimator assuming correctly-specified GPS model,

$$\hat{\nu}_z = \left(\sum_{i=1}^{n} \left\{ \frac{w(\boldsymbol{X}_i)D_i(z)}{\hat{e}_z(\boldsymbol{X}_i)} \right\} \right)^{-1} \sum_{i=1}^{n} \left\{ \frac{w(\boldsymbol{X}_i D_i(z)Y_i)}{\hat{e}_z(\boldsymbol{X}_i)} \right\}.$$

The ATE between treatment z and z' can then be estimated by $\hat{\nu}_{z'} - \hat{\nu}_z$. Different choices of $w(\boldsymbol{x})$ result in ATE with respect to different populations. In particular, $w(\boldsymbol{x}) = 1$ corresponds to the

inverse probability weighting (IPW) estimator, whose target population is the combined population all sampled groups. The target population of ATT discussed in section 1.2.1 is represented by units in a particular treatment group, say treatment J, and can be estimated by setting $w(\boldsymbol{x})$ to $e_J(\boldsymbol{x})$.

Li and Greene [12] proposed to specify $w(\boldsymbol{x})$ as the minimum of the probabilities of receiving treatment and control in the binary case, which they call matching weights (MW). MW can be extended to the case with more than two treatments [13] with weights $w_{MW}(\boldsymbol{x}) = \min\{e_1(\boldsymbol{x}), \cdots, e_J(\boldsymbol{x})\}$. For the three treatment case, the MW estimator uses weights to mimic the 1:1:1 matching procedure without replacement and yields more efficient estimation of τ^*_{ATE} [12, 13]. The MW estimator and the estimator from 1:1:1 matching without replacement have asymptotically the same estimand [13], and therefore the corresponding target population of the MW estimator is the "matched" population of units that can be matched in 1:1:1 matching.

Li et al. [14] and Li and Li [26] proposed weighting by the overlap weights (OW), $w_{OW}(\boldsymbol{x}) = \left\{\sum_{j=1}^{J} 1/e_j(\boldsymbol{x})\right\}^{-1}$. We refer to the corresponding population as the overlap population. Both MW and OW upweight the units whose GPS is in the middle range, which have approximately equal chances of being assigned to any of the candidate treatments.

Inversely-weighted estimators have a number of issues. The first is that their variance may be inflated if the weights are highly variable. The second issue is that they rely heavily on the correct specification of the propensity score model for valid inference. In addition, the inference for treatment group z is made only based on individuals with $D_i(z) = 1$, with individuals in other treatment groups not contributing. To improve the robustness to model misspecification and make more effective use of the available data, augmented versions of these estimators have been proposed [12, 38]. The estimating equation (1.2) is augmented by an extra term that involves a function of x. The resulting estimating equation is

$$\sum_{i=1}^{n} \left\{ \frac{w(\boldsymbol{X}_i) D_i(z)(Y_i - \nu_z)}{\hat{e}_z(\boldsymbol{X}_i)} - \frac{w(\boldsymbol{X}_i)[e_z(\boldsymbol{X}_i) - D_i(z)]}{e_z(\boldsymbol{X}_i)} h(\boldsymbol{X}_i) \right\} = 0$$

The resulting estimator $\hat{\nu}_z$ achieves the smallest asymptotic variance when $h(\boldsymbol{X}_i) = E(Y_i - \nu_z | Z_i = z, \boldsymbol{X}_i)$ [39]. We label the augmented versions of IPW, MW, and OW estimators as AIPW, AMW, and AOW, respectively. Besides asymptotic efficiency, as shown in the original set of papers [12, 14], for any scalar outcome, the corresponding estimator has the property of double robustness, which means that only one of the propensity score and outcome models need to be correctly specified to obtain a consistent estimator for ν_z. Semiparametric theory shows that these estimators are asymptotically normal, and variances can be estimated using sandwich-type estimators or the bootstrap [3, 38].

1.4.3 Outcome regression model methods

The methods based on outcome regression directly models the relationship between the outcome and pre-treatment covariates by treatment groups. The unconfoundedness assumption implies that the ATE can be identified by positing a parametric model for $E[Y_i|Z_i = z', \boldsymbol{X}_i]$ and $E[Y_i|Z_i = z, \boldsymbol{X}_i]$, obtaining the predicted values of Y_i under each treatment group for each $\boldsymbol{X}_i$, and taking the average over the observed and predicted values for each treatment. For a binary outcome Y_i, predictions can be based on a logistic regression model:

$$\log \frac{pr(Y_i = 1|Z_i, \boldsymbol{X}_i)}{pr(Y_i = 0|Z_i, \boldsymbol{X}_i)} = \gamma + \boldsymbol{X}_i^T \boldsymbol{\alpha} + \sum_{z=1}^{J-1} \theta_j D_i(z), \tag{1.3}$$

where treatment J is considered as the reference group. The coefficients $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{J-1})$ and $\boldsymbol{\alpha}$ can be replaced by maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$. Many applied studies that use this conventional covariate-adjustment method report $\hat{\boldsymbol{\theta}}$'s which represent the odds ratios conditional on $\boldsymbol{x}$, as the estimated effect measure. Outcome regression (OREG) then estimates the risk difference between treatment z and z' as $\hat{\tau}_{OREG}(z, z') = \hat{\mu}_{z'} - \hat{\mu}_z$, where

$$\hat{\mu}_z = n^{-1} \sum_{i=1}^{n} \text{expit}(\hat{\gamma} + \hat{\theta}_z + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}})$$

for $z = 1, \cdots, J-1$, and

$$\hat{\mu}_z = n^{-1} \sum_{i=1}^{n} \text{expit}(\hat{\gamma} + \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}})$$

for $z = J$. The associated standard error can be estimated via bootstrap.

Utilizing this idea, Zhou et al. [17] proposed PENCOMP, which estimates causal effects comparing two treatments for a continuous outcome by imputing unobserved potential outcomes from the corresponding predictive distributions. PENCOMP incorporates splines of propensity scores as predictors in the outcome model, which gives it a double robustness property for a continuous outcome such that the estimator for the marginal mean is consistent if a) the prediction models are correctly specified, or b) the propensity model and the relationship between the outcome and the splines are correctly specified. We extend PENCOMP at a single time point to more than two treatments and a binary outcome, calling the method PEN-GAM. The double robustness property for PEN-GAM has not yet been theoretically established. However, our simulation studies shed light on its finite sample performance. The steps for PEN-GAM can be summarized as follows:

(a) Generate a bootstrap sample $S^{(b)}$ for $b = 1, \cdots, B$, stratified on treatment groups, from the

original data set. For each $S^{(b)}$, repeat steps (b)-(d).

(b) Estimate the GPS, possibly from a multinomial logistic regression model. Denote the estimated values as $\hat{\boldsymbol{e}}_i = \left\{\hat{e}_1\left(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_1^{(b)}\right), \cdots, \hat{e}_{J-1}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_{J-1}^{(b)}\right)\right\}$, where $\hat{e}_z\left(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_z^{(b)}\right) = pr\left(Z_i = z|\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}_z^{(b)}\right)$ and $\hat{\boldsymbol{\beta}}_z^{(b)}$ is the maximum likelihood estimate of $\boldsymbol{\beta}_z$ for sample $S^{(b)}$.

(c) For $z = 1, \cdots, J$, fit a generalized linear regression model

$$\log \frac{pr\{Y_i(z) = 1|Z_i = z, \boldsymbol{X}_i, \boldsymbol{\theta}_z, \boldsymbol{\alpha}_z\}}{pr\{Y_i(z) = 0|Z_i = z, \boldsymbol{X}_i, \boldsymbol{\theta}_z, \boldsymbol{\alpha}_z\}} = s(\hat{\boldsymbol{e}}_i^*|\boldsymbol{\theta}_z) + g(\boldsymbol{X}_i, \hat{\boldsymbol{e}}_i^*; \boldsymbol{\alpha}_z), \tag{1.4}$$

where $s(\hat{\boldsymbol{e}}_i^*|\boldsymbol{\theta}_z)$ denotes a penalized spline with fixed knots, and $g(\cdot)$ denotes a parametric function of the covariates and propensity scores and has to be constrained to ensure identifiability. In this case we assume truncated linear basis, namely, $s(\hat{\boldsymbol{e}}_i^*|\boldsymbol{\theta}_z) = \sum_{z=1}^{J-1}\{\theta_{0z} + \theta_{1z}\hat{e}_{iz}^* + \sum_{k=1}^{K}\theta_{1zk}(\hat{e}_{iz}^* - Q_k)_+\}$, where $Q_1, \cdots, Q_K$ are fixed knots, and $(\hat{e}_{iz}^* - Q_k)_+ = \hat{e}_{iz}^* - Q_k$ if $\hat{e}_{iz}^* > Q_k$, and $(\hat{e}_{iz}^* - Q_k)_+ = 0$ otherwise. Note that following [17], we fit different spline functions in (1.4) for each treatment level z. For linear regression of $Y_i(z)$, the coefficients in the spline model can be estimated in a linear mixed model framework [49] and implemented using standard statistical software, as was done in [17]. In principal, the coefficients of a generalized linear model with penalized spline terms as (1.4) can be obtained by fitting a generalized linear mixed models (GLMM). However, to the best of our knowledge, current GLMM implementation in R either does not allow the specification of the structure of the covariance matrices or will take unreasonable running time. Therefore, we instead fit a generalized additive model (GAM) using the gam function in the mgcv package in R [50].

(d) For $z = 1, \cdots, J$, impute the values of $Y(z)$ for subjects with $D(z) = 0$ in the original dataset with draws from the Bernoulli distribution with predictive probability $pr\{Y_i(z) = 1|Z_i = z, \boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_z^{(b)}, \hat{\boldsymbol{\alpha}}_z^{(b)}\}$, where $\hat{\boldsymbol{\theta}}_z^{(b)}$ and $\hat{\boldsymbol{\alpha}}_z^{(b)}$ are estimates for the coefficients $\boldsymbol{\theta}_z^{(b)}$ and $\boldsymbol{\alpha}_z^{(b)}$, respectively, for the bth bootstrap replicate. For subjects with $D_i(z) = 1$, $Y_i(z) = Y_i$.

(e) Derive the estimated treatment effects and associated standard error using Rubin's Rules [51].

For all methods discussed in this section, we refer the readers to the corresponding R packages developed by the authors (Table 1.1). In the cases where there are no R packages available, we provide accessible code for easier implementation at https://github.com/youfeiyu/multiTreatment.

Method	Reference	R package/author generated code	Which unconfoundedness assumption is made
NAIVE	N/A	https://github.com/youfeiyu/multiTreatment	N/A
OREG	N/A	https://github.com/youfeiyu/multiTreatment	Assumption 2*
PENCOMP	Zhou et al. [17]	https://github.com/youfeiyu/multiTreatment	Assumption 2
Propensity Score Matching			
MCOV	Abadie and Imbens [31]	Matching [52], Matchit [53, 54]	Assumption 2
MGPSV	Yang et al. [24]	https://github.com/youfeiyu/multiTreatment	Assumption 2
MGPSS	Yang et al. [24]	Multilevelmatching (https://github.com/shuyang1987/multilevelMatching/)	Assumption 2*
Propensity Score Weighting			
IPW, AIPW	Lunceford and Davidian [3], among others	https://github.com/youfeiyu/multiTreatment	Assumption 2*
MW, AMW	Li and Greene [12], Yoshida et al. [13]	https://github.com/youfeiyu/multiTreatment	Assumption 2*
OW, AOW	Li and Li [26]	PSweight [55], or https://github.com/youfeiyu/multiTreatment	Assumption 2*

Table 1.1: Causal inference methods under comparison and their corresponding R implementation. NAIVE estimator refers to the direct comparison of the proportions of each treatment group. PENCOMP was developed in the context of binary treatment and continuous outcome. We extend it to the case of multiple treatment and binary outcome. Assumptions 2 and 2^* are the strong and weak unconfoundedness assumption, respectively.

1.5 Simulation Studies

We conducted simulation studies to assess the finite sample properties of the twelve estimators listed in Table 1.1 combined with the two GPS estimation methods (GLMPS and CBPS) discussed in Section 1.3. We used direct comparison of the proportions of each group as a benchmark, which is referred to as the naive estimator. We considered two levels of covariate overlap (good and poor), two functional forms for the true propensity score model (linear and nonlinear in covariates), two levels of associations for the outcome model (strong and weak), two levels of overall marginal outcome prevalence (common [0.3] and rare [0.1]) and two sample sizes (300 and 1500). Simulation results are presented in terms of bias from ATE, empirical standard deviation, average standard error, root mean squared error (RMSE), average width of 95% confidence intervals (CI), and 95% coverage rate.

1.5.1 Simulation Design

Each simulated dataset contains six covariates. $(X_{i1}, X_{i2}, X_{i3})^T$ follows a multivariate normal distribution with mean $(0,0,0)^T$ and covariance matrix $[(2,1,-1)^T, (1,1,-0.5)^T, (-1,-0.5,1)^T]$, $X_{i4} \sim Bernoulli(0.5)$, $X_{i5} \sim Bernoulli\{0.75X_{i4} + 0.25(1 - X_{i4})\}$, and X_{i6} follows a chi-squared distribution with 1 degree of freedom. Let $X_i = (1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})^T$. Three treatment groups were compared, and the true GPS model was given by $Z_i \sim$ Multinomial $\left\{e_1(\tilde{\boldsymbol{X}}_i), e_2(\tilde{\boldsymbol{X}}_i), e_3(\tilde{\boldsymbol{X}}_i)\right\}$, where $\tilde{\boldsymbol{X}}_i$ is a function of $\boldsymbol{X}_i$ that corresponds to a model specification and

The GPS were estimated in two ways, the first using a multinomial logistic regression and the second using the CBPS framework that incorporates covariate balancing conditions [45]. Since PENCOMP is computationally intensive, we only implemented GLMPS (not CBPS) for this method. We used 10 equally-spaced knots on the logit scale for each GPS component. We used 200 imputed datasets to estimate treatment effects and the associated standard errors and confidence intervals.

For each scenario, we generated 2000 Monte Carlo datasets for each of two sample sizes, 300 and 1500. The true 1000$\times$ATEs (risk differences) for the estimands $\tau_{ATE}(1,2)$, $\tau_{ATE}(1,3)$, and $\tau_{ATE}(2,3)$ were respectively 56, 46, and -10 for scenario 3, -1, -24, and -23 for scenario 5, and 234, 76, and -158 for the other three scenarios, which were determined over 10^6 sample units.

For estimation methods that involve only the GPS or the outcome model (IPW, MW, OW, MGPSV, MGPSS, and OREG), we studied their performance when the corresponding model is correctly (c) and incorrectly (m) specified, respectively. For augmented estimators (AIPW, AMW, AOW, PEN-GAM), we considered the following four cases:

(1) both GPS and outcome models are correctly specified denoted by (c, c),

(2) the GPS model is correct while the outcome model is incorrect denoted by (c, m),

(3) the outcome model is correct while the GPS model is incorrect denoted by (m, c),

(4) both models are misspecified denoted by (m, m).

For the first three scenarios, the misspecification of both models is caused by removing one of the confounders, X_{i6}, from the corresponding models. For scenario 4 where the true GPS model is nonlinear in $\boldsymbol{X}_i$, the misspecified outcome model omits X_{i6}, while the incorrect GPS model incorporates the whole set of covariates ($\boldsymbol{X}_i$) but ignores the higher order and interaction terms. Similarly, we evaluated the performance of MCOV, which is free of parametric modeling, when matching on all elements $\tilde{\boldsymbol{X}}_i$, and on a subset of $\tilde{\boldsymbol{X}}_i$, where the subset being the same as the set of variables adjusted in the GPS model.

The 95% confidence intervals were calculated using: (1) bootstrapped standard errors from 200 bootstrap samples for OREG, IPW, AIPW, MW, AMW, OW, AOW, and CBPS-based MGPSS; (2) Wald-type confidence interval based on original data for NAIVE; (3) Abadie and Imbens (2006) confidence interval for MCOV and both GLMPS- and CBPS-based MGPSV [24, 31]; (4) Abadie and Imbens (2016) confidence interval for GLMPS-based MGPSS [24, 48]; (5) Rubin's imputation rule for PEN-GAM [17].

1.5.2 Simulation Results

The main results of the simulation studies for sample size 1500 are summarized in Figures 1.1-1.7. The complete results are presented in Tables A.3-A.9 for sample size 1500, and Figures A.3-A.9 and Tables A.10-A.16 for sample size 300. In all scenarios, all estimators for τ_{ATE} with at least one model correctly specified yielded smaller empirical bias compared to the naive estimator.

Three key takeaways from the simulation studies are summarized below:

1. The improvement in precision was limited for AIPW and PEN-GAM compared to IPW when a) there was sufficient covariate overlap or b) the prevalence of the outcome was low.

2. With moderate prevalence of the outcome (0.3 in our simulation setting) or relatively poor covariate overlap, AIPW and PEN-GAM outperformed IPW and AI-type matching algorithms considered in this study in terms of RMSE across the scenarios, as AIPW and PEN-GAM incorporate the outcome information, which tended to provide efficiency gains over IPW and AI-type matching.

3. For a relatively small sample size, PEN-GAM with at least one model being correctly specified were noted to be slightly biased away from the true risk difference. Moreover, PEN-GAM tended to show over-coverage and produce wider confidence width than IPW when the outcome is sparse. One reason is that the fitting of spline models in PEN-GAM is more unstable with low outcome prevalence and small sample size. The empirical bias and over-coverage tended to disappear as the outcome prevalence and sample size increased.

Results of RMSE for each of the treatment comparisons averaged over 2000 datasets for sample size 1500 across all methods that estimate τ_{ATE} are presented in Figures 1.1-1.3. Note that the corresponding estimands for MW, AMW, OW, and AOW were in general different from τ_{ATE}, and the RMSE for these estimators are shown in Tables A.3 and A.10. We report the ratio of RMSE to the RMSE of the GLMPS-based IPW estimator with correctly specified GPS model. When both models were correctly specified and the overlap in covariate distributions was good (Figure 1.1, scenario 1), OREG, IPW, AIPW, and PEN-GAM had similar RMSE. Matching methods had larger RMSE than GLMPS-based IPW, with the ratios ranging from 1.1 to 1.2. In this case, AIPW and PEN-GAM had similar empirical standard deviation (and therefore RMSE) to IPW (Table A.4). A study conducted by Austin showed similar results that AIPW provided little efficiency gain over IPW [56].

In the presence of poor covariate overlap (Figure 1.1, scenarios 2-4), OREG had the smallest RMSE, followed by PEN-GAM and AIPW. We observed 6.3%-16.5% reduction in RMSE for AIPW and PEN-GAM compared to GLMPS-based IPW when the associations between the outcome and covariates was weak (scenario 2). Greater reduction (14.1%-40.1%) was noted as the associations became stronger (scenario 3). When the prevalence of the outcome was low (scenario 5), AIPW barely reduced RMSE compared to IPW, and PEN-GAM had larger RMSE than IPW. The increased RMSE for PEN-GAM may result from the instability of model fitting with low prevalence. MGPSS had larger RMSE than MGPSV, which was also observed for the scenario with good covariate overlap. For all scenarios considered, RMSEs of GLMPS-based estimators were close to those of their CBPS-based counterparts (Figure 1.1 and Table A.3). One exception is that for IPW, the use of CBPS tended to reduce RMSE compared with GLMPS when the covariate overlap was poor.

When only the GPS model was correctly specified (Figure 1.2), PEN-GAM and AIPW in general had the lowest RMSEs across the scenarios with moderate prevalence, and the RMSEs for PEN-GAM were close to or lower than those for AIPW. When only the outcome was modeled correctly (Figure 1.3), the RMSEs for AIPW and PEN-GAM remained similar to or lower than those for IPW with correctly specified GPS model. In scenario 4 where the misspecification of the GPS model was caused by incorrect functional form, the use of GLMPS may lead to substantial RMSE for IPW (Figure 1.3) and AIPW with misspecified outcome model (Table A.7) due to large

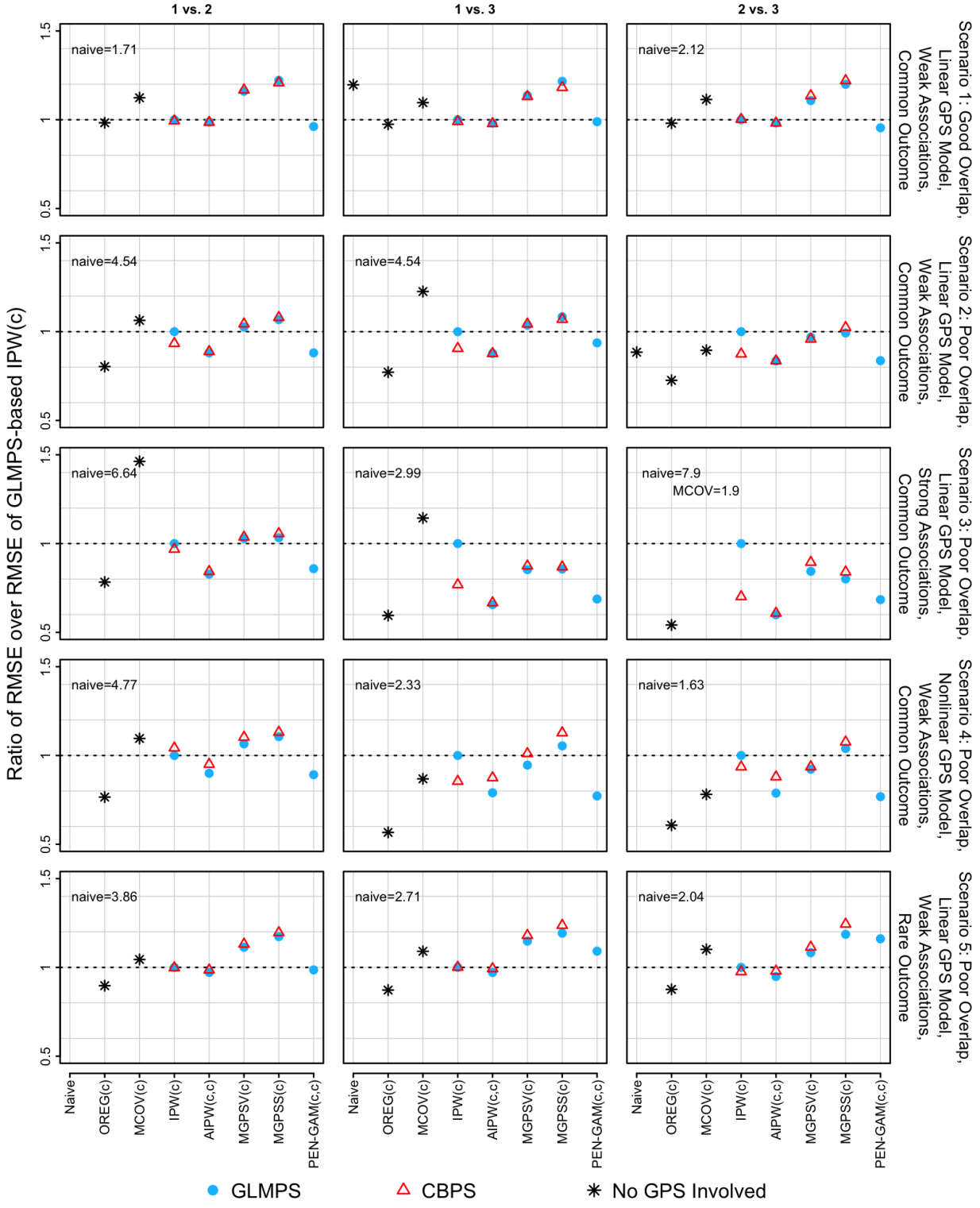

Figure 1.1: Ratio of RMSE over RMSE of GLMPS-based IPW(c) for sample size 1500 across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

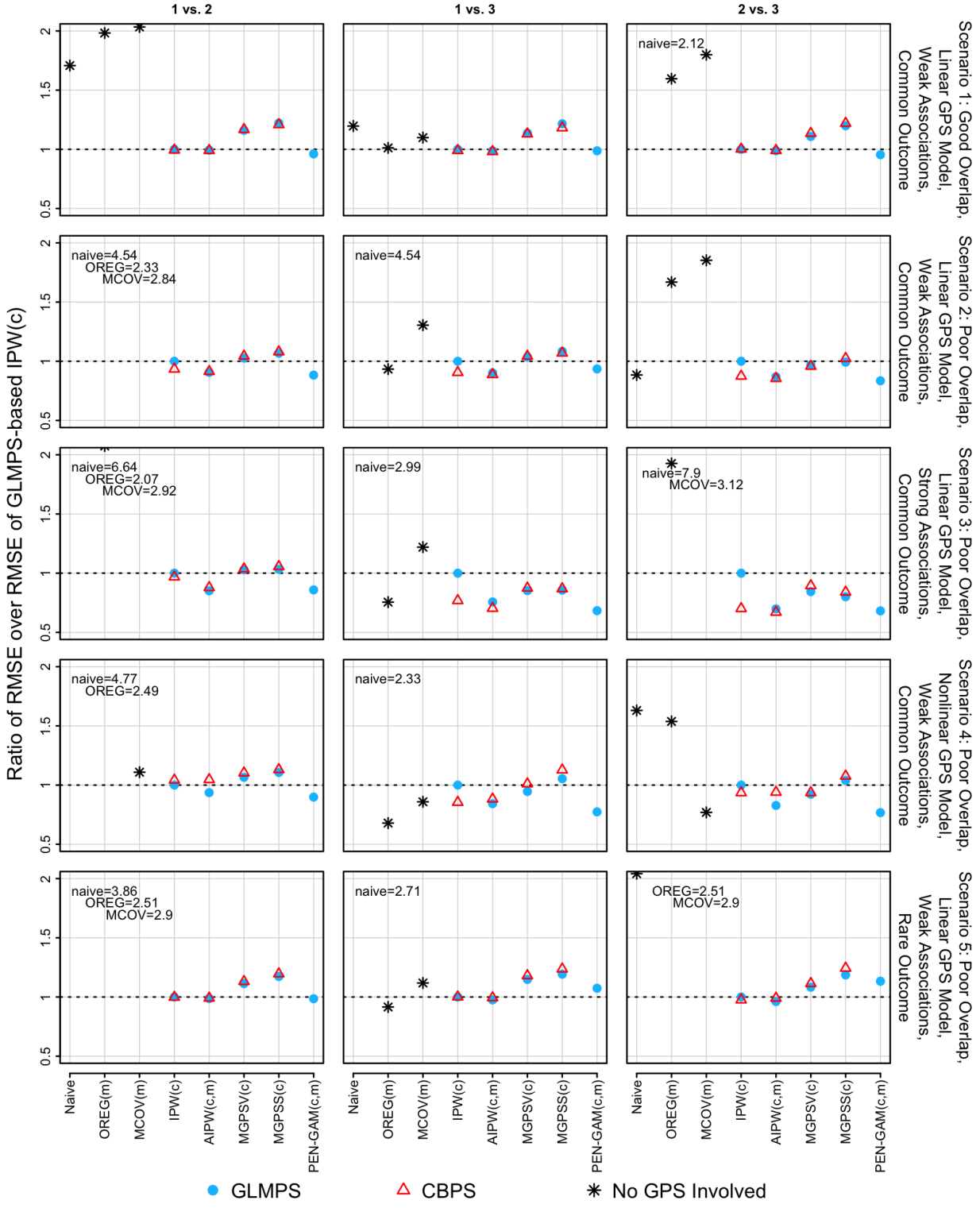

Figure 1.2: Ratio of RMSE over RMSE of GLMPS-based IPW(c) for sample size 1500 across methods based on a correctly specified propensity model only. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

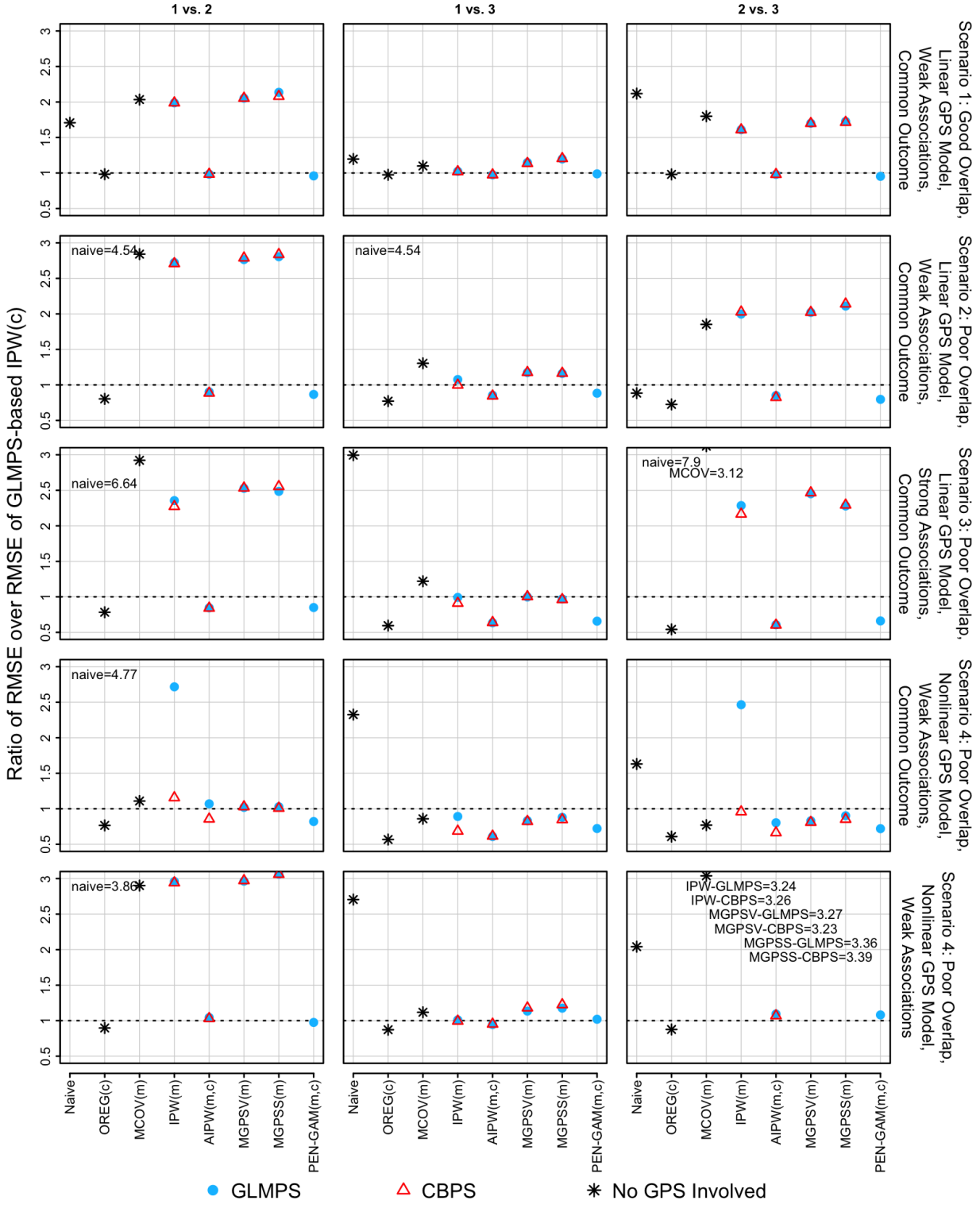

Figure 1.3: Ratio of RMSE over RMSE of GLMPS-based IPW(c) for sample size 1500 across methods based on a correctly specified outcome model only. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

empirical bias, which is consistent with previous findings [44, 45]. The bias was greatly reduced and became close to zero when GLMPS were replaced by CBPS with misspecified functional form, which leaded to smaller RMSEs. The RMSEs of the AI-type matching methods (MGPSS, and MGPSV) were noted to be smaller than those of GLMPS-based IPW in scenario 4, since the matching methods yielded approximately unbiased estimates of ATE (Table A.7) even when the GPS model was incorrect but adjusted for the whole set of confounders, which indicates that matching methods are more robust to the omission of higher order and interaction terms in the GPS model than IPW.

The empirical coverage rates of 95% confidence interval for sample sizes 1500 with both models correctly specified and either one of the models misspecified are shown in Figures 1.4 and 1.5, respectively. The true values for MW, AMW, OW, and AOW were determined using the true GPS based on 10^6 sample units and used to evaluate the corresponding coverage rates. In general, when both models were correctly specified (Figures 1.4), all methods except MCOV had close to nominal coverage of 95% for moderate prevalence. Coverage for MCOV was far below nominal in scenarios 2 and 3 with moderate and strong confounding, respectively. This under-coverage was primarily the result of empirical bias (Tables A.5 and A.6).

With the outcome model being misspecified (Figure 1.5), all of the augmented estimators showed reasonable coverage. Note that the corresponding estimands of MW, OW, and their augmented versions depend on the actual values of GPS. Therefore, different specifications of GPS model lead to different estimands, while the estimands based on the true GPS model were used for evaluating the coverage rates, which explains the under-coverage of AMW and AOW in some scenarios when the GPS model was misspecified (Figure 1.5). For a small sample size ($n = 300$) or sparse outcome (scenario 5), we consistently observed over-coverage for PEN-GAM methods across all scenarios regardless of the specifications of the models, with some of the CIs achieving 99% coverage (Figures A.6 and A.7, and scenario 5 in Figures 1.4 and 1.5). This finding agrees with the overestimation of the standard errors for PEN-GAM observed in Tables A.8 and A.11-A.15. The under-coverage for GLMPS-based MGPSS in scenario 3 (Figure A.6) was caused by the underestimation of the standard errors using the asymptotic formula provided in Yang et al. [24]. Such under-coverage was remedied as the sample size increased.

The average 95% CI widths for sample size 1500 are shown in Figures 1.6 and 1.7. When both models were correctly specified (Figure1.6), the average widths of OREG, AIPW, and PEN-GAM were close to or smaller than those of GLMPS-based IPW for common outcome. MGPSS and MGPSV tended to have wider confidence intervals than IPW across all scenarios. The average widths of CBPS-based estimators tended to be larger than those of their corresponding GLMPS-based ones. Figure 1.7 displays the results for the augmented estimators with either one of the models being misspecified. The relative relationships among IPW, AIPW, and PEN-GAM were

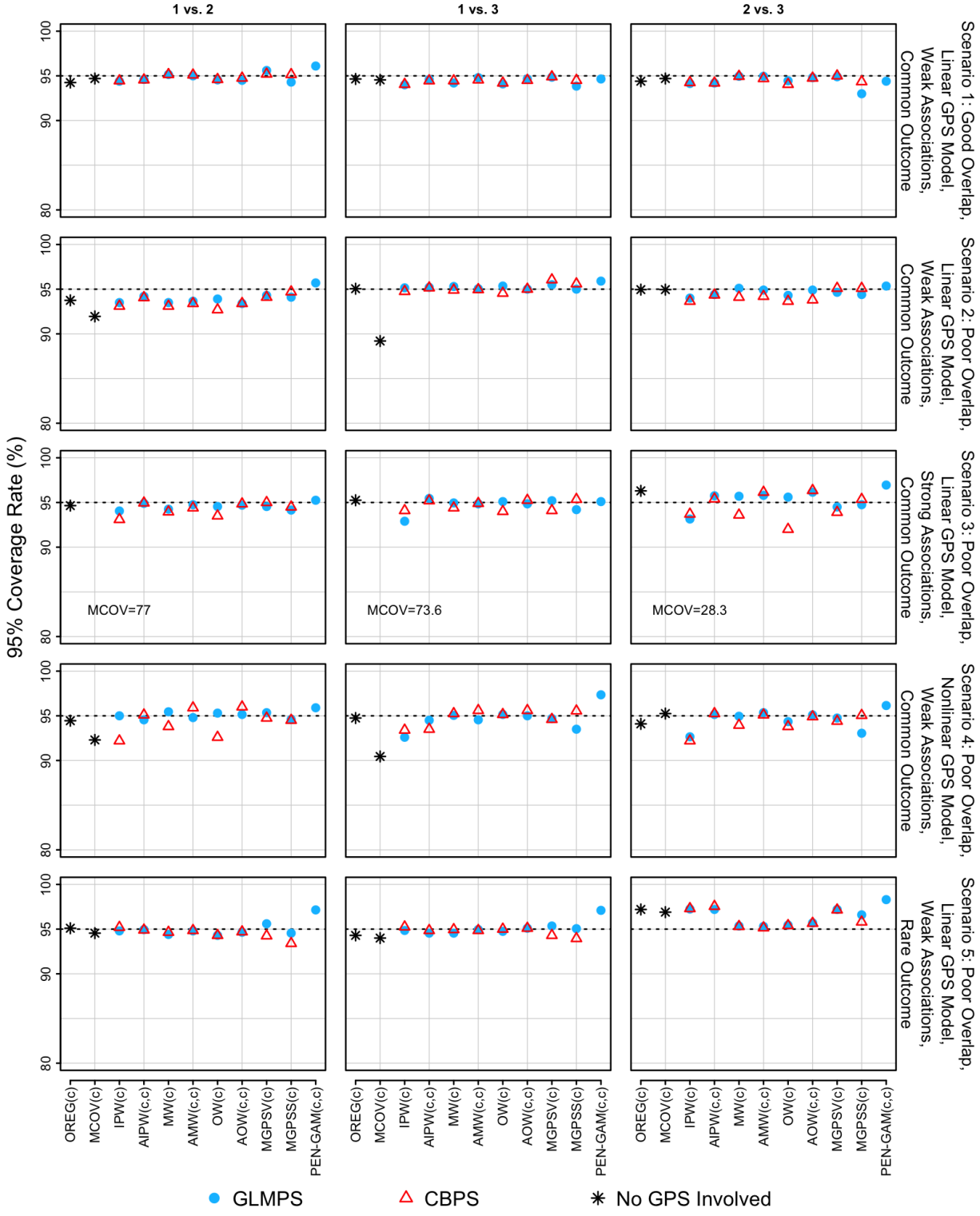

Figure 1.4: 95% Coverage probability for sample size 1500 across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

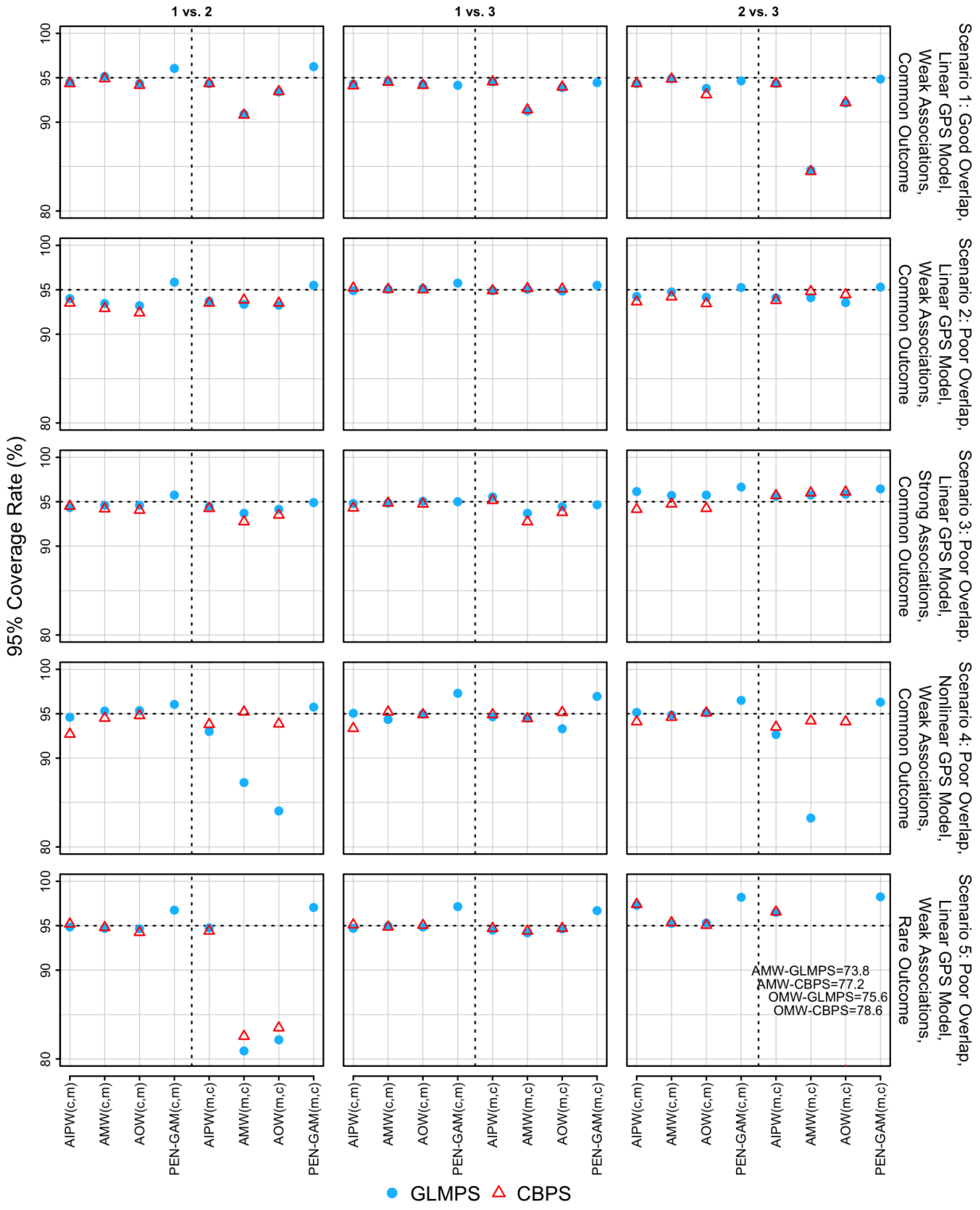

Figure 1.5: 95% Coverage probability for sample size 1500 across methods based on a correctly specified propensity score or outcome model. For methods that involve both models, the first and second letter in the parentheses correspond to the propensity and outcome model, respectively. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

similar to the ones in Figure 1.6 where both models were correct. In general, for all estimators considered in Figure 1.7, the CIs were wider when the outcome model was misspecified compared to the case with a misspecified GPS model only. For $n = 300$, the CIs for PEN-GAM were in general wider than those of IPW (Figures A.8 and A.9). The average standard errors of PEN-GAM were greater than their corresponding Monte Carlo standard deviations for all scenarios (Tables A.11-A.15), suggesting that PEN-GAM tends to be more sensitive to small sample size in terms of standard error estimation compared to IPW and AIPW.

MW, OW estimators and their augmented version provide stable estimates of τ^*_{ATE}, regardless of the overlap status in the covariate distribution of the original population (Tables A.4-A.8 and A.11-A.15). This is as expected since MW and OW artificially downweight the units with extreme GPS and upweight the units whose GPS for each treatment are similar, the latter of which tend to have a common support in their covariate distribution.

1.6 Data Analysis

1.6.1 Data Analysis Methods

We applied the methods in Table 1.1 to claims data of patients with metastatic castration-resistant prostate cancer (mCRPC), which was obtained from a large national private health insurance network (Optum Clinformatic Data Mart). Our data consisted of a subset of a previously identified cohort [57, 58, 59], which included patients who had at least one diagnosis of prostate cancer from January 1, 2010 to September 30, 2016 and used at least one of the six focus drugs (docetaxel, abiraterone, enzalutamide, sipuleucel-T, cabazitaxel, and radium-233) after the diagnosis. Since radium-233 were approved by FDA and released to the market later than the other five drugs, we restricted our cohort to patients who initiated treatment after January 1, 2014 to give them a fair comparison and make the results more generalizable to the current mCRPC population. We observed that the cabazitaxel and radium-233 groups had much fewer samples ($n_{cabazitaxel} = 11$ and $n_{radium} = 57$) than the other four groups, and therefore we further dropped those patients who received the two drugs as their first-lines therapy from our analysis. We assessed the safety of the four remaining drugs for mCRPC with the outcome being the occurrence of post-prescription emergency room (ER) visits during a fixed period of time. Specifically, we evaluated the risk difference of ER visits among the four drugs within 180-day time window of the initiation of each therapy.

Medical and pharmacy claims pertaining to ER visits were identified by procedure code and type of service variables in the database. In this study, we did not consider treatment sequence and hence were only interested in ER visits associated with the first drug used. Patients who

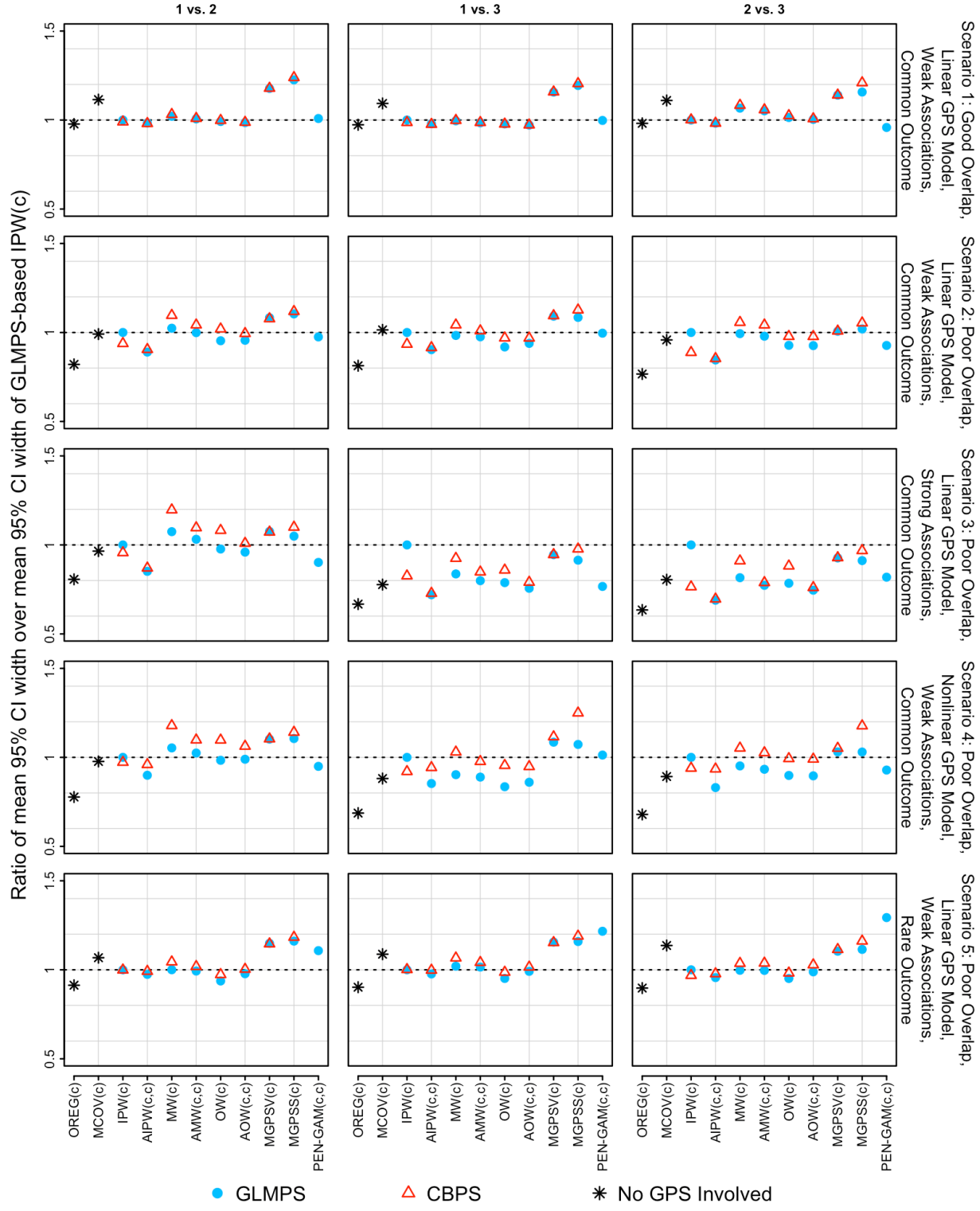

Figure 1.6: Ratio of mean 95% CI width over mean 95% CI width of GLMPS-based IPW(c) for sample size 1500 across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

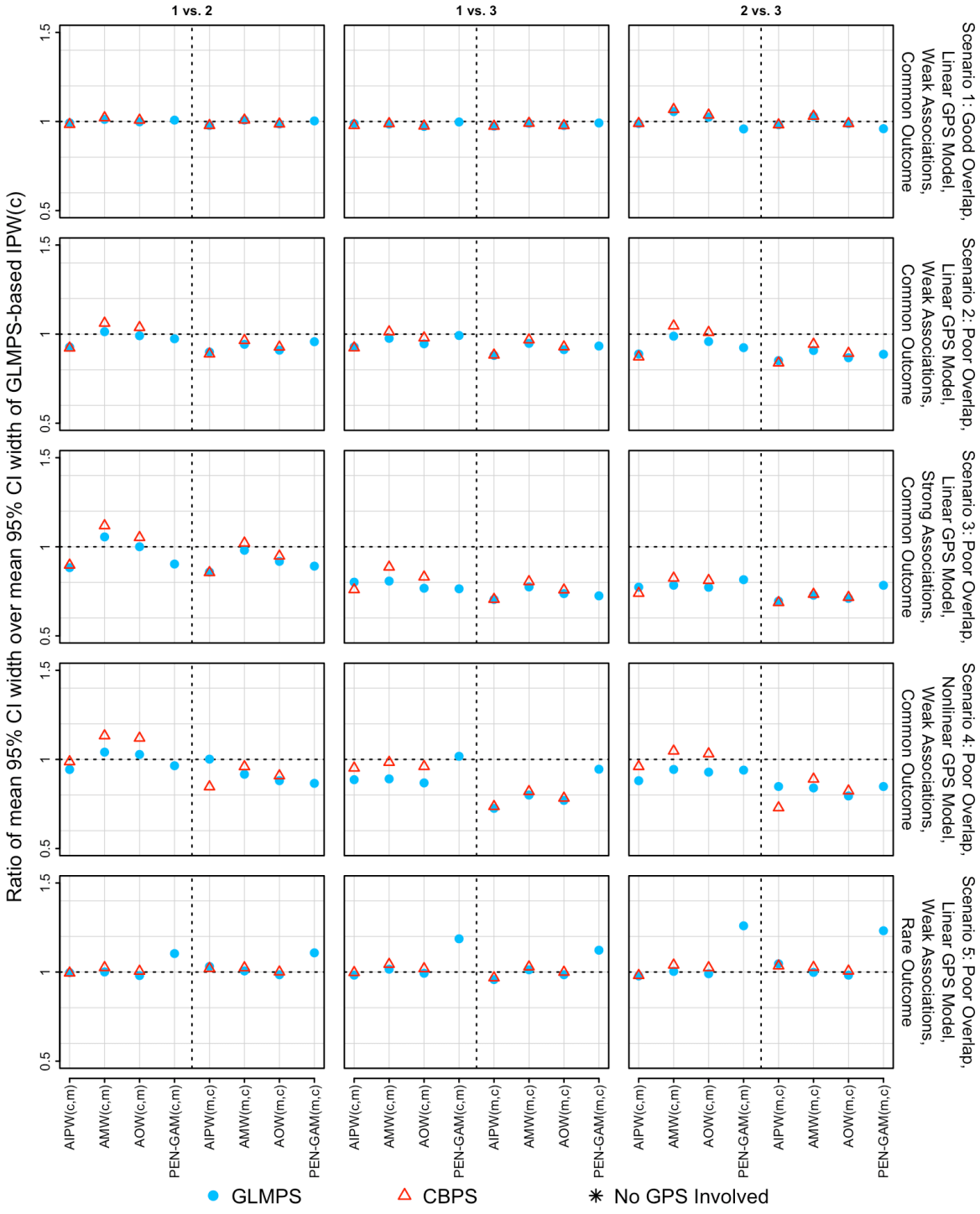

Figure 1.7: Ratio of mean 95% CI width over mean 95% CI width of GLMPS-based IPW(c) for sample size 1500 across methods based on a correctly specified propensity score or outcome model. For methods that involve both models, the first and second letter in the parentheses correspond to the propensity and outcome model, respectively. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

switched treatment or dropped out of the insurance plan within 180 days of the first prescription with no events (i.e. ER visits) occurring during the follow-up period were regarded as being censored. Censored patients exhibited similar demographic and baseline clinical characteristics to uncensored ones (Table A.17) and were dropped from the analysis. We first calculated the crude risks of at least one ER visit for 180-day follow up for each of the four focus drugs, and compared the risk among the four treatment groups using causal inference methods described in the previous section.

The GPS for each subject was estimated from a multinomial logistic regression model adjusting for age, race, education level, household income, geographic region, insurance product type, whether the insurance plan is administrative services only, metastatic status of cancer, year of first prescription, comorbid conditions, and provider type. All covariates were binary or categorical, and the categorization was summarized in Table A.18. We observed insufficient overlap among the four treatment groups in terms of the logit propensity of receiving docetaxel, especially at the left end of the distribution (Figure A.10A), which indicates that we may not be able to find a good match in docetaxel users for some patients receiving abiraterone, enzalutamide, or sipuleucel-T. Similar patterns occurred for the logit propensity of receiving the other three drugs (Figures A.10C, A.10E and A.10G). One can use trimming methods that discard the tails of propensity score distributions to remedy the lack of overlap. Several trimming criteria for three or more treatment groups are discussed in the literature [24, 25, 60]. In our case, we trimmed the data using the criteria described in Lopez and Gutman [25]. In brief, for each treatment $z \in \{1, 2, 3, 4\}$, where $l_z = \max_j\{\min_i(pr(Z_i = z|Z_i = j, \boldsymbol{X}_i))\} =$ and $u_z = \min_j\{\max_i(pr(Z_i = z|Z_i = j, \boldsymbol{X}_i))\}$, where $pr(Z = z|Z = j, \boldsymbol{X})$ is the treatment assignment probability for z among those receiving treatment j. Subjects with $e_z(\boldsymbol{x}) \notin [l_z, u_z]$ for any z were discarded. GPS were recalculated using the remaining subjects. One important step in propensity score modeling is balance checking. Ways to check for balance in covariates and their corresponding results for the methods considered are described in Section A.1 in the Appendix. The log odds of the outcome was modeled as a linear combination of the same set of covariates adjusted in the GPS model for each treatment group. The confidence intervals for each method were obtained in the same way as described in the simulation studies. Specifically, 200 bootstrap replicates were used for OREG, PEN-GAM, and all weighting-based methods.

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

1.6.2 Data Analysis Results

A total of 2628 mCRPC patients with at least 180 days of continuous enrollment prior to the receipt of the first focus drug were identified. The average and median length of the enrollment period that covers January 1, 2014 is 6.16 and 4.75 years, respectively. Among the 2,628 patients, 670 (25.5%) were censored and 4 (0.2%) had incomplete covariates. We further excluded these patients from the analysis. The demographic and baseline clinical data of the remaining 1,955 patients are presented in Table A.18. Table 1.2 presents the crude risks of at least one ER visit during 180-day follow up among uncensored patients for each of the four treatment groups. The unadjusted risk was the highest in the docetaxel group (51.5%), followed by Sipuleucel-T group (44.3%). Enzalutamide users had the lowest risk (25.5%) of at least one ER visit within 180 days.

First-line therapy	Total number of patients	Number of uncensored patients with complete covariates	At least 1 ER visit (%) within 180 days*
Docetaxel (Taxotere, Decefrez)	728	565	291 (51.5)
Abiraterone (Zytiga)	1039	783	314 (40.1)
Enzalutamide (Xtandi)	639	476	163 (34.2)
Sipuleucel-T (Provenge)	222	131	58 (44.3)

Table 1.2: Emergency room visits following the first prescription ($N = 2628$). Percentage was calculated using the number uncensored patients as the denominator.

We observed imbalance in some of the covariates (Table A.1 and Table A.18). For example, patients who received abiraterone or enzalutamide tend to be older than those receiving docetaxel. Sipuleucel-T users tend to have more pre-treatment osteoporosis (16.0%) than patients receiving the other three drugs (5.3% for docetaxel, 8.4% for abiraterone, and 9.0% for enzalutamide).

To improve the covariate overlap among the treatment groups, we applied data trimming with criteria discussed previously, which left us with 1777 subjects. Results of data analysis are presented in Figure 1.8 and Table A.19. Direct comparison of the four groups (naive method) revealed that docetaxel users had significantly higher risk of at least one ER visits within 180 days of follow up than users of abiraterone (risk difference = 0.130 [0.073, 0.186]), enzalutamide (risk difference = 0.177 [0.115, 0.239]), and sipuleucel-T (risk difference = 0.099 [0.001, 0.197]). The directions of the average effects between docetaxel and the other drugs were preserved for the other methods, though the effect sizes varied. The 95% CIs for the average causal effects between docetaxel and enzalutamide consistently excluded 0 for all methods. However, for the Sipuleucel-T-docetaxel comparison, only MCOV showed a significant difference. For the enzalutamide-abiraterone comparison, all methods considered indicated a higher risk for enzalutamide, while none of these estimated risk differences were significant. For the sipuleucel-T-abiraterone comparison, PEN-GAM yielded negative point estimates (indicating higher risk for abiraterone), while the other methods

indicated a reversed relationship. Again, none of the corresponding CIs excluded 0. In general, there was a larger uncertainty in regard to the direction and magnitude of the risk differences that involve the Sipuleucel-T group due to its smaller sample size. Notably, PEN-GAM tended to have wider CIs than the other methods, which was consistent with the simulation results for small sample size. The results of MW, AMW, OW, and AOW were close to one another in terms of point estimates as well as standard errors for all pairwise comparisons, possibly because their corresponding target populations were similar. This finding aligns with what was observed in the simulation studies. The results of our data analysis agree well with the clinical evidence in current literature [57, 58, 59]. The naive method yielded results that were highly consistent with those of the methods that adjust for potential confounding, suggesting that the treatment effects were relatively strong compared to the confounding effects.

1.7 Discussion

This paper has reviewed and compared a set of causal inference strategies that account for confounding for multiple treatment comparison with a binary outcome variable. Some of these methods, for example, MGPSS [24] and PENCOMP [17], were recently proposed and less explored under the setting of binary outcome in current literature. Our simulation studies show that when there is sufficient overlap in covariate distributions, MGPSS, and in general all AI-type matching methods, are less efficient than the conventional inverse probability weighted (IPW) estimator. The gain in precision of AIPW over IPW that has been observed for continuous outcomes [3, 61] was less evident in our simulations for a binary outcome and good covariate overlap. Thus, while augmentation was still useful for the robustness of estimating the causal effect, it was less useful for improving efficiency. When there was lack of common support, PEN-GAM and AIPW provided more precise estimation than IPW. The improvement in precision increased as the associations of the outcome with baseline covariates became stronger. With moderate outcome prevalence, PEN-GAM tended to perform better than AIPW in terms of RMSE when only the propensity model was correctly specified. One possible reason was that when the covariate overlap is poor, the weights tend to have large variations and some individuals may receive extreme weights, which results in highly variable estimates. PEN-GAM avoids weights by adjusting for the splines of propensity scores (in logit scale) in the outcome model. When the outcome model was misspecified, the estimates relied more on the use of propensity scores. On the other hand, when the outcome was sparse, the fitting of the spline models tended to be unstable, which leads to larger RMSE for PEN-GAM than AIPW.

For propensity score-based methods, correctly modeling the propensity scores is the key to yielding valid inference. The generalized linear model based on maximum likelihood (GLMPS) is

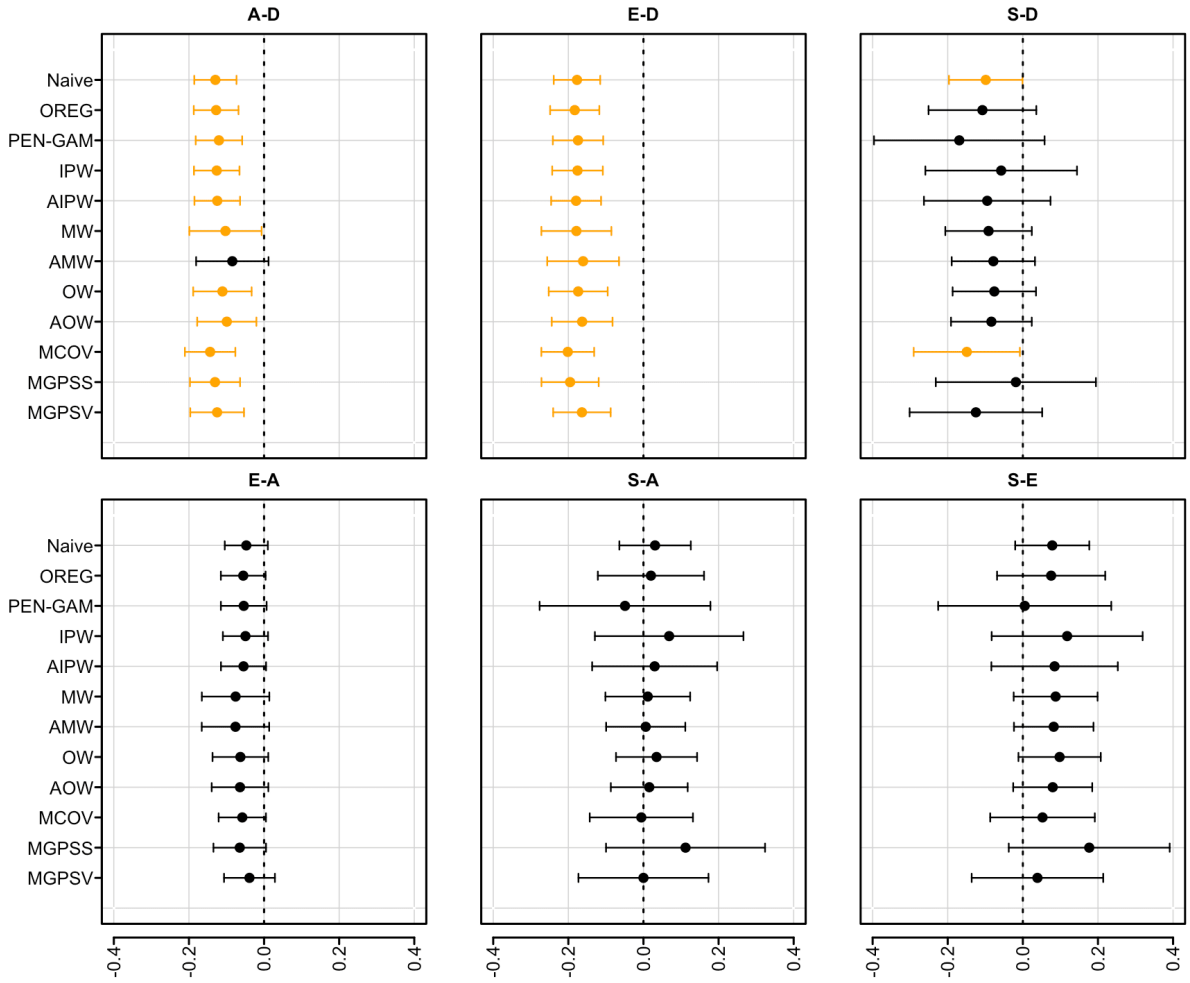

Figure 1.8: Differences in 180-day risks of experiencing at least one emergency room visit among the four focus drugs and the associated 95% confidence intervals. Data were obtained from Optum Clinformative Data Mart, with the outcome interest being the occurrence of emergency room visit within 180 days of treatment initiation. Total sample size is $N = 1777$ ($N_A = 699$, $ND = 519$, $N_E = 438$, $N_S = 121$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

sensitive to both unmeasured confounders and misspecified functional form, which tend to lead to large bias in ATE estimation. Efforts have been made to improve the robustness of propensity score estimation and the Covariate Balancing Propensity Scores (CBPS), which utilizes the covariate balancing property of the propensity scores and achieves robustness in the presence of incorrect functional forms, in one of the examples [45]. In particular, when the GPS model has misspecified functional form but adjusts for the whole set of confounders, the use of CBPS can reduce the bias of the ATE estimates compared to using GLMPS. In addition to CBPS, methods based on machine learning technique have also been proposed for propensity score estimation [62].

Our focus in this paper has remained on simple parametric models. There is extensive literature on using machine learning methods [63, 64, 65] to capture potential nonlinearities and higher-order interactions. The relative gain by using such flexible methods depends on the sample size, the number of predictors, and the true structure of the underlying models (the propensity model or the outcome model).

The computational time for each of the methods considered in the simulation studies for a sample size of 1500 and 3 treatment groups is reported in Table A.20. All simulations were run on an Intel® Xeon® Gold 6138 Processor (2.00 GHz). The average run time of over-identified CBPS was almost twice as much as that of just-identified CBPS. The average run time of PEN-GAM for one bootstrap replicate was around 2 seconds. The projected computational time for 200 bootstrap replicates is approximately 7 minutes.

The methods examined in this study only accounts for the selection bias associated with differences in the covariates. However, the outcome of the data we used is also subject to censoring, which may introduce another layer of selection bias. In particular, approximately 30% of the patients in our data set were censored due to treatment switch or dropout within 180 days of treatment initiation. Weighting-based methods have been proposed to achieve unbiased estimation of average causal effect in the presence of right-censored observations under certain assumptions [66, 67, 68].

CHAPTER 2

An Inverse Probability Weighted Regression Method for a Binary Outcome that Accounts for Right-censoring

2.1 Introduction

Data from observational studies, in which treatments are usually not randomly assigned, have been increasingly used to evaluate comparative effectiveness of treatments in the real world. Without randomization, it is difficult to make causal interpretations concerning the effect of a treatment on an outcome of interest, as the estimation of the average treatment effect tends to be biased by the presence of confounders. A commonly used tool that controls for confounding is the propensity score, defined as the conditional probability of treatment given a set of potential confounders [10]. There is a large body of work on propensity score-based methods that adjust for the differences in baseline covariates among treatment groups, and readers can find some reviews of these methods in Stuart [28], Hu et al. [63], and Yu et al. [69].

It usually takes some period of follow-up time, which may or may not be the same for all subjects, to observe the post-treatment outcome of interest after the subjects enter the study. This type of outcome is sometimes referred to as time-lagged response [68]. One example is whether the event of interest happens within a pre-specified time window, which results in a binary outcome that cannot be ascertained if the subject has been censored prior to the end of the window. Censoring occurs when the information about the response is not completely available due to dropout, study termination, or treatment switch. Censoring together with confounding arises in many applications, such as insurance claims databases, where the participants are not randomly enrolled nor randomly assigned to treatments, and could potentially leave the insurance plan or be switched to another treatment prior to the occurrence of the event of interest. In this paper, we consider a study that evaluates the possible adverse effects of four candidate drugs prescribed for metastatic castration-resistant prostate cancer (mCRPC) using data from Optum Clinformatic Data Mart, a national

private health insurance network. Some of these drugs are chemotherapy, while some others are hormone therapies, and the prior expectation is that the hormone therapies will lead to less acute adverse events. We separately examine two endpoints: emergency room (ER) visit and all-cause hospitalization post first-line therapy with one of these drugs. We intend to compare patients' risks of experiencing at least one event of interest (i.e., ER visit or hospitalization) within a 180-day, 270-day, and 360-day time window after treatment initiation. Therefore, the outcome we focus on is binary that is subject to right-censoring.

Leaving out censored observations when conducting a comparative analysis can result in a biased effect estimate even after proper adjustment for confounding. Special techniques that adjust for both confounders and censoring are required to make valid inference on the causal effects. Anstrom and Tsiatis [68] proposed an inverse probability weighted estimator of average treatment effect for the time-lagged response, which can be used to estimate the difference in risks of event occurrence. To improve the robustness and statistical efficiency of the estimator of Anstrom and Tsiatis [68], Wang et al. [66] proposed an augmented inverse probability weighted estimator that makes use of the information about the outcome model for censored medical cost data. The estimator of Wang et al. [66] can conveniently be extended to the case of a binary outcome by replacing the linear regression model for the outcome with a logistic regression model. Since one minus the survival function of the event of interest describes the risk of an event over the duration of follow-up, one can also use methods for time-to-event outcomes. These will include approaches that model the whole survival curve, and obtain the probability of surviving to the end of the pre-specified time window. For example, Zhang and Schaubel [67] proposed an estimator for the cumulative hazard function, which incorporates the information from Cox models for the time-to-event outcome into the estimating equations. Their method was originally developed for estimation of restricted mean lifetimes, but can be easily adapted to estimate the average treatment effect on a possibly censored binary outcome, since one can think of the survival probability at a given time point as a fixed time 'snapshot' of the whole survival curve. The augmented estimators such as Wang et al. [66] and Zhang and Schaubel [67] are known to possess the double robustness property, such that the estimator remains consistent as long as either the model corresponding to the outcome, or the models corresponding to the weights in the estimating equations are correct. Another line of work in causal inference for censored data involves using pseudo-observations in replacement of the original outcomes that are possibly incompletely observed [70]. For a binary outcome, the causal treatment effect is estimated by computing pseudo survival probability for each subject, followed by standard causal inference method for completely observed outcomes, such as inverse probability weighting or direct standardization. We leave the details of the aforementioned methods to Section 2.4. Among these approaches that address right-censoring, some assume conditional independence of the censoring and survival time given treatment only [e.g.,

66, 68, 70], while others require less restrictive conditions such that censoring and survival times are independent given treatment and baseline covariates [e.g., 67].

We propose a method that directly models the binary outcome using logistic regression, with confounding and censoring properly accounted for by weighting. The risk of event occurrence (and therefore the average treatment effect) is estimated based on standardization, which averages the outcome predictions obtained from the logistic regression model across all subjects. The treatment assignment and censoring mechanism together can be viewed as a special case of coarsening, a process that prevents one from observing the desired data structure [71]. We explain how our problem can be described using the concept of coarsening later in Section 2.3. Coarsening is handled by applying inverse probability of not being coarsened as weights to the score function of the logistic regression model. We call the method inverse probability weighted regression-based estimator, CIPWR for short, with the letter C highlighting the censoring component. Specifically, three sets of working models are constructed, one for the treatment assignment, one for the treatment specific censoring distribution, and the other for the outcome of interest. We show that the CIPWR estimator is doubly robust in the sense that consistency of the estimator can be achieved if either the outcome or the coarsening mechanism is correctly modeled. As Zhang and Schaubel [67], this method makes the less restrictive assumption about censoring than Wang et al. [66], Anstrom and Tsiatis [68], Andersen et al. [70]. Unlike Wang et al. [66], Zhang and Schaubel [67] that are based on the general approach of augmented inverse probability weighting, our method is a standardization method. Also, unlike Zhang and Schaubel [67] that estimates the whole survival curve, this method targets the binary outcome of interest and may lead to improved efficiency in some situations.

The rest of the paper is organized as follows. In Section 2.2, we introduce the statistical framework and the notations used. In Section 2.3, we describe our proposed method and establish its asymptotic properties. We compare the finite sample performance of the CIPWR estimator to that of several alternative approaches through simulation studies and results are presented in Section 2.5. The proposed method is then applied to the prostate cancer treatment comparison example from the insurance claims database in Section 2.6. Conclusions and discussions for future research are presented in Section 2.7.

2.2 Notations and Assumptions

For individual i, where $i = 1, \cdots, n$, let $\tilde{\boldsymbol{X}}_i$ be a set of baseline variables, and Z_i be the treatment received. We assume that Z_i is nominal with J levels, i.e., $Z_i = j \in \{1, \cdots, J; J \geq 2\}$, and let $D_{ij} \equiv I(Z_i = j)$. Let T_i denote the underlying lag time to the first event of interest, which will always be observed if there were no censoring. In this study, the outcome of interest, denoted by

Y_i, is whether the event of interest occurs within a pre-specified time window d. By this definition, $Y_i = I(T_i < d)$. We adopt the counterfactual framework to formulate the problem of causal comparison [41]. Each individual is associated with a set of potential outcomes $\{Y_i^{(1)}, \cdots, Y_i^{(J)}\}$, where $Y_i^{(j)} = I\{T_i^{(j)} < d\}$ and $T_i^{(j)}$ is defined as the potentially observed time to the first event of interest had the patient received treatment j. Under the Stable Unit Treatment Value Assumption (SUTVA, defined later), only the outcome under the actual treatment received, $Y_i = \sum_{j=1}^{J} D_{ij} Y_i^{(j)}$, can be observed.

In practice, the time to event T_i may not be completely observed due to right-censoring, in which case the outcome variable Y_i is therefore subject to coarsening. Let C_i denote the censoring time and $R_i = I\{C_i \geq \min(T_i, d)\}$. Then Y_i is observed if the individual has not been censored before d, i.e., $R_i = 1$. We further let $\Delta_i = I(T_i \leq C_i)$ and $L_i = \min(T_i, C_i, d)$. Note that the outcomes Y_i of those whose T_i are censored ($\Delta_i = 0$) are not necessarily missing at time d ($R_i = 0$).

Interest lies in estimating the average treatment effect $\tau(j, j') = E\left\{Y^{(j')} - Y^{(j)}\right\}$, which equals the risk difference $pr\{Y^{(j')} = 1\} - pr\{Y^{(j)} = 1\}$ for a binary outcome. We seek to estimate $E\left[Y^{(j)}\right]$ separately for $j = 1, \cdots, J$. To connect the counterfactual framework to the observable data and establish a causal interpretation, we make the following assumptions.

(A1) (*Random sampling*) The individuals in the study are randomly sampled from the population.

(A2) (*Stable Unit Treatment Value Assumption, or SUTVA*) For any individual i, $i = 1, \cdots, n$, if $Z_i = j$, then $Y_i = Y_i^{(j)}$, for all $j = 1, \cdots, J$.

(A3) (*Unconfoundedness*) $\{Y_i^{(1)}, \cdots, Y_i^{(J)}\} \perp\!\!\!\perp Z_i | \tilde{\boldsymbol{X}}_i$.

(A4) (*Overlap*) For all values of j and $\tilde{\boldsymbol{x}}$, $0 < \pi_j(\tilde{\boldsymbol{x}}) < 1$, where $\pi_j(\tilde{\boldsymbol{x}}) = pr(Z_i = j | \tilde{\boldsymbol{x}})$.

(A5) (*Censoring at random*) $C_i \perp\!\!\!\perp \{T_i^{(1)}, \cdots, T_i^{(J)}\} \Big| (Z_i, \tilde{\boldsymbol{X}}_i)$.

2.3 Proposed Method: Inverse Probability Weighted Regression that Accounts for Right-Censoring

We note that instead of directly evaluating $E\{Y_i^{(j)}\}$, it is theoretically more convenient to work with the survival function

$$\mu_j \equiv E[I\{T_i^{(j)} \geq d\}] = 1 - E\{Y_i^{(j)}\},$$

and we let $\tilde{Y}_i^{(j)} = I\{T_i^{(j)} \geq d\}$. The counterfactual parameter μ_j can be represented using the observed data, $\mu_j = E_X[E\{\tilde{Y}_i^{(j)} | \tilde{\boldsymbol{X}}_i\}] = E_X[E\{\tilde{Y}_i | \tilde{\boldsymbol{X}}_i, Z_i = j\}]$, where the second equation

follows from the unconfoundedness assumption (A3). Had there been no right-censoring, μ_j could be estimated by averaging the predicted potential outcomes, $\hat{\mu}_j = n^{-1}\sum_{i=1}^{n}\hat{E}(\tilde{Y}_i|\tilde{\boldsymbol{X}}_i, Z_i = j)$, for $j = 1, \cdots, J$, where $\hat{E}(\tilde{Y}_i|\tilde{\boldsymbol{X}}_i, Z_i = j)$ are usually fitted values in a parametric regression model. For a binary outcome, a logistic regression model is a popular choice to fit $\tilde{Y}$ in group j, specified as

$$\text{logit}\{E(\tilde{Y}_i|\tilde{\boldsymbol{X}}_i, Z_i = j)\} = \boldsymbol{X}_i^T\boldsymbol{\beta}_j, \tag{2.1}$$

where $\boldsymbol{X}_i$ is a vector-valued function of $\tilde{\boldsymbol{X}}_i$ with an intercept and possibly interactions and non-linear terms. For notational convenience, we define $m_{ij}(\boldsymbol{\beta}_j) = \text{expit}(\boldsymbol{X}_i^T\boldsymbol{\beta}_j)$. When the outcome Y (and therefore $\tilde{Y}$) is completely observed for all individuals in the sample, $\boldsymbol{\beta}_j$ is commonly estimated by solving the score equations

$$\sum_{i=1}^{n}\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j)\} = 0, \tag{2.2}$$

the solution of which is the maximum likelihood estimator.

From the missing data perspective, the potential outcome $Y_i^{(j)} = I\{T_i^{(j)} < d\}$ would be missing at baseline ($t = 0$) if individual i were assigned to the treatment other than j. When censoring comes into play, $Y_i^{(j)}$ is subject to missingness at any time $0 < t < d$ because of censoring. In this case, we consider the more general notion of coarsening of data [71, 72, 73], which describes the case where one only gets to observe a many-to-one function of the full data for some of the individuals in the sample, and different many-to-one functions are allowed for different individuals. In the context of estimating μ_j, the full data one would like to observe for individual i is $\{Y_i^{(j)}, \tilde{\boldsymbol{X}}_i\}$. When $D_{ij} = 0$, $T_i^{(j)}$ (and therefore $Y_i^{(j)}$) is completely missing, and we only observe $\tilde{\boldsymbol{X}}_i$. When $D_{ij} = 1$ and $C_i = t < \min(T_i^{(j)}, d)$, we observe $\{I(T_i^{(j)} > t), \tilde{\boldsymbol{X}}_i\}$, where $t < d$. When $D_{ij} = 1$ and $C_i = t \geq \min(T_i^{(j)}, d)$, there was no coarsening at all, and we observe the full data $\{Y_i^{(j)}, \tilde{\boldsymbol{X}}_i\}$. In summary, there are two layers of missingness in our setting, one due to treatment assignment and the other due to censoring. Therefore, we can inversely weight the unbiased estimating equations (2.2) by the probability of not being coarsened to make inference about the target population where no subjects were coarsened. The weighted estimating equations for $\boldsymbol{\beta}_j$ is given by

$$\sum_{i=1}^{n}\frac{D_{ij}R_i\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j)\}}{\pi_{ij}(\boldsymbol{\alpha})\exp\{-\Lambda_{ij}(L_i)\}} = 0, \tag{2.3}$$

where $\pi_{ij}(\boldsymbol{\alpha}) = pr(Z_i = j|\tilde{\boldsymbol{X}}_i)$ is the propensity score for treatment j, and $\Lambda_{ij}(t)$ is the cumulative hazard function of C_i at t for treatment j. We denote the solution to (2.3) by $\hat{\boldsymbol{\beta}}_j$. The total weights

that account for the coarsening mechanism consist of two weighting components. The first is the propensity of being assigned to treatment j, and the second is (informally) the conditional probability of not being censored. Therefore, equation (2.3) can be regarded as weighting the score functions (2.2) by the inverse probability of observing the complete cases. A regression-based estimator for μ_j indicated by (2.1) is

$$\hat{\mu}_j = n^{-1} \sum_{i=1}^{n} \text{expit}(\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_j) = n^{-1} \sum_{i=1}^{n} m_{ij}(\hat{\boldsymbol{\beta}}_j),$$

and we call it inverse probability weighted regression-based estimator that accounts for right-censoring (CIPWR).

In the literature, an outcome that is subject to censoring is sometimes handled using survival models, such as Cox regression model [67, 74]. CIPWR, on the other hand, directly models the binary outcome of interest using the logistic regression, which is relatively more intuitive and straightforward to implement for empirical researchers. Our proposed estimator can be implemented using standard statistical software, such as the *glm* function in R with the *weights* argument being specified.

In practice, $\pi_{ij}(\boldsymbol{\alpha})$ and $\Lambda_{ij}(t)$ in (2.3) are usually unknown and need to be estimated from the data. We build working models for these two nuisance components. Let $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ be vector-valued functions of $\tilde{\boldsymbol{X}}_i$, which are allowed to be different from $\boldsymbol{X}_i$. We assume that the treatment assignment mechanism is governed by a multinomial logistic regression model

$$\log \frac{pr(Z_i = j|\boldsymbol{V}_i)}{pr(Z_i = J|\boldsymbol{V}_i)} = \boldsymbol{V}_i^T \boldsymbol{\alpha}_j, \;\; j = 1, \cdots, J-1,$$

where J is the reference treatment level. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{J-1})^T$, and its estimated value $\hat{\boldsymbol{\alpha}}$ can be obtained through maximum likelihood estimation. With respect to censoring, for each treatment $j = 1, \cdots, J$, we assume a Cox proportional hazards model, specified as

$$\lambda_{ij}(t|\boldsymbol{W}_i, \boldsymbol{\gamma}_j) = \lambda_{0j}(t) \exp(\boldsymbol{W}_i^T \boldsymbol{\gamma}_j),$$

where $\lambda_{0j}(t)$ is an unspecified treatment-specific baseline hazard function of C. The estimates for $\boldsymbol{\gamma}_j$ and $\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(s)ds$ can be determined by the maximum partial likelihood estimator, $\hat{\boldsymbol{\gamma}}_j$, and the Breslow estimator, $\hat{\Lambda}_{0j}(t)$, respectively. Then the probability of remaining uncensored at t for individual i is given by $\exp\{-\hat{\Lambda}_{0j}(t) \exp(\boldsymbol{W}_i^T \hat{\boldsymbol{\gamma}}_j)\}$. In a typical survival study, one only gets to observe the minimum of C and T, which is usually referred to as *observation time*, and the observed data can be represented as $\{\Delta_i, \min(T_i, C_i)\}$. However, for our data example, time to treatment switch or the end of insurance coverage can always be identified from the claims data,

regardless of whether the event of interest happens or not. In this case, censoring time C is always available for all subjects, and $\Delta_i = 0$ for all $i, i = 1, \cdots, n$. Therefore, one can alternatively estimate the probability of remaining uncensored by replacing $\{\Delta_i, \min(T_i, C_i)\}$ with $\{0, C_i\}$.

2.3.1 Consistency and Double Robustness

Under suitable regularity conditions, $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}_j$, and $\hat{\Lambda}_{0j}$ converge in probability to well-defined limits, denoted by $\boldsymbol{\alpha}^*$, $\boldsymbol{\gamma}_j^*$, and $\boldsymbol{\Lambda}_{0j}^*$, respectively, which can be different from their corresponding true values $\boldsymbol{\alpha}^0$, $\boldsymbol{\gamma}_j^0$, and $\boldsymbol{\Lambda}_{0j}^0$ [75, 76]. We denote the true values for $\boldsymbol{\beta}_j$ and μ_j by $\boldsymbol{\beta}_j^0$ and μ_j^0, respectively. For notational convenience, we also define $\Lambda_{ij}^*(t) = \Lambda_{0j}^*(t)\exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}_j^*)$ and $\Lambda_{ij}^0(t) = \Lambda_{0j}^0(t)\exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}_j^0)$.

We first show that $\hat{\mu}_j = n^{-1}\sum_{i=1}^n m_{ij}(\hat{\boldsymbol{\beta}}_j)$, a function of $\hat{\boldsymbol{\beta}}_j$, is consistent when the outcome model for treatment j is correct. Using the theory of M-estimator, the consistency of $\hat{\boldsymbol{\beta}}_j$ can be established by showing that the estimating function is unbiased [77], that is,

$$
\begin{aligned}
0 =& E\left[\frac{D_{ij}R_i\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j^0)\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}}\right] \\
=& E\left[\frac{D_{ij}R_i\boldsymbol{X}_i\tilde{Y}_i}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}}\right] && \text{(2.4a)} \\
& - E\left[\frac{D_{ij}R_i\boldsymbol{X}_i m_{ij}(\boldsymbol{\beta}_j^0)}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}}\right]. && \text{(2.4b)}
\end{aligned}
$$

Applying the law of iterated expectation and using the Assumption (A5),

$$
\begin{aligned}
(2.4a) &= E\left\{E\left[\frac{D_{ij}R_i\boldsymbol{X}_i\tilde{Y}_i}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}}\middle|\boldsymbol{X}_i, Z_i = j\right]\right\} \\
&= E\left[\frac{D_{ij}\boldsymbol{X}_i E\left\{I(C_i > d)|\boldsymbol{X}_i, Z_i = j\right\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(d)\}}E(\tilde{Y}_i|\boldsymbol{X}_i, Z_i = j)\right],
\end{aligned}
$$

where the second equation is derived from the formula $R_i\tilde{Y}_i = I\{C_i > \min(T_i, d)\}\tilde{Y}_i = I\{C_i > d\}\tilde{Y}_i$. Since when $T_i > d$, $R_i/\exp\{-\Lambda_{ij}^*(L_i)\} = I(C_i > d)/\exp\{-\Lambda_{ij}^*(d)\}$, using similar tech-

niques,

$$(2.4b) = E\left\{E\left[\frac{D^{(j)}R_i\boldsymbol{X}_i m_{ij}(\boldsymbol{\beta}_j^0)}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}}\middle|\boldsymbol{X}_i, Z_i = j, T_i > d\right]\right\}$$
$$= E\left\{\frac{D_{ij}\boldsymbol{X}_i E\left[I(C_i > d)|\boldsymbol{X}_i, Z_i = j\right]}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(d)\}} m_{ij}(\boldsymbol{\beta}_j^0)\right\}.$$

Since $E(\tilde{Y}_i|\boldsymbol{X}_i, Z_i = j) = m_{ij}(\boldsymbol{\beta}_j^0)$ when the outcome model is correctly specified, the estimating function is shown to be unbiased, which implies that $\hat{\boldsymbol{\beta}}_j$ obtained by solving (2.3) converges in probability to the truth $\boldsymbol{\beta}_j^0$. Therefore, $\hat{\mu}_j = n^{-1}\sum_{i=1}^n m_{ij}(\hat{\boldsymbol{\beta}}_j) \xrightarrow{p} E\left\{m_{ij}\left(\boldsymbol{\beta}_j^0\right)\right\} = \mu_j^0$.

We then show the consistency of $\hat{\mu}_j$ when the coarsening mechanisms (i.e., treatment and censoring models) are correctly specified, in which case $\pi_{ij}(\hat{\boldsymbol{\alpha}}) \xrightarrow{p} \pi_{ij}(\boldsymbol{\alpha}^0)$ and $\hat{\Lambda}_{ij}(t) \xrightarrow{p} \Lambda_{ij}^0(t)$. Under suitable regularity conditions, $\hat{\boldsymbol{\beta}}_j \xrightarrow{p} \boldsymbol{\beta}_j^*$, where $\boldsymbol{\beta}_j^*$ is a well-defined limit, and then

$$\hat{\mu}_j = n^{-1}\sum_{i=1}^n m_{ij}\left(\hat{\boldsymbol{\beta}}_j\right) \xrightarrow{p} E\left\{m_{ij}\left(\boldsymbol{\beta}_j^*\right)\right\}.$$

We consider the intercept term in $\boldsymbol{X}_i$ and rearrange equation (2.3),

$$n^{-1}\sum_{i=1}^n \frac{D_{ij}R_i\tilde{Y}_i}{\pi_{ij}(\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}_{ij}(L_i)\}} = n^{-1}\sum_{i=1}^n \frac{D_{ij}R_i m_{ij}(\hat{\boldsymbol{\beta}}_j)}{\pi_{ij}(\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}_{ij}(L_i)\}}. \tag{2.5}$$

The left-hand side of (2.5) converges in probability to μ_j^0, because

$$\begin{aligned} n^{-1}\sum_{i=1}^n \frac{D_{ij}R_i\tilde{Y}_i}{\pi_{ij}(\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}_{ij}(L_i)\}} &\xrightarrow{p} E\left[\frac{D_{ij}R_i\tilde{Y}_i}{\pi_{ij}(\boldsymbol{\alpha}^0)\exp\{-\Lambda_{ij}^0(L_i)\}}\right] \\ &= E\left[\frac{D_{ij}E\{I(C_i > d)\tilde{Y}_i^{(j)}|\tilde{\boldsymbol{X}}_i, Z_i = j\}}{\pi_{ij}(\boldsymbol{\alpha}^0)\exp\{-\Lambda_{ij}^0(d)\}}\right] \\ &= E\left[\frac{D_{ij}E\{I(C_i > d)|\tilde{\boldsymbol{X}}_i, Z_i = j\}E\{\tilde{Y}_i^{(j)}|\tilde{\boldsymbol{X}}_i, Z_i = j\}}{\pi_{ij}(\boldsymbol{\alpha}^0)\exp\{-\Lambda_{ij}^0(d)\}}\right] \end{aligned} \tag{2.6}$$

where (2.6) follows from Assumption (A5). With correct specification of the treatment and censoring models, $E\{I(C_i > d)|\tilde{\boldsymbol{X}}_i, Z_i = j\} = \exp\{-\Lambda_{ij}^0(d)\}$ and $E(D_{ij}|\tilde{\boldsymbol{X}}_i) = \pi_{ij}(\boldsymbol{\alpha}_0)$, and therefore (2.6) can be reduced to $E[E\{\tilde{Y}_i^{(j)}|\tilde{\boldsymbol{X}}_i, Z_i = j\}]$, which is equivalent to μ_j^0.

Using similar techniques, one can show that the right-hand side of (2.5) converges in probability

to $E\left\{m_{ij}(\boldsymbol{\beta}_j^*)\right\}$, since

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n}\frac{D_{ij}R_i m_{ij}(\hat{\boldsymbol{\beta}}_j)}{\pi_{ij}(\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}_{ij}(L_i)\}} &\xrightarrow{p} E\left[\frac{D_{ij}R_i m_{ij}(\boldsymbol{\beta}_j^*)}{\pi_{ij}(\boldsymbol{\alpha}^0)\exp\{-\Lambda_{ij}^0(L_i)\}}\right] \\
&= E\left[\frac{D_{ij}m_{ij}(\boldsymbol{\beta}_j^*)E\{I(C_i>d)|\boldsymbol{X}_i, Z_i=j, T_i>d\}}{\pi_{ij}(\boldsymbol{\alpha}^0)\exp\{-\Lambda_{ij}^0(d)\}}\right] \\
&= E\left\{m_{ij}(\boldsymbol{\beta}_j^*)\right\}.
\end{aligned}
$$

It follows that $\mu_j^0 = E\left\{m_{ij}(\boldsymbol{\beta}_j^*)\right\}$, and $\hat{\mu}_j \xrightarrow{p} \mu_j^0$ when the treatment and censoring models are correctly specified. Note that when the stronger independence assumption ($C \perp\!\!\!\perp T|Z$) holds for survival and censoring times, only the treatment model is required to be correct. We have shown that the proposed estimator exhibits the so-called double robustness property.

2.3.2 Asymptotic Properties

In this section, we establish the asymptotic properties of our proposed estimator $\hat{\mu}_j$. For $j = 1, \cdots, J$, through a Taylor series expansion of $\hat{\mu}_j = n^{-1}\sum_{i=1}^{n} m_{ij}(\hat{\boldsymbol{\beta}}_j)$ about $\boldsymbol{\beta}_j^*$,

$$
n^{1/2}(\hat{\mu}_j - \mu_j^0) = n^{-1/2}\sum_{i=1}^{n}\left\{m_{ij}(\boldsymbol{\beta}_j^*) - \mu_j^0\right\} + \boldsymbol{A}_j(\boldsymbol{\beta}_j^*)n^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) + o_p(1), \tag{2.7}
$$

where $\boldsymbol{A}_j(\boldsymbol{\beta}_j^*) = E\left[\boldsymbol{X}_i^T m_{ij}(\boldsymbol{\beta}_j^*)\{1 - m_{ij}(\boldsymbol{\beta}_j^*)\}\right]$.

Equation (2.7) indicates that to characterize the asymptotic distribution of $n^{1/2}(\hat{\mu}_j - \mu_j^0)$, one first needs to identify the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)$, which further depends on the asymptotic results for the parameters of the treatment and censoring models. Under some suitable regularity conditions, $\hat{\boldsymbol{\alpha}}_l \xrightarrow{p} \boldsymbol{\alpha}_l^*$ for $l = 1, \cdots, J-1$, and the estimator of the treatment model parameter is asymptotically normal with

$$
n^{1/2}(\hat{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_l^*) = \boldsymbol{H}_l^{-1}(\boldsymbol{\alpha}^*)n^{-1/2}\sum_{i=1}^{n}\boldsymbol{V}_i\left\{D_{il} - \pi_{il}(\boldsymbol{\alpha}^*)\right\} + o_p(1), \tag{2.8}
$$

where $\boldsymbol{H}_l(\boldsymbol{\alpha}^*) = E\left[\sum_{i=1}^{n}\boldsymbol{V}_i\boldsymbol{V}_i^T\pi_{il}(\boldsymbol{\alpha}^*)\left\{1 - \pi_{il}(\boldsymbol{\alpha}^*)\right\}\right]$ with $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_{J-1}^*)^T$.

For the asymptotic distributions of the estimators $\hat{\gamma}_j$ and $\hat{\Lambda}_{ij}$, we define the relevant notations $\boldsymbol{s}_j^{(q)}(t;\boldsymbol{\gamma}_j)$ for $q = 0, 1, 2$, $\overline{\boldsymbol{w}}_j(t;\boldsymbol{\gamma}_j)$, $d\Lambda_{0j}^*(t)$, and $dM_{ij}^*(t)$ in section B.1 in Appendix B. We further denote the counting process by $N_{ij}(t) = D_{ij}I\{\min(T_i, C_i) \le t, \Delta_i = 1\}$ and the at-risk process by $Y_{ij}(t) = D_{ij}I\{\min(T_i, C_i) \ge t\}$. Let δ be the time point that satisfies $P\{\min(T_i, C_i) \ge \delta\} > 0$ for $i = 1, \cdots, n$, which is practically set to the maximum observation time. Lin and Wei [76] showed that under some regularity conditions, $\hat{\boldsymbol{\gamma}}_j \xrightarrow{p} \boldsymbol{\gamma}_j^*$, and $n^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*)$ converges in distribution to a

normal distribution

$$n^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*) = \boldsymbol{\Omega}_j^{-1}(\boldsymbol{\gamma}_j^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) + o_p(1), \tag{2.9}$$

where $\boldsymbol{\Omega}_j(\boldsymbol{\gamma}_j^*) = \int_0^\delta \left\{ \frac{\boldsymbol{s}_j^{(2)}(t;\boldsymbol{\gamma}_j^*)}{s_j^{(0)}(t;\boldsymbol{\gamma}_j^*)} - \overline{\boldsymbol{w}}_j(t;\boldsymbol{\gamma}_j)^{\otimes 2} \right\} E\{Y_{ij}(t)\lambda_{ij}(t)\} dt$ and $\boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) = \int_0^\delta \{\boldsymbol{W}_i - \overline{\boldsymbol{w}}(t;\boldsymbol{\gamma}_j^*)\} dM_{ij}^*(t)$. Using (2.9), one can show that

$$\begin{aligned} n^{1/2}\{\hat{\Lambda}_{ij}(t) - \Lambda_{ij}^*(t)\} = \boldsymbol{K}_{ij}^T(t;\boldsymbol{\gamma}_j^*) \boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma}_j^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) \\ + \exp(\boldsymbol{W}_i^T \boldsymbol{\gamma}_j^*) n^{-1/2} \sum_{i=1}^{n} \int_0^t \frac{dM_{ij}^*(u)}{s^{(0)}(u;\boldsymbol{\gamma}_j^*)} + o_p(1), \end{aligned} \tag{2.10}$$

where $\boldsymbol{K}_{ij}(t;\boldsymbol{\gamma}_j^*) = \int_0^t \{\boldsymbol{W}_i - \overline{\boldsymbol{w}}_j(t;\boldsymbol{\gamma}_j^*)\} d\Lambda_{ij}^*(u)$.

By a sequence of Taylor series expansion of $n^{-1}\sum_{i=1}^n \frac{D_{ij}R_i\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\hat{\boldsymbol{\beta}}_j)\}}{\pi_{ij}(\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}_{ij}(L_i)\}}$ (see section B.1 in the Appendix) and combining the results of (2.8), (2.9), and (2.10), it follows that

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) = \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) n^{-1/2} \sum_{i=1}^{n} \frac{D_{ij}R_i\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j^*)\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}} \\ + \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \sum_{l=1}^{J-1} \left[\boldsymbol{F}_{jl}\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*\right) \boldsymbol{H}_l^{-1}(\boldsymbol{\alpha}^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{V}_i\{D_{il} - \pi_{il}(\boldsymbol{\alpha}^*)\} \right] \\ + \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \boldsymbol{P}_j(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \boldsymbol{\Omega}_j^{-1}(\boldsymbol{\gamma}_j^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) \\ + \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \boldsymbol{Q}_j(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}, \Lambda_{ij}^*) n^{-1/2} \sum_{i=1}^{n} \int_0^t \frac{dM_{ij}^*(u)}{s^{(0)}(u;\boldsymbol{\gamma}_j^*)} + o_p(1). \end{aligned} \tag{2.11}$$

where $\boldsymbol{B}_j$, $\boldsymbol{F}_{jl}$, $\boldsymbol{P}_j$, and $\boldsymbol{Q}_j$ are defined in section B.1 in Appendix B.

Plugging (2.11) into (2.7), we can represent $n^{1/2}(\hat{\mu}_j - \mu_j)$ as $n^{-1/2}\sum_{i=1}^n \psi_{ij} + o_p(1)$, where

$$\begin{aligned} \psi_{ij} =& m_{ij}(\boldsymbol{\beta}_j^*) - \mu_j + \boldsymbol{A}_j(\boldsymbol{\beta}_j^*) \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \frac{D_{ij}R_i\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j^*)\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda_{ij}^*(L_i)\}} \\ &+ \boldsymbol{A}_j(\boldsymbol{\beta}_j^*) \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \sum_{l=1}^{J-1} \boldsymbol{F}_{jl}\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*\right) \boldsymbol{H}_l^{-1}(\boldsymbol{\alpha}^*) \boldsymbol{V}_i\{D_{il} - \pi_{il}(\boldsymbol{\alpha}^*)\} \\ &+ \boldsymbol{A}_j(\boldsymbol{\beta}_j^*) \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \boldsymbol{P}_j(\boldsymbol{\beta}_j, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \boldsymbol{\Omega}_j^{-1}(\boldsymbol{\gamma}_j^*) \boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) \\ &+ \boldsymbol{A}_j(\boldsymbol{\beta}_j^*) \boldsymbol{B}_j^{-1}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \boldsymbol{Q}_j(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) \int_0^{L_i} \frac{dM_{ij}^*(u)}{s^{(0)}(u;\boldsymbol{\gamma}_j^*)}. \end{aligned}$$

By the central limit theorem, $n^{-1/2}\sum_{i=1}^{n}\psi_{ij}$ converges in distribution to a normal distribution with mean 0 and variance $E(\psi_{ij}^2)$.

2.4 Methods under Comparison

We compare our proposed method with several alternative approaches. The first is to leave out censored subjects and apply standard inverse probability weighted (IPW) method to the data with completely observed outcome only. The corresponding estimator for the average potential outcome in treatment group j is given by

$$\hat{\mu}_{j,\text{IPW}} = n^{-1}\sum_{i=1}^{n}\frac{D_{ij}R_i\tilde{Y}_i}{\pi_{ij}(\hat{\boldsymbol{\alpha}})}.$$

The second approach considered builds on the IPW estimator and, along the line of Anstrom and Tsiatis [68], further weights the subjects by the inverse probability of not being coarsened, namely, CIPW estimator

$$\hat{\mu}_{j,\text{CIPW}} = n^{-1}\sum_{i=1}^{n}\frac{D_{ij}R_i\tilde{Y}_i}{\pi_{ij}(\hat{\boldsymbol{\alpha}})\exp\{-\hat{\Lambda}_{ij}(L_i)\}}.$$

The third is the estimator of [66], which is a doubly robust estimator for average treatment effect using an augmented inverse probability weighted method, and we label it CAIPW-Wang. Let $h_{ij}(\boldsymbol{\omega}_j)$ be a posited model, in this case a logistic regression model, for $E(\tilde{Y}_i|Z_i = j, \tilde{\boldsymbol{X}}_i)$. The estimates for the parameter $\boldsymbol{\omega}_j$, denoted by $\hat{\boldsymbol{\omega}}_j$, are obtained by solving the score functions weighted by the inverse probability of not being censored. The final estimator is given by

$$\hat{\mu}_{j,\text{CAIPW-Wang}} = \left(\sum_{i=1}^{n}w_i\right)^{-1}\sum_{i=1}^{n}w_i\left\{\frac{D_{ij}\tilde{Y}_i}{\pi_{ij}(\hat{\boldsymbol{\alpha}})} - \frac{D_{ij}-\pi_{ij}(\hat{\boldsymbol{\alpha}})}{\pi_{ij}(\hat{\boldsymbol{\alpha}})}h_{ij}(\hat{\boldsymbol{\omega}}_j)\right\},$$

where $w_i = \sum_{i=1}^{n}\{\Delta_i/\sum_{j=1}^{J}D_{ij}\hat{K}_j[\min(T_i,C_i)]\}$ and $\hat{K}_j(t)$ is the treatment-specific Kaplan-Meier (KM) estimator.

The fourth is to apply standard causal inference methods, such as IPW, to pseudo-values of the outcome [70], and we call it Pseudo-IPW. Suppose that the parameter of interest is $\theta = E\{I(T_i \geq d)\}$. The pseudo-observation for subject i is defined as $\theta_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$, where $\hat{\theta}$ is the KM estimator and $\hat{\theta}^{-i}$ is the estimator applied to the sample from which subject i is excluded. The method is implemented using the *pseudo* package in R [78]. The pseudo-observations can be viewed as a replacement for the (possibly incompletely observed) outcome variable, and censoring is taken care of in the computation of pseudo-observations. In this case, the pseudo-observations

are calculated assuming the independence of T and C given Z.

The fifth is the estimator of Zhang and Schaubel [67], which is originally designed for estimating the restricted mean lifetimes, and involves modeling the entire survival curve of the event time using Cox proportional hazards models, rather than focusing on a fixed time point as in the aforementioned approaches and our proposed method. It first estimates the cumulative hazard function for the event time T for each group j, denoted by $\hat{\Lambda}_j(t)$, by augmenting an inverse probability weighted estimating equation with additional terms that involve outcome models. Then one can estimate the survival probability $\mu_j(t)$ at any time point t by $\hat{\mu}_j(t) = e^{-\hat{\Lambda}_j(t)}$. Therefore, the method of Zhang and Schaubel [67] can also be used for evaluating the risk at a specific time point d. We label their approach based on the inverse probability weighted estimating function as CIPW-ZS, and the augmented version as CAIPW-ZS. CAIPW-ZS is doubly robust under Assumption (A5), such that the estimator is consistent when either the time-to-event or the coarsening mechanism is correctly modeled.

The method of Zhang and Schaubel [67] is developed within the general framework of augmented inverse probability weighting, whereas the proposed method is a standardization method. Another key difference between these two methods is that the former uses the Cox models but the proposed method uses logistic regression models as working models to improve efficiency of the treatment effect estimator. If the interest only lies in a binary outcome (i.e., the occurrence of the event within a specific time period), theoretically one only needs to model the relationship of the binary outcome with covariates to improve efficiency. The method of Zhang and Schaubel [67] requires modeling the relationship of the hazard function (equivalently, the survival curve), not limited to a specific time point, with covariates. This tends to be an overkill for our purpose and increase the chance of model misspecification. When the Cox model is severely misspecified, the method of Zhang and Schaubel [67] may provide little efficiency gain for the treatment effect estimates. We illustrate this point in one of our simulation settings.

We further consider the simple difference in the average outcome of each treatment group, as a benchmark, and call it the Naive estimator. The Naive estimator ignores both confounding and censoring. Sample code for the methods described in this section can be found at https://github.com/youfeiyu/CIPWR.

Among these methods, Naive and IPW estimators fail to account for censoring. Pseudo-IPW and AIPW-Wang assume that T and C are independent conditional on Z. CIPW-ZS and CAIPW-ZS, along with our proposed method CIPWR, rely on a more relaxed assumption that T and C are independent conditional on Z and $\tilde{\boldsymbol{X}}$. CAIPW-Wang, CAIPW-ZS, and CIPWR leverage the information about the outcome model, which asymptotically improves the precision of the estimates. Moreover, the double robustness property of CAIPW-Wang and CAIPW-ZS are provided by the augmentation terms in the estimating equations, while the proposed CIPWR achieves double

robustness through standardization of the weighted outcome model.

2.5 Simulation Studies

We compared the finite sample performance of our proposed method to the six alternative approaches described in Section 2.4 through simulation studies. Specifically, we considered two settings that varied in degrees of nonproportionality with respect to the hazard functions for our simulation. The first setting assumed a logistic distribution for the time to event such that the hazard functions did not cross. The second setting concerned hazard functions that crossed at a certain time point for subjects with different covariate values, in which case Cox proportional hazards model tended to perform poorly in terms of improving precision.

2.5.1 Simulation Setting I: Non-crossing Hazards

For the first setting, each simulated data set contained five baseline covariates. X_1, X_2, and X_3 were independently sampled from a standard normal distribution. $X_4 \sim$ Bernoulli(0.4) and $X_5 \sim$ Uniform$(-2, 2)$. The treatment assignment Z was simulated from a categorical distribution with the probability of receiving treatment j being

$$\frac{\exp(\alpha_{j0} + \alpha_{j1}X_1 + \alpha_{j2}X_2 + \alpha_{j4}X_4)}{\sum_{z=1}^{3} \exp(\alpha_{z0} + \alpha_{z1}X_1 + \alpha_{z2}X_2 + \alpha_{z4}X_4)}$$

for $j = 1, 2, 3$. The potential time to event $T^{(j)}$ was sampled from a logistic distribution with mean function $\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \beta_{j3}X_3$ and scale parameter $s = 7$. The potential outcome $Y^{(j)}$ is defined as $Y^{(j)} = I\{T^{(j)} < d\}$, where $d = 130$. We generated the censoring time C using inverse transform sampling [79]. In particular, we assumed a Cox proportional hazard model with the baseline hazard following a Weibull distribution,

$$C^{(j)} = \{\lambda^{-1} \exp(\gamma_{j0} + \gamma_{j1}X_1 + \gamma_{j2}X_2 + \gamma_{j5}X_5)^{-1} \log u\}^{1/\nu},$$

where the scale parameter $\lambda = 0.01$, the shape parameter $\nu = 7$, and u was randomly sampled from a uniform distribution with interval [0,1].

We defined a 'baseline' scenario where the outcome was weakly associated with the covariates and the proportion of being censored by $d = 130$ was 30%. Then we varied the corresponding parameters to induce three proportions of being censored by $d = 130$ (20%, 30%, and 40%) and two levels of associations with the outcome (strong and weak). We also considered a scenario where censoring was independent of the covariates, referring to it as the random censoring scenario, with

30% of the subjects being censored by $d = 130$. The values of the parameters chosen in Setting I are listed in Table B.1 in Appendix B. The true values for the estimands $E\{Y^{(1)}\}$, $E\{Y^{(2)}\}$, and $E\{Y^{(3)}\}$ were respectively 0.36, 0.50, and 0.63 for the scenarios of weak outcome associations, and 0.40, 0.59, and 0.50 when the outcome associations were strong.

The propensity scores were estimated using a multinomial logistic regression model, and the probability of remaining uncensored at d was estimated by a Cox proportional hazards model. For the random censoring scenario, the probability of remaining uncensored was also estimated using the treatment-specific KM estimator. For the scenario with the largest proportion of censored observations ($\sim$40%), we further considered the case in which the censoring time was observed for all subjects, as was the case in our data example, and evaluated the performance of CIPWR based on the observed censoring time (in contrast to the observation time). Across the scenarios, we considered three sets of model specifications for the CIPWR estimator: (1) correctly specified models for outcome, treatment, and censoring, (2) correctly specified models for treatment and censoring only, and (3) a correctly specified outcome model only. The misspecification for each model was caused by removing the confounder X_2. The CAIPW-ZS estimator assumed a Cox proportional hazard model for the survival time, and therefore the outcome model was always misspecified in this case. The CAIPW-Wang method only considered an outcome model and a propensity score model, since the survival function of the censoring time was estimated by the KM estimator.

2.5.2 Simulation Setting II: Crossing Hazards

Two covariates independently sampled from the standard normal distribution, X_1 and X_2, were considered for the setting of crossed hazard functions. We assumed a multiphase model for the event time, where the effects of risk factors on the hazards differed by phases. The varying effects over time of risk factors are often seen in the setting of surgery. Specifically, the event time was generated such that the hazard functions crossed at some time point, and the equations we used to obtain the event time are listed in section B.3 in Appendix B. The probability of being assigned to treatment j was $\exp(\alpha_{j1}X_1 + \alpha_{j2}X_2)/\sum_{z=1}^{3}\exp(\alpha_{z1}X_1 + \alpha_{z2}X_2)$. Censoring was generated using $C = -\lambda^{-1}\exp\{\gamma_1 X_1 + \gamma_2 X_2 + \theta_1 I(Z = 2) + \theta_2 I(Z = 3)\}^{-1}\log u$. In the first scenario, we assumed that the treatment assignment and censoring time only depended on X_1, such that $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})^T = (0, 0)^T$, $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22})^T = (0.2, 0)^T$, $\boldsymbol{\alpha}_3 = (\alpha_{31}, \alpha_{32})^T = (0.3, 0)^T$, and $(\lambda, \gamma_1, \gamma_2, \theta_1, \theta_2)^T = (0.8, 1, 0, 0.2, 0.4)^T$. In the second scenario, we let X_2 come into play, such that $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})^T = (0, 0)^T$, $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22})^T = (0.2, 0.2)^T$, $\boldsymbol{\alpha}_3 = (\alpha_{31}, \alpha_{32})^T = (0.3, 0.3)^T$, and $(\lambda, \gamma_1, \gamma_2, \theta_1, \theta_2)^T = (0.7, -0.5, 0.5, 0.4, 0.2)^T$. The cutoff point d was chosen to be 0.5 and 0.3 for the first and second scenario, respectively, which led to 30.7% and 13.2% of censored

observations for the corresponding d. The true values for $E\{Y^{(1)}\}$, $E\{Y^{(2)}\}$, and $E\{Y^{(3)}\}$ were 0.68, 0.62, and 0.52 for the first scenario, and 0.54, 0.45, and 0.41 for the second scenario.

The models for the treatment assignment and censoring were correctly specified in this setting. The logistic regression model and the Cox model for the outcome were misspecified such that squared terms of the covariates and interactions (if applicable) were included in the model.

2.5.3 Evaluation Metrics

For each scenario, we generated 2000 Monte Carlo data sets, each with $n = 1500$ subjects. For CIPWR, the standard errors and 95% confidence intervals (CIs) were estimated using the formula for asymptotic variance derived in Section 2.3.2. For the other methods considered, 200 bootstrap replicates were used to estimate the standard errors and 95% CI. Simulation results are presented in terms of bias, empirical standard deviation, root mean squared error (RMSE) and coverage rate of 95% CI.

2.5.4 Simulation Results

Naive and IPW estimators were expectedly biased away from the true risk differences in all scenarios (Figures B.1 and B.2), as they failed to accommodate censoring. All other methods had close to zero empirical bias when censoring was unrelated to covariates (Figure B.1). On the other hand, when censoring depended on the covariates, the bias for Pseudo and CAIPW-Wang, which relied on the strict assumption $C \perp\!\!\!\perp T|Z$, became non-negligible (Figures B.1 and B.2).

The RMSE results for Setting I are displayed in Figures 2.1 and 2.2. We report the ratio of RMSE to the RMSE of CIPW with correctly modeled coarsening mechanism. Figure 2.1 summarizes the results for different censoring mechanisms and proportions of being censored. In general, CAIPW-ZS and CIPWR had the smallest RMSE among the methods considered across all treatment pair comparisons. As the censoring proportions increased, we observed larger gain in efficiency for CAIPW-ZS and CIPWR over CIPW. Cox model-based and KM estimator-based CIPWR yielded similar RMSE in the random censoring scenario (Figure B.6 in Appendix B), with the former having slightly smaller RMSE than the latter in some cases. This finding suggests that estimating the probability of remaining uncensored from the data, even if the value is known, may actually lead to smaller variance for the CIPWR estimator than using the true value, which is consistent with the theoretical results in the literature [71]. Moreover, estimating the probability of remaining uncensored using observation time tended to reduce the RMSE compared with using observed censoring time (Figure B.7 in Appendix B). Figure 2.2 displays the RMSE results for different levels of outcome associations. When the associations between the covariates and the outcome were weak, we observed 2.6%-3.7% reduction in RMSE for CIPWR and CAIPW-ZS.

Greater reduction (7.6%-10.5%) was noted as the associations became stronger. The RMSE results for Setting II where crossing hazards existed are presented in Figure 2.3. Note that in this setting both the logistic regression model and the Cox model for the outcome were misspecified. Again, we observed lower RMSE for CIPWR and CAIPW-ZS compared to CIPW. Furthermore, CAIPW-ZS produced larger variability (and therefore larger RMSE) than CIPWR when the hazard functions had a crossing (Figure 2.3 and Table B.7 in Appendix B). For example, the ratios of the variance of CAIPW-ZS to that of CIPWR ranged from 1 to 1.06 in the first scenario, and from 1.05 to 1.08 in the second scenario.

All methods that were approximately unbiased in the presence of nonrandom censoring (i.e., CIPW, CIPW-ZS, CAIPW-ZS, and CIPWR) achieved close to nominal coverage of 95% across all scenarios considered (Table B.2-B.6).

2.6 Application to Comparison of Treatments for Prostate Cancer using Medical Claims

2.6.1 Data Analysis Methods

We applied our proposed method to a dataset comprised of patients with metastatic castration-resistant prostate cancer (mCRPC), which was obtained from a large national private health insurance network (Optum Clinformatic Data Mart). The study cohort included patients who used at least one of the six drugs (docetaxel, abiraterone, enzalutamide, sipuleucel-T, cabazitaxel, and radium-223) approved to treat mCRPC from January 1, 2014, to December 31, 2019. Among these drugs, docetaxel and cabazitaxel are chemotherapies, abiraterone and enzalutamide are oral hormone therapies, sipuleucel-T is an immunotherapy, and radium-223 is a radioactive drug. We excluded the patients who received cabazitaxel ($n = 56$) or radium-223 ($n = 28$) as their first-line therapy from our analysis, since there were much fewer samples in these two groups than the other four. We examined the occurrence of ER visits and all-cause hospitalization within 180, 270, and 360 days of treatment initiation, respectively. Patients who switched to another treatment or dropped out of the insurance plan prior to the event of interest within the pre-specified time window were considered as being censored.

The treatment was modeled using a multinomial logistic regression adjusting for age, race, education level, household income, geographic region, insurance product type, whether the insurance plan is administrative services only (ASO), metastatic status of cancer, year of first prescription, comorbid conditions, and provider type [58]. All covariates were binary or categorical. To improve the common support of the covariate distributions, we followed the criteria discussed in Lopez and Gutman [25] and discarded the tails of the propensity score distributions. The outcome model

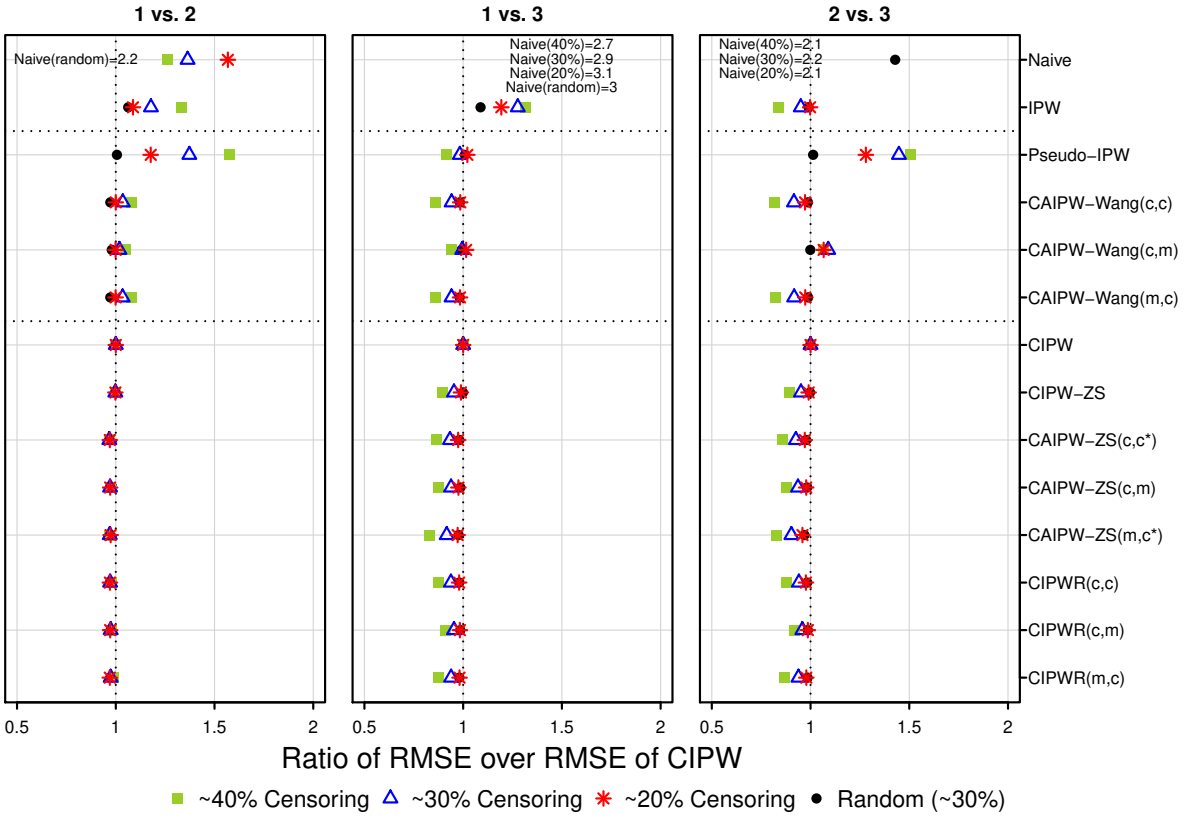

Figure 2.1: RMSE over RMSE of CIPW with correctly specified propensity and censoring models for different proportions of censoring in Setting I. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c* to denote the case where the true predictors for the outcome were included in the model. Propensity model is correctly specified for IPW, Pseudo-IPW, CIPW, and CIPW-ZS. Numbers that fall outside the range of x-axis are labeled in the figure. Sample size was 1500. Results were obtained using 2000 simulated datasets.

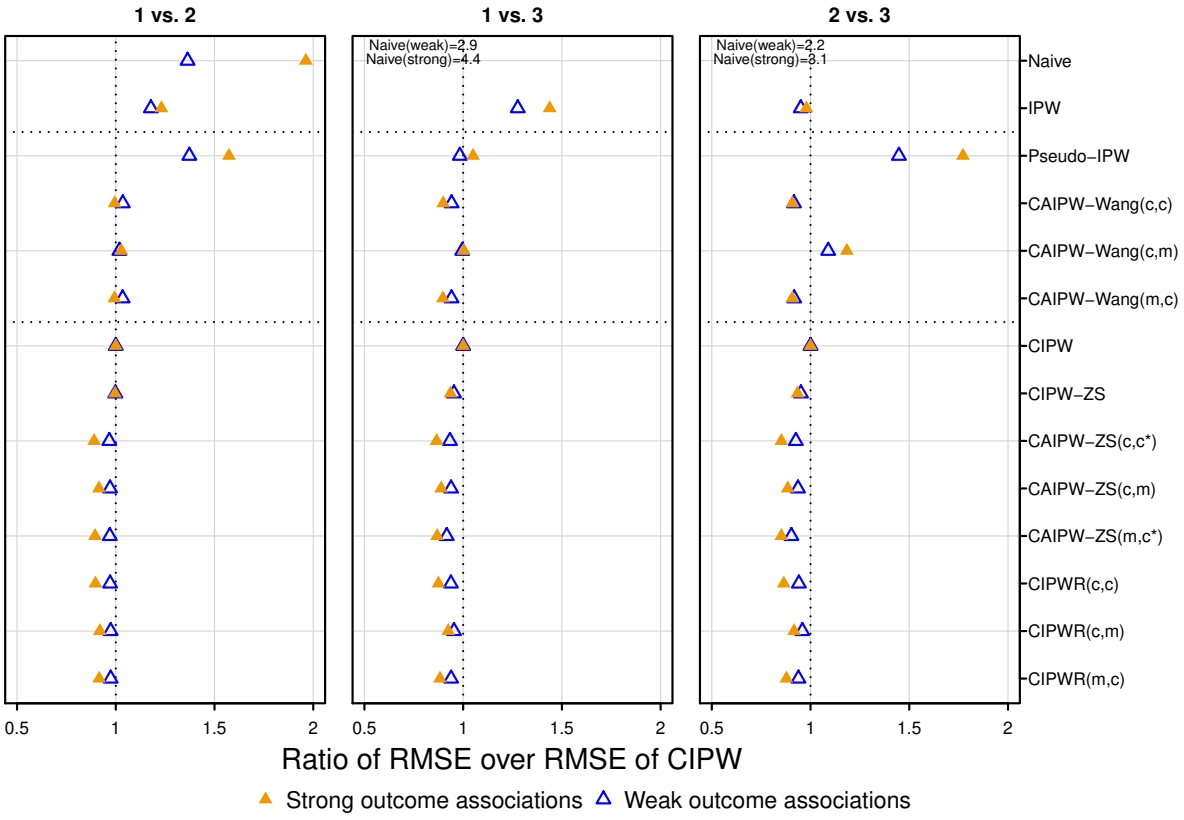

Figure 2.2: RMSE over RMSE of CIPW with correctly specified propensity and censoring models for different levels of outcome associations in Setting I. Censoring depended on covariates and censoring proportion was 30%. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c* to denote the case where the true predictors for the outcome were included in the model. Propensity model is correctly specified for IPW, Pseudo-IPW, CIPW, and CIPW-ZS. Numbers that fall outside the range of x-axis are labeled in the figure. Sample size was 1500. Results were obtained using 2000 simulated datasets.

and the censoring model adjusted for the same set of covariates that were controlled for in the treatment model. The CIs were obtained using (1) Wald-type CIs based on original data for the Naive method, and (2) bootstrap standard errors based on 200 bootstrap samples for the rest of the methods.

2.6.2 Data Analysis Results

Patients who had less than 180 days of continuous enrollment prior to the first prescription, had missing covariates, or experienced the event of interest on the same day as the first prescription were removed from the analysis. In the end, we identified 7678 and 7709 mCRPC patients for ER visit and hospitalization, respectively, and calling them ER visit cohort and hospitalization cohort. The sample sizes of the two cohorts differed because ER visit and initial treatment prescription for the first time were more likely to occur on the same day than subsequent hospitalization. The proportions of overall and cause-specific censoring within each specified time window for the two outcomes are reported in Table B.8 in Appendix B. In general, hospitalization (24.6%-40.6%) was associated with greater overall proportion of censored observations than ER visits (20.8%-32.6%). The proportions of patients being censored and the unadjusted risks ignoring censored patients for each treatment group are presented in Table B.9 in Appendix B. In particular, Sipuleucel-T group had larger percentage of censored patients than the other three groups. Docetaxel group had the highest crude risks of ER visits (53.6%, 64.6%, and 71.8% within 180, 270, and 360-day time windows, respectively) and hospitalization (41.1%, 52.3%, and 60.2% within 180, 270, and 360-day time windows, respectively). The baseline demographic and clinical characteristics of the ER visit cohort stratified by treatment groups are presented in B.10 in Appendix B. The covariate distributions of the hospitalization cohort were close to those of the ER visit cohort (data not shown).

To improve the covariate overlap among the treatment groups for the comparative analysis, we applied data trimming with criteria discussed in Lopez and Gutman [25], which left us with 7003 and 7045 patients for ER visit and hospitalization, respectively.

Figure 2.4 shows the differences in 360-day risks among the four treatment groups for both outcomes of interest. Results for 180-day and 270-day risks are presented in Figures B.8 and B.9 in Appendix B. When confounding and censoring were both ignored, docetaxel users had significantly higher risk of at least one ER visit within 360 days of treatment initiation than users of abiraterone, enzalutamide, and sipuleucel-T. Similar directional results for docetaxel vs. abiraterone and docetaxel vs. enzalutamide comparisons were noted for methods that accounts for both confounding and censoring, and the corresponding 95% CIs consistently excluded zero across the methods. For example, the risk difference estimated by the CIPWR estimator using observation

time (CIPWR1) was -0.082 (95% CI [-0.118, -0.046]) for docetaxel vs. abiraterone comparison, and -0.156 (95% CI [-0.198, -0.114]) for docetaxel vs. enzalutamide comparison. These findings agree with the clinical evidence that oral therapies abiraterone and enzalutamide tend to have fewer side effects than docetaxel, a chemotherapy [80]. Sipeucel-T was identified to have lower risk of ER visit than docetaxel, though the differences were not significant for some of the methods (e.g., risk difference=-0.158, 95% CI [-0.295, -0.021] for CAIPW-ZS; risk difference=-0.073, 95% CI [-0.172, 0.025] for CIPWR1). For the two oral drugs, Enzalutamide was identified to have lower risk of ER visit than Abiraterone within each specified specified period of time when both confounding and censoring were accounted for. The Naive and IPW methods indicated significantly higher 360-day risk of ER visit for Enzalutamide than Siputeucel-T, while the methods that account for both confounding and censoring showed that there was no significant difference. Similar patterns were observed for the risks of all-cause hospitalization for each time window considered.

In general, CIPWR based on observed censoring time (CIPWR2) tended to have wider CIs than CIPWR using observation time (CIPWR1), which is consistent with our simulation results (Figure B.7 in Appendix B). For example, the ratios of confidence widths of CIPWR1 over CIPWR2 for 360-day hospitalization ranged from 54.9% to 86.9%. Greater differences in the width of CIs were noted as the duration of time window became longer and the proportion of censored patients increased (Figures B.8 and B.9 in Appendix B and Figure 2.4). In most cases, CAIPW-ZS and CIPWR yielded narrower CIs than the CIPW estimator. For example, the percentage of reduction in width ranged from 58.8% to 92.4% for the risk of ER visits estimated by CIPWR1. CAIPW-ZS was noted to have wider CIs than CIPWR for treatment pairs that involved Sipuleucel-T for ER visits (Figure 2.4). Greater differences in point estimates between CAIPW-ZS and CIPWR, the former of which modeled the entire survival curve over time, was noted as the ending time point moved farther away from the treatment initiation.

Differences between methods that ignore and account for censoring increased as the time window was extended. Results for methods that rely on different independence assumptions on the censoring mechanism were similar. One possible reason is that censoring may only depend on the treatment in this cohort, and the restrictive version of the assumption is satisfied, as the censored patients within each time window exhibited similar demographic and baseline clinical characteristics to uncensored ones (Table B.11 in Appendix B), and most covariates were not significantly associated with censoring (Table B.12 in Appendix B).

2.7 Discussion

We present an inverse probability weighted regression-based estimator, CIPWR, for average treatment effect for a binary outcome that is subject to right-censoring. This method is based on the

intuitively simple standardization idea, where we model the binary outcome given the observed covariates using the familiar logistic regression model for each treatment separately and then averaging predictions for all patients. The CIPWR method improves robustness by accounting for confounding due to nonrandomized treatment and censoring using the inverse probability weighting approach. Therefore, the proposed method is a hybrid of the two general approaches (standardization and weighting) in the missing data and causal inference literature that handle missingness, confounding and censoring. Like the well-studied augmented inverse probability weighting approach (e.g., CAIPW-Wang and CAIPW-ZS), the proposed method enjoys a double robustness property such that the estimator is consistent if either the (binary) outcome, or both treatment assignment and censoring are correctly modeled. However, in this method the double robustness and improvement in efficiency are not through direct augmentation. Instead, it achieves double robustness by combining two approaches in different steps, with each step based on popular models and the natural ideas of standardization and weighting. The proposed method is conceptually straightforward to understand and easy to implement using standard statistical software for practitioners.

Simulation studies show that in finite sample, CIPWR yielded approximately unbiased estimates and close to nominal coverage of 95% across the scenarios considered, particularly when censoring depends on the covariates. CIPWR also provides efficiency gain over CIPW by exploiting the information from the outcome model. The proposed method was applied to claims data for comparing the average treatment effects of multiple treatments.

Time-to-event data are often analyzed using approaches that model the whole survival curve from baseline to the end of follow-up, such as the Cox proportional hazards model used for CAIPW-ZS [67]. In the case where interest only lies in the risk difference over a pre-specified period of time (e.g., 180-day risk of ER visit), a method that directly targets the survival function at the fixed time point can lead to better efficiency than general methods that estimate the whole survival curve. The problem can be reduced to estimating the marginal expectation of a (possibly censored) binary indicator of event occurrence, which we propose to solve by utilizing logistic regression that directly targets the binary outcome. In general, the difficulty of correctly modeling the time-to-event outcome given observed covariates increases as the time window becomes longer, and the incorrect model for the survival time tends to result in more severe problems than misspecification of the outcome model at a fixed time point. In our simulation setting where the hazard functions did not cross, CIPWR, which directly modeled the binary indicator of event occurrence, performed similarly to CAIPW-ZS, which utilized outcome information accumulated over time, in terms of RMSE. In the presence of crossing hazards, when both the logistic regression model and Cox model were misspecified, CIPWR realized more efficiency gain over CIPW than CAIPW-ZS, as the latter failed to capture the associations between the outcome and covariates. The efficiency gain of CIPWR over CAIPW-ZS was also observed in the data example, where CAIPW-ZS had

wider CI than CIPWR for some treatment pairs (Figure 2.4), and the difference in confidence width increased as the time window became larger (Figure B.8 in Appendix B and Figure 2.4).

In the presence of right censoring, discarding the censored observations may result in biased estimates for the treatment effects, even if the censoring was completely random, as shown in our simulation studies (Figure B.1 in Appendix B). In this paper, we assumed the independence of survival and censoring time conditional on treatment and baseline covariates, and used Cox model to estimate the probability of remaining uncensored, which was then inverted and used as weights in the estimating equation. Under the more restrictive conditional independence assumption given treatment only, it is sufficient to use the non-parametric KM estimator for estimating the probability of remaining uncensored. However, simulation results showed that when censoring was random, CIPWR based on Cox model was still unbiased for the treatment effect, and could possibly be more efficient than CIPWR based on KM estimator.

In this study, we focus on low-dimensional covariates and only consider simple parametric working models for the outcome and treatment. As researchers gain increasing access to large databases with a substantial collection of covariates, variable selection techniques for causal inference has been an emerging topic of interest [e.g., 81, 82]. Another possible extension to our method is to replace the parametric models with modern machine learning methods that can capture potential nonlinearities and nonadditivities. For example, neural networks and methods based on recursive partitioning have been suggested as promising alternatives to logistic regression for estimating propensity scores when the true model structure is complex [83, 84]. In addition, treatment switching was treated in the same way as dropout and study termination. That is, an observation was considered censored when someone switched treatment, which may not be optimal and studying treatment sequences will be another challenge.

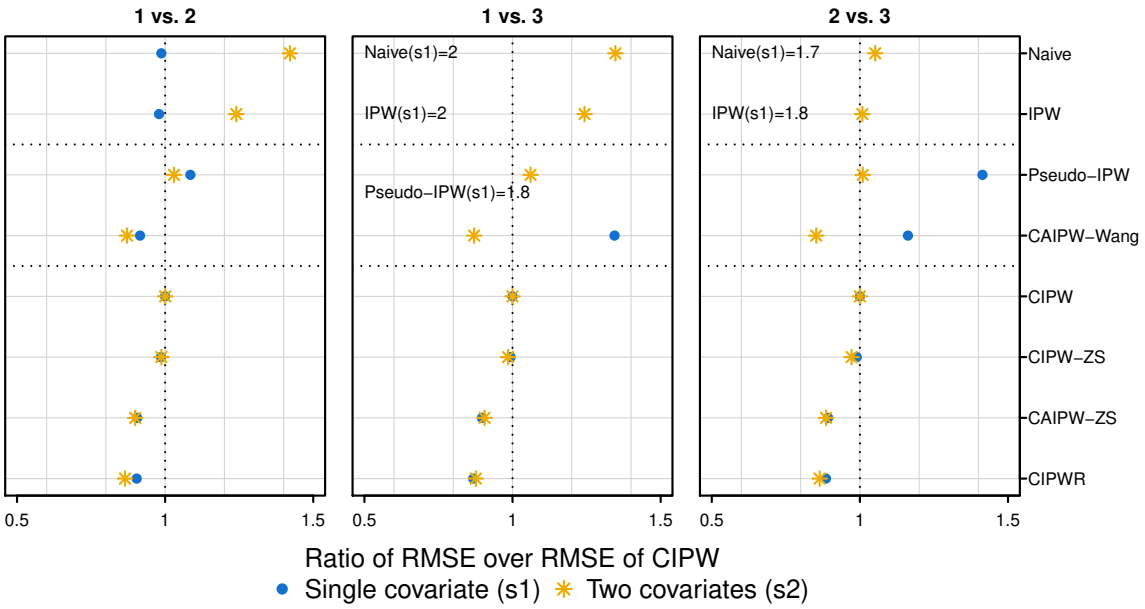

Figure 2.3: RMSE over RMSE of CIPW with correctly specified propensity and censoring models in the presence of crossing hazards in Setting II. The models for the coarsening mechanism were correctly specified. The outcome model was always misspecified in this setting. Numbers that fall outside the range of x-axis are labeled in the figure. Sample size was 1500. Results were obtained using 2000 simulated datasets.

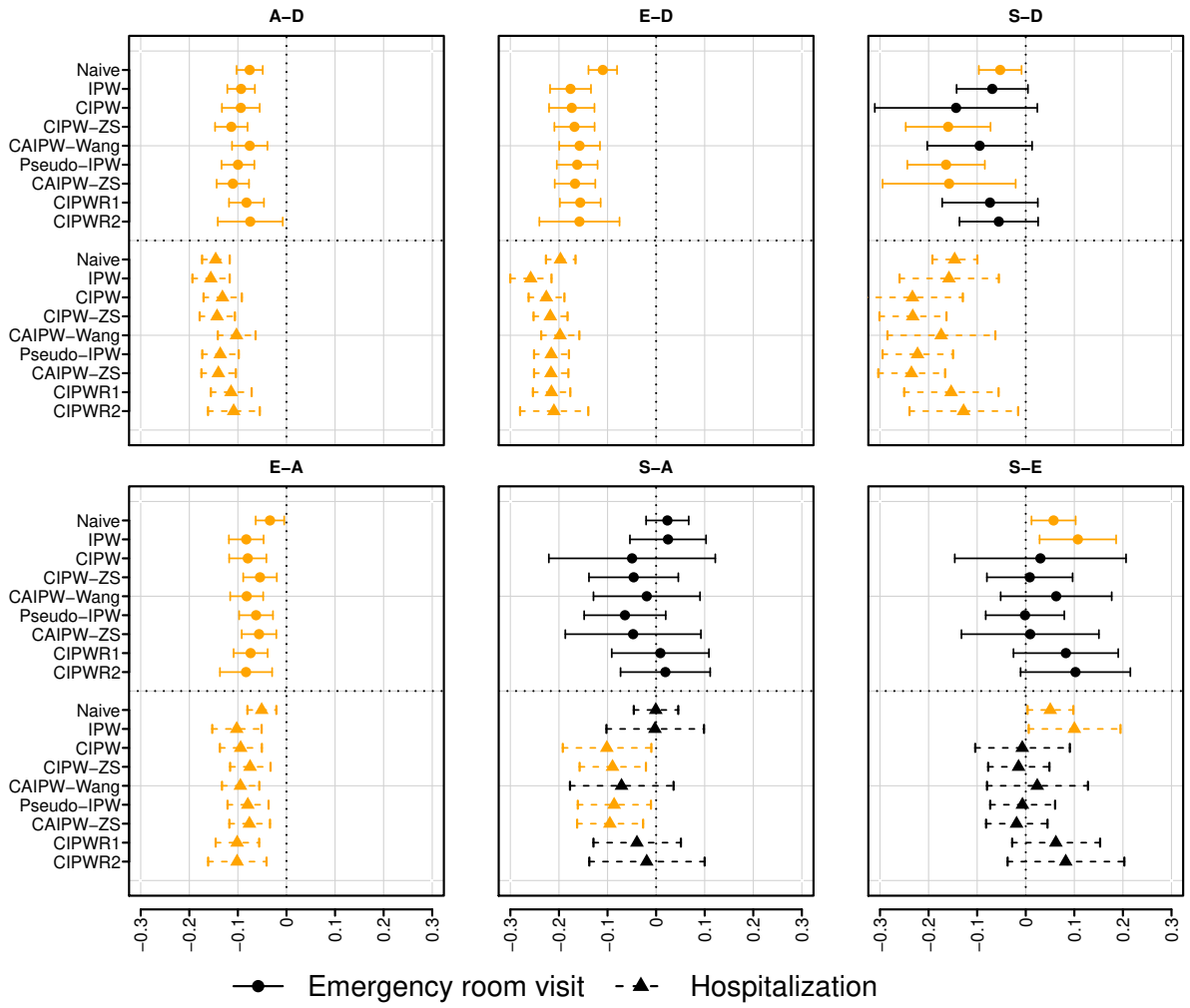

Figure 2.4: Differences in 360-day risks of experiencing at least one emergencyroom visit among the four focus drugs and the associated 95% confidence intervals. Data were obtained from Optum Clinformative Data Mart, with the outcome interest being the occurrence of emergency room visit within 180 days of treatment initiation. Total sample size is $N = 7003$ ($N_A = 2458$, $N_D = 2162$, $N_E = 1833$, $N_S = 550$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

CHAPTER 3

Outcome-Adaptive Propensity Methods for Handling Censoring and High Dimensionality: Application to Insurance Claims

3.1 Introduction

To obtain an unbiased estimate for the causal effect of a treatment using data from observational studies, it is important to control for confounding by adjusting for the differences in pre-treatment baseline covariates between treatment groups. A commonly used tool for confounder adjustment is the propensity score, defined as the conditional probability of receiving a specific treatment given a set of covariates [10], which is usually unknown in observational studies and needs to be estimated from the data. The estimated propensity scores can be used in matching [24, 28], weighting [3], regression adjustment [17], among others, for estimating the causal treatment effect. Regardless of the propensity score-based methods chosen, making valid inference on the treatment effect relies on good estimation for the propensity scores.

There has been a large body of work focusing on the estimation and use of propensity scores in a low-dimensional setting, where the sample size is far greater than the number of candidate covariates. In this setting, empirical researchers normally estimate the propensity scores by fitting a logistic regression model for binary treatment, or a multinomial logistic regression model for more than two treatment groups. In the era of 'big data', large databases are increasingly being used to conduct comparative effectiveness research. For example, insurance claims data were used to compare the safety of four drugs on the market prescribed for patients with metastatic castration-resistant prostate cancer [69]. A massive collection of candidate covariates, such as demographics, socioeconomic status, clinical measurements, and diagnosis codes, can be ascertained from patients' health care or insurance claims data, and the number of covariates can possibly be large compared to the sample size in each treatment group. Standard parametric regression models may become problematic when the dimensionality increases, and a key challenge is to identify

variables to be included in the propensity score model from a high-dimensional set of measured covariates. There has been considerable interest in developing methods that perform variable selection for estimating propensity/balancing scores in the high-dimensional setting. For instance, Schneeweiss et al. [85] proposed the high-dimensional propensity score (hd-PS) algorithm that selects covariates to facilitate high-dimensional propensity score adjustment using health care claims data. Specifically, they rank each covariate based on its potential for controlling confounding by assessing the covariate's prevalence and univariate association with the treatment and outcome, and then pick the top k covariates for inclusion in the propensity score modeling. Athey et al. [82] circumvented propensity score modeling and addressed the problem of high dimensionality using balancing weights. They proposed a two-stage algorithm called approximate residual balancing (ARB). In the first stage, one fits a regularized linear model, such as elastic net, for the outcome. In the second stage, one reweighs the first-stage residuals using the weights with minimum variance that optimize the balance of covariates.

Parametric models requires empirical researchers to impose structures to the models in terms of variable selected and their functional forms. A practical difficulty is that misspecification of the model can result in substantial bias of the estimated treatment effect [44]. Machine learning methods are well-known for their powerful predictive performance and ability to handle complex and nonlinear relationship between the response and predictors. There has been a growing interest in using machine learning techniques for propensity score modeling [83, 84, 86, 87]. Setoguchi et al. [83] compared several data mining techniques that optimize the prediction of treatment status, including classification and regression trees (CART), pruned CART, and neural networks, in the context of propensity score matching with a continuous outcome. Lee et al. [84] extended the work of Setoguchi et al. [83] to the setting with a binary outcome, and evaluated the performance of CART, pruned CART, bootstrap aggregated (bagged) CART, random forests, and boosted CART with regard to propensity score weighting. Both works focused on the low-dimensional setting, and demonstrated that machine learning methods, such as CART and neural networks, are promising alternatives to parametric modeling for the estimation of propensity scores in the presence of non-additivity and/or nonlinearity in the true model. However, numerical evaluations of these machine learning techniques for high-dimensional covariates remain limited.

These aforementioned approaches only consider the treatment-covariate relationship when modeling the propensity/balancing scores, and fail to incorporate the outcome models into the treatment modeling process. Studies illustrate that using covariates that are associated with the treatment but not the outcome will inflate the variance of the estimators of average treatment effects (ATE) without reducing the bias [81, 88]. On the other hand, adding covariates explaining the outcome but not the treatment to the propensity score model can improve the precision of the treatment effect estimates [81, 89]. These findings suggest that a highly predictive model for treat-

ment assignment will not necessarily lead to efficient estimators of treatment effects. Therefore, standard variable selection methods designed for prediction, which rely only on the relationship between treatment and covariates, may yield suboptimal results in the context of causal inference. There is an expanding literature on variable selection methods for causal inference that account for the information about the outcome-covariate relationship. Shortreed and Ertefaie [81] proposed the Outcome-Adaptive Lasso (OAL) method, which adopts the adaptive lasso framework [90] and places smaller adaptive weights on the outcome predictors, for selecting covariates to be included in the propensity score model. As a result, heavier penalties are imposed on variables relevant to treatment only than variables predictive of outcome only. Other works that allow the outcome information to contribute to variable selection for propensity score modeling include Outcome Highly Adaptive Lasso proposed by Ju et al. [91] and Bayesian Adjustment for Confounding proposed by Zigler and Dominici [92]. However, how information about the outcome-covariate relationship can be incorporated into tree-based machine learning methods, such as CART, and to what extent can outcome models contribute to the efficiency gain in high-dimensional data settings have been less studied.

In this Chapter, we evaluate multinomial logistic regression (combined with lasso penalty) as well as several popular machine learning algorithms for variable selection and propensity score estimation in the context of high-dimensional covariates. We estimate the causal treatment effects using the inverse probability weighting (IPW) estimator, although the estimated propensity scores considered in this study are applicable to any propensity score-based methodology, such as propensity score matching. As in Chapter 2, we also aim to account for censoring at the same time, where the bias due to censoring is controlled for by applying the inverse probability of remaining uncensored as weights to the outcome. We focus on estimating the treatment effects on a binary outcome (that is possibly censored) among multiple treatment groups. To the best of our knowledge, the literature lying within the intersection of high-dimensionality, complexity of the associations between treatment and covariates, and outcome-adaptive propensity score modeling is largely absent.

In Section 3.2, we introduce the notations and basic setup of the problem. Section 3.3 describes the algorithm we consider for selecting variables to included in the treatment model. Section 3.4 outlines the methods considered for the final treatment model that is used to estimates the propensity scores. Section 3.5 presents simulation studies that compare the methods considered in setting with high-dimensional covariates and potentially complex underlying treatment model (i.e., model with nonlinearity and/or nonadditivity). We also consider a setting where censoring exists. We illustrate the efficiency gain provided by leveraging the information about the outcome-covariate relationship when estimating the propensity scores. Section 3.6 presents a data example that compares the risks of hospitalization and emergency room (ER) visits of four prostate cancer treatments

using data from an insurance claims database. The binary outcome of this data example is subject to censoring, as the follow-up period tended to terminate early due to drop-out or treatment switch. We conclude with a discussion section.

3.2 Definition of the Problem and Notations

3.2.1 Notations and Assumptions

We consider n independent individuals, indexed by i, with $\boldsymbol{X}_i$ being a p-dimensional vector of covariates measured prior to receiving the treatment Z_i, where $Z_i = j \in \{1, \cdots, J\}$. We focus on the context in which the ratio of p to n is relatively large but smaller than 1. We let $\mathcal{C}$, $\mathcal{Z}$, $\mathcal{Y}$, and $\mathcal{S}$ be the indices of confounders (i.e., covariates associated with both treatment and outcome), covariates predictive of treatment only, covariates predictive of outcome only, and covariates unrelated to both treatment and outcome (i.e., spurious covariates), respectively. Suppressing the index by i, $\boldsymbol{X}_{\mathcal{C}}$, $\boldsymbol{X}_{\mathcal{Z}}$, $\boldsymbol{X}_{\mathcal{Y}}$, and $\boldsymbol{X}_{\mathcal{S}}$ are mutually exclusive and $\boldsymbol{X}_i = \boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Z}} \cup \boldsymbol{X}_{\mathcal{Y}} \cup \boldsymbol{X}_{\mathcal{S}}$. We further let $|\mathcal{C}|$, $|\mathcal{Z}|$, $|\mathcal{Y}|$, and $|\mathcal{S}|$ denote the cardinality of the corresponding set. We let T_i be the underlying lag time to the first event of interest for each individual, and C_i be the censoring time. The outcome of interest is whether the event of interest occurs before a prespecified time point d, defined by $Y_i = I(T_i < d)$, which results in a possibly censored binary outcome. The information on Y_i may not be completely available due to dropout, study termination, or treatment switch. In the absence of censoring, T_i (and therefore Y_i) would be observed for all individuals, and the set of complete data is then $(\boldsymbol{X}_i, Z_i, Y_i)$. When the outcome variable is subject to right-censoring, Y_i is observed only if the individual has not been censored before d, and we let $R_i = I\{C_i \geq \min(T_i, d)\}$ be the indicator of observing Y_i.

Under the potential outcome framework [41], each individual is associated with a set of potential outcomes $\{Y^{(1)}, \cdots, Y^{(J)}\}$, where $Y^{(j)}$ denotes the potential outcome had the individual received treatment j. The causal parameter of interest is the marginal ATE on the outcome between j and j', denoted $\tau(j, j') = E\{Y^{(j')}\} - E\{Y^{(j)}\}$. The following assumptions are required for valid inferences on causal effects using the observable data.

(A1) (*Random sampling*) The individuals in the study are randomly sampled from the population.

(A2) (*Stable Unit Treatment Value Assumption, or SUTVA*) For any individual i, $i = 1, \cdots, n$, if $Z_i = j$, then $Y_i = Y_i^{(j)}$, for all $j = 1, \cdots, J$.

(A3) (*Unconfoundedness*) $\{Y_i^{(1)}, \cdots, Y_i^{(J)}\} \perp\!\!\!\perp Z_i | \boldsymbol{X}_i$.

(A4) (*Overlap*) For all values of j and $\boldsymbol{x}$, $0 < \pi_j(\boldsymbol{x}) < 1$, where $\pi_j(\boldsymbol{x}) = pr(Z_i = j | \boldsymbol{x})$.

(A5) (*Censoring at random*) $C_i \perp\!\!\!\perp \{T_i^{(1)}, \cdots, T_i^{(J)}\} \Big| (Z_i, \boldsymbol{X}_i)$.

3.2.2 Underlying Models for Outcome, Treatment, and Censoring, and Estimators for Average Treatment Effects

We assume that the true outcome model for $Z_i = j$ is a logistic regression model,

$$\text{logit}\, P(Y_i = 1 | \boldsymbol{W}_i, Z_i = j) = \boldsymbol{W}_i^T \boldsymbol{\beta}_j$$

where $\boldsymbol{W}_i$ is a p_w-dimensional function of $(\boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{Y}}^T)^T$. The treatment assignment mechanism is governed by a multinomial logistic regression

$$\log \frac{P(Z_i = j | \boldsymbol{V}_i)}{P(Z_i = J | \boldsymbol{V}_i)} = \boldsymbol{V}_i^T \boldsymbol{\alpha}_j, \quad j = 1, \cdots, J,$$

where J is the reference level and $\boldsymbol{V}_i$ is a p_v-dimensional function of $(\boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{Z}}^T)^T$. Both $\boldsymbol{W}_i$ and $\boldsymbol{V}_i$ may contain nonlinear terms and interactions. We assume that $|\mathcal{C}|+|\mathcal{Y}| \ll p$ and $|\mathcal{C}|+|\mathcal{Z}| \ll p$, where $|\mathcal{C}|$, $|\mathcal{Y}|$, and $|\mathcal{Z}|$ are the numbers of variables in $\boldsymbol{X}_{\mathcal{C}}$, $\boldsymbol{X}_{\mathcal{Z}}$, and $\boldsymbol{X}_{\mathcal{Y}}$, respectively. Note that in practice, $\boldsymbol{X}_{\mathcal{C}}$, $\boldsymbol{X}_{\mathcal{Y}}$, and $\boldsymbol{X}_{\mathcal{Z}}$ are generally unknown and a common practice is to include all candidate covariates in the model. In the case where the ratio of p to n is relatively large, traditional regression models based on maximum likelihood estimation may not converge, and some variable selection procedure is required for model fitting.

With respect to censoring, we assume a proportional hazard model for treatment $j = 1, \cdots, J$,

$$\lambda_j(t | \boldsymbol{U}_i) = \lambda_{0j}(t) \exp(\boldsymbol{U}_i \boldsymbol{\gamma}_j),$$

where, $\lambda_{0j}(t)$ is the treatment-specific baseline hazard function and $\boldsymbol{U}_i$ is a p_u-dimensional function of $\boldsymbol{X}_i$. The censoring model is assumed to have moderate number of predictors ($p_u \ll n$) that are known *a priori*.

We estimate the ATE using the IPW estimator,

$$\hat{\tau}(j, j') = \frac{\sum_{i=1}^n \hat{w}_i I(Z_i = j') Y_i}{\sum_{i=1}^n \hat{w}_i I(Z_i = j')} - \frac{\sum_{i=1}^n \hat{w}_i I(Z_i = j) Y_i}{\sum_{i=1}^n \hat{w}_i I(Z_i = j)}$$

where $\hat{w}_i = \sum_{j=1}^J 1/\hat{\pi}_j(\boldsymbol{X}_i)$ and the propensity scores $\hat{\pi}_j(\boldsymbol{X}_i)$, $j = 1 \cdots, J$, are required to be estimated from the data in an observational study. In the presence of censoring, the outcome is

further weighted by the inverse probability of remaining uncensored at d,

$$\hat{\tau}(j,j') = \frac{\sum_{i=1}^{n} \hat{w}_i' R_i I(Z_i = j') Y_i}{\sum_{i=1}^{n} \hat{w}_i' R_i I(Z_i = j')} - \frac{\sum_{i=1}^{n} \hat{w}_i' R_i I(Z_i = j) Y_i}{\sum_{i=1}^{n} \hat{w}_i' R_i I(Z_i = j)}.$$

The weights $\hat{w}_i' = \sum_{j=1}^{J} (\hat{\pi}_j(\boldsymbol{X}_i) \exp\{\Lambda_{ij}(\min(T_i, C_i, d))\})^{-1}$, where $\Lambda_{ij}(t)$ is the cumulative hazard function of C_i at t for treatment j.

3.3 Variable Selection for Dimensionality Reduction for the Propensity Score Model

When the dimension of the covariate vector is high, it tends to be infeasible to fit an unrestricted parametric model, such as a multinomial logistic regression model, for the treatment using all the available covariates. A practical problem for empirical researchers is to identify a subset of covariates to be conditioned on to control for confounding. Typically, in a medical research, a list of important covariates will be suggested based on the evidence in the literature and/or experts' opinion. However, as the number of the available covariates increases, it becomes extremely difficult for human experts the check manually which variables are potential confounders. Alternatively, one can turn to data-driven variable selection approaches, such as lasso [93], which automatically select important variables for treatment predictions from all the available covariates. Figure 3.1 displays a flowchart of several possible routes that can be followed to identify the set of covariates to be included in the final treatment model for estimating the propensity scores. A commonly chosen route is to apply shrinkage methods directly to the original reservoir of covariates (route [1] in Figure 3.1), and we label the set of covariates `All`. In this case, variable selection and propensity score estimation are conducted simultaneously in a single step. As was noted later in our simulation, following route (1) can result in substantially biased effect estimates. One possible reason is that large number of noise variables slow down the rate of convergence for the lasso. For sparse high-dimensional data, large value of tuning parameter is necessary to select a parsimonious model. However, large penalties at the same time increase the shrinkage of non-zero components, leading to less optimal estimation [94]. Therefore, reducing the number of spurious covariates entering the final treatment model through variable selection may help improve the performance of the propensity score estimation methods.

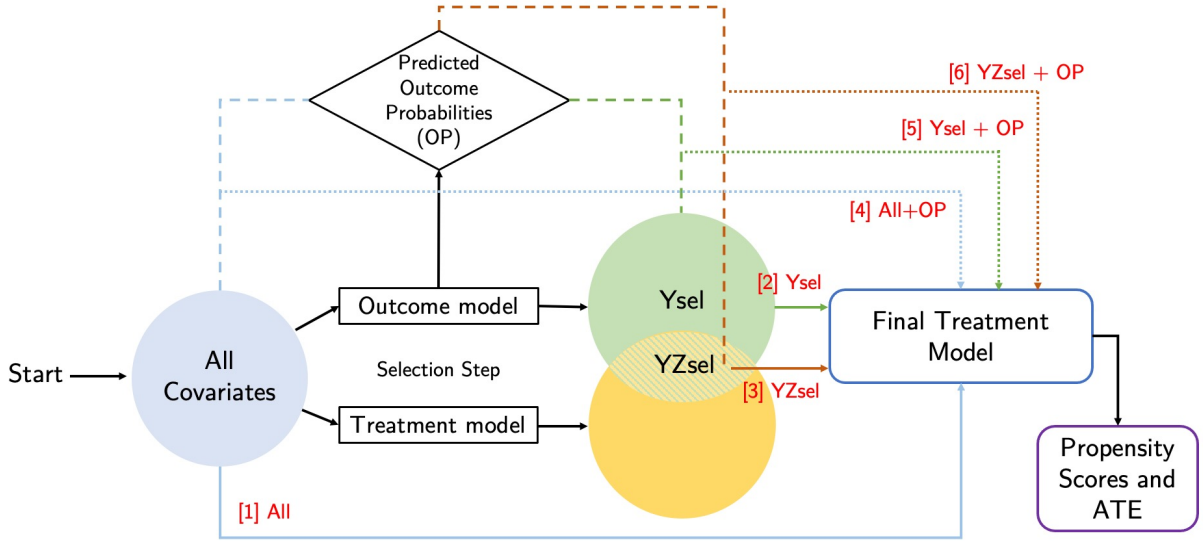

Figure 3.1: Flowchart for variable selection and propensity score estimation. Routes [1]-[6] correspond to different sets of input variables: [1] All, [2] Ysel, [3] YZsel, [4] OP+All, [5] OP+Ysel, [6] OP+YZsel.

3.3.1 Using Outcome Model for Variable Selection

Theoretically, adjusting for $\boldsymbol{X}_{\mathcal{C}}$ alone in the treatment model is sufficient to remove the confounding bias, which suggests that one targets $\boldsymbol{X}_{\mathcal{C}}$ when selecting covariates for inclusion in the propensity score. To identify variables in the set $\boldsymbol{X}_{\mathcal{C}}$, we apply a pre-selection procedure to the original set of covariates (`All`), where a regularized model (in this case we choose lasso) is fitted separately for both the outcome and the treatment. Specifically, the model for a binary outcome is specified as

$$\text{logit}P(Y_i = 1|\boldsymbol{X}_i) = \boldsymbol{X}_i^T\boldsymbol{\theta}, \tag{3.1}$$

the coefficients of which are estimated based on the lasso penalty

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\{-Y_i \boldsymbol{X}_i^T\boldsymbol{\theta} + \ln(1 + e^{\boldsymbol{X}_i^T\boldsymbol{\theta}})\} + \lambda \sum_{k=1}^{p} \|\theta_k\|_1 \right],$$

where the tuning parameter λ is chosen using cross-validation. Here we exclude the treatment variable Z from the outcome model. For the treatment assignment mechanism, we assume a multi-

nomial logistic regression model,

$$P(Z_i = j|\boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i^T \boldsymbol{\psi}_j)}{\sum_{l=1}^{J} \exp(\boldsymbol{X}_i^T \boldsymbol{\psi}_l)} \tag{3.2}$$

which is fitted by minimizing the negative penalized log-likelihood

$$N^{-1} \sum_{i=1}^{N} \log P(Z_i = j|\boldsymbol{X}_i) + \lambda \sum_{j=1}^{J} \|\boldsymbol{\psi}_j\|_1,$$

where the different $\boldsymbol{\psi}_j$ correspond to the vectors of coefficients for the J different treatment groups [95]. This penalty function ensures that each variable in $\boldsymbol{X}_i$ is selected or excluded for all J levels, as opposed to each level having its own set of selected variables. We use `YZsel` to label the intersection of the two sets of variables selected by Model 3.1 and Model 3.2. In an ideal case, `YZsel` is identical to $\boldsymbol{X}_{\mathcal{C}}$. Route [3] in Figure 3.1 corresponds to the case where `YZsel` is used as predictors for propensity scores estimation.

Traditional variable selection techniques applied to the treatment model targets $\boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Z}}$, and variables included in the propensity score models are selected based on the goodness-of-fit for the treatment mechanism itself. However, it has been shown that the use of $\boldsymbol{X}_{\mathcal{Z}}$ for propensity score modeling may inflate the variance of the estimated ATE. On the other hand, including $\boldsymbol{X}_{\mathcal{Y}}$ in the propensity score model can improve the precision of the ATE estimates [81, 89]. Later in our simulation studies, we also observed that using outcome predictors as input for the propensity score models yielded smaller variability for the effect estimates than using treatment predictors or confounders alone. Therefore, the ideal propensity score model adjusts for $\boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Y}}$ while excluding $\boldsymbol{X}_{\mathcal{Z}}$. The OAL method proposed by Shortreed and Ertefaie [81] intends to achieve this by fitting an adaptive lasso for the treatment [90], where the adaptive weights are computed based on the coefficients from the outcome model, with smaller coefficients corresponding to larger weights. OAL discourages $\boldsymbol{X}_{\mathcal{Z}}$ from being selected and encourages the inclusion of $\boldsymbol{X}_{\mathcal{Y}}$ by imposing heavier penalties on $\boldsymbol{X}_{\mathcal{Z}}$ than $\boldsymbol{X}_{\mathcal{Y}}$. There are two limitations of the work of Shortreed and Ertefaie [81]. The first is that the coefficients of the outcome model were estimated using unpenalized linear regression, which may be fitted for a continuous outcome in the context where the ratio of p to n is relatively large. However, for a binary outcome as considered in our study, it tends to be more difficult for a standard logistic regression to converge. We extend their method by considering a lasso-fitted outcome model to compute the adaptive weights, in which case the coefficients of covariates not predictive of the outcome can be zero, and therefore the covariates with zero coefficients will be excluded from the treatment model. Another limitation is that this idea cannot be conveniently extended to machine learning methods that do not rely on regularization.

OAL incorporates the outcome-covariate associations as adaptive weights when selecting variables for propensity score estimation. An alternative is to directly include the covariates predictive of the outcome in the treatment model. We call the set of variables selected by the outcome model (Model 3.1) `Ysel`. The treatment can then be modeled as a function of `Ysel` (route [2] in Figure 3.1).

One possible problem of using `Ysel` as input for the final treatment model is that the dimensionality of `Ysel` can still remain relatively high for sample size n. To improve propensity score modeling by utilizing the variables predictive of the outcome only while maintaining a reasonable dimensionality, we propose to incorporate the information contained in $\boldsymbol{X}_{\mathcal{Y}}$ into the propensity score estimation process in the form of predicted probabilities of the outcome, Outcome Probabilities (OP) for short, as opposed to directly including those variables in their original form. We estimate OP for each subject by $\hat{p}_i = \exp(\boldsymbol{X}_i^T\hat{\boldsymbol{\theta}})/\{1+\exp(\boldsymbol{X}_i^T\hat{\boldsymbol{\theta}})\}$, and the logit scale of $\hat{p}_i$ is denoted $\hat{p}_i^* = \log\{\hat{p}_i/(1-\hat{p}_i)\}$. OP summarizes the information about the outcome-covariate relationship and reduces the dimensionality of the outcome predictors to a one-dimensional vector. After obtaining `YZsel` which targets $\boldsymbol{X}_{\mathcal{C}}$, we add OP back to the set of input variables to capture the information in $\boldsymbol{X}_{\mathcal{Y}}$ (route [6] in Figure 3.1, and we denote the combination `OP+YZsel`). In addition to `YZsel`, as was noted later in the simulation studies, OP can also be combined with all the available covariates and `Ysel` as input for the final treatment model in Figure 3.1 to improve the effect estimates. We denote the combinations `OP+All` and `OP+Ysel`, which correspond to route [4] and route [5] in Figure 3.1, respectively. We discuss how OP can be used as predictors to estimate the propensity scores for different treatment modeling techniques in Section 3.4, where we outline several possible choices for the final treatment model.

In summary, both `Ysel` and `YZsel` use the outcome model for selecting covariates to account for confounding. Variable selection and propensity score estimation are conducted separately in two steps for methods based on `Ysel`, `YZsel`, `OP+Ysel`, and `OP+YZsel`.

3.4 Possible Choices for the Final Treatment Model for Propensity Score Estimation

We consider five different methods (Table 3.1) for estimating the propensity score given a set of candidate predictors. We leave out the OP for now and only consider the alternatives for the final treatment model for routes [1]-[3] in Figure 3.1. The first is a multinomial logistic regression

model (LOGIS), specified as

$$pr(Z_i = j|\tilde{\boldsymbol{X}}_i) = \frac{\exp(\tilde{\boldsymbol{X}}_i^T \boldsymbol{\alpha}_j)}{\sum_{l=1}^{J} \exp(\tilde{\boldsymbol{X}}_i^T \boldsymbol{\alpha}_l)},$$

where $\tilde{\boldsymbol{X}}_i$ is the set of (possibly selected) variables entering the final treatment model. When the dimension of $\tilde{\boldsymbol{X}}_i$ is sufficiently small, the coefficients can be estimated using the maximum likelihood estimator. In the case where dimension reduction is needed, the estimates $\hat{\boldsymbol{\alpha}}$ can be obtained by minimizing the negative log-likelihood with lasso penalty

$$N^{-1} \sum_{i=1}^{N} \log P(Z_i = j|\tilde{\boldsymbol{X}}_i) + \lambda \sum_{j=1}^{J} \|\boldsymbol{\alpha}_j\|_1,$$

which indicates that each covariate in $\tilde{\boldsymbol{X}}_i$ is associated with all or none of the J levels.

The second is the classification and regression tree (CART) method which performs recursive binary splitting on the feature space in a top-down fashion. At each split, CART agnostically searches for a variable X and a cutpoint such that the response values in each of the resulting nodes lead to the greatest homogeneity [96]. In that sense, CART intrinsically conduct variable selection while growing the tree, as variables that are not predictive of the treatment are less likely to be chosen at each split. We used the Gini index, a measure of the total variance across the J classes, as the metrics for node splitting. Small value of Gini index indicates that observations in this node are predominated by a single class. CART tends to overfit the data. To address the overfitting issue, the common strategy is to first grow a large tree and prune it back in order to retain only part of the tree, as simpler trees tend to be less sensitive to the noises in the data. This method is referred to as pruned CART.

The single tree implementation of both CART and pruned CART, sometimes known as weak learners, may give poor predictions on their own. The ensemble methods, which combine multiple weak learners into one predictive model, have been developed to enhance the predictions. One example is the bootstrap aggregation of the CART algorithm (bagged CART). The bootstrap step of bagged CART involves randomly drawing n observations (i.e., the same size as the original sample) with replacement from the original sample and fitting a CART separately for each bootstrap replicate. The bagging estimates of the probability of subjects being assigned to each class are obtained by averaging the predicted class probabilities from each of the single trees. Another popular ensemble method is the random forests. Similar to bagged CART, random forests build trees based on bootstrap samples of the original observations. What is different from bagged CART is that random forests only considers a random sample of m predictors ($m < p$) at each split, and typically $m \approx \sqrt{p}$ is chosen for classification problems in practice [97]. In our simulation studies

and data example, we choose to grow a relatively large number of trees in order to stabilize the out-of-bag error rate.

3.4.1 Implementation of Incorporating the Predicted Outcome Probabilities

The methods listed in Table 3.1 can be applied to `All`, `Ysel`, and `YZsel` for estimating the propensity scores, which are then used to compute the ATE. The OP is employed for propensity score estimation in different ways for different final treatment models. For the LOGIS method, the propensity scores are estimated by regressing the treatment variable on the union of the input covariates (`All`, `Ysel`, or `YZsel`) and $\hat{p}_i^*$. When the regression model is regularized, no penalty is imposed on $\hat{p}_i^*$. In this way the outcome information is guaranteed to be utilized in the propensity score model.

For the tree-based machine learning methods such as CART, there is no straightforward way to force the OP into the tree growing process, where variable selection is intrinsically conducted. We instead fit a logistic regression model for the treatment as a function of $\hat{p}_i^*$ and the propensity scores obtained in route [1], [2], or [3]:

$$\log \frac{P(Z_i = j|\tilde{\boldsymbol{X}}_i)}{P(Z_i = J|\tilde{\boldsymbol{X}}_i)} = \hat{\boldsymbol{\pi}}(\tilde{\boldsymbol{X}}_i)^T \boldsymbol{\eta}_j + \phi \hat{p}_i^*, \tag{3.3}$$

where $\hat{\boldsymbol{\pi}}(\tilde{\boldsymbol{X}}_i) = \{\hat{\pi}_1(\tilde{\boldsymbol{X}}_i), \cdots, \hat{\pi}_{j-1}(\tilde{\boldsymbol{X}}_i)\}^T$ is the set of propensity scores estimated using the tree-based methods. The coefficients $(\phi, \boldsymbol{\eta}_1^T, \cdots, \boldsymbol{\eta}_{j-1})^T$ can be estimated using maximum likelihood. The final propensity scores that take OP into account are then obtained by calculating the predicted probabilities from model (3.3).

For `OP+Ysel` (route [5]) and `OP+YZsel` (route [6]), the associations between the covariates and the outcome are used twice in the entire estimation process, one for variable selection using the outcome model, and the other for propensity score estimation using the OP.

Methods for Constructing the Treatment Model	Simultaneous Variable Selection and Estimation	Ways to Incorporate OP into the Estimation Process	R package
Logistic regression (standard or penalized)	Yes for penalized logistic regression	Used as a regressor for logistic regression. No penalty is imposed on OP for penalized logistic regression.	glmnet*, gcdnet
CART	Yes	A multinomial logistic regression for the treatment is fitted as a function of estimated propensity scores and OP	rpart*
Pruned CART	Yes	Same as above	rpart*
Bagged CART	Yes	Same as above	ipred*
Random forests	Yes	Same as above	randomForest*, ranger

Table 3.1: Possible choices for the final treatment model.
* R packages used to implement the methods in this paper.

3.5 Simulation Studies

3.5.1 Implementation of Methods under Comparison

The comparative methods were implemented in R with default parameters unless otherwise specified. The lasso algorithm was implemented using the R package *glmnet*. The tuning parameter λ was determined using 10-fold cross-validation with the `lambda.1se` criterion for selecting variables in the pre-selection step, as the goal was to narrow down the number of covariates entering the final treatment model. For LOGIS based on `All`, which does not involve the pre-selection step, the `lambda.min` criterion was used. CART was implemented using the *rpart* package with the complexity parameter (cp) being 0.001, which encourages a large and complex tree structure. For pruned CART, the cp that corresponded to the smallest 10-fold cross-validated error was used to determine the best trimmed tree. Bagged CART was implemented using the *ipred* package with 200 bootstrap replicates. Random forests were implemented using the *randomForest* package with 1000 bootstrap replicates. The minimum size of terminal nodes (the `nodesize` parameter) was set to be 7 in order to make it consistent with the parameters used for CART. For bagged CART and random forests, propensity scores were estimated based on the out-of-bag predictions.

We also extended the OAL approach to three-treatment comparison. Following Shortreed and Ertefaie [81], we considered a set of possible values for the tuning parameter λ_n, $\{n^{-20}, n^{-15}, n^{-10}, n^{-5}, n^{-3}, n^{-1}, n^{-0.75}, n^{-0.5}, n^{-0.25}, n^{0.25}, n^{0.49}\}$, and λ_n was selected by minimizing a weighted absolute mean difference between treatment groups, a quantity that combines the weighted difference in covariates and the absolute values of the coefficients corresponding to the covariates in the outcome model. Since the adaptive weights for the covariates excluded by the outcome model got inflated to infinity, these covariates were not used to fit the adaptive lasso for variable selection and propensity score estimation. In that sense, OAL in the high-dimensional setting is equivalent to a lasso based on `Ysel` with additional weights imposed on the covariates for variable selection.

3.5.2 Simulation Setup

For each simulated dataset, $J = 3$ treatment groups were compared and $p = 100$ covariates were considered, with $|\mathcal{C}|$ confounders, $|\mathcal{Z}|$ related to treatment only, $|\mathcal{Y}|$ related to outcome only, and $|\mathcal{S}|$ spurious predictors. Covariates $\boldsymbol{X}_i$ were generated as follows unless otherwise specified. Half of the covariates (rounded down) in $\boldsymbol{X}_{\mathcal{C}}$, $\boldsymbol{X}_{\mathcal{Z}}$, $\boldsymbol{X}_{\mathcal{Y}}$, and $\boldsymbol{X}_{\mathcal{S}}$ were generated from a binomial distribution with a probability of 0.3, and the other half were generated from a multivariate normal distribution with a p'-dimensional vector of means $\mathbf{0}_{p'}$ and a covariance matrix Σ, where p' is the number of continuous covariates in each subset ($\boldsymbol{X}_{\mathcal{C}}$, $\boldsymbol{X}_{\mathcal{Z}}$, $\boldsymbol{X}_{\mathcal{Y}}$, or $\boldsymbol{X}_{\mathcal{S}}$) and Σ is an identity matrix.

We considered three different simulation settings. Our first setting assumes additivity and linearity for the treatment generating model,

$$Z_i \sim \text{Multinomial}\{\pi_1(\boldsymbol{V}_i), \pi_2(\boldsymbol{V}_i), \pi_3(\boldsymbol{V}_i)\}, \tag{3.4}$$

where $\pi_j(\boldsymbol{V}_i) = \exp(\boldsymbol{V}_i^T\boldsymbol{\alpha}_j)/\sum_{l=1}^3 \exp(\boldsymbol{V}_i^T\boldsymbol{\alpha}_l)$ is the probability of receiving treatment j, with $\boldsymbol{V}_i = (1, \boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{Z}}^T)^T$. We assumed heterogeneous treatment effects on the outcome, and sampled the potential outcome $Y_i^{(j)}$ from a binomial distribution with probability

$$pr\{Y_i^{(j)} = 1|\boldsymbol{W}_i\} = \text{expit}\{\beta_{0j} + \boldsymbol{W}_i^T\boldsymbol{\beta}_j\},$$

where $\boldsymbol{W}_i = (\boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{Y}}^T)^T$, $(\beta_{01}, \beta_{02}, \beta_{03}) = (0, 0.6, 0.4)$, and $\boldsymbol{\beta}_j \propto (1, \cdots, 1)^T$. We considered three levels of sparsity for the treatment and outcome models, with $|\mathcal{C}| = |\mathcal{Z}| = |\mathcal{Y}| = 5$, 10, and 20 for the scenarios with sparse, moderately sparse, and dense models, respectively. As a result, the dimension of $\boldsymbol{\alpha}_j$ differed across the scenarios. The parameter $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T)^T$ was scaled such that $\|\boldsymbol{\alpha}\|_2 = 5$. Similarly, the signal strength of the outcome model was scaled such that $\|\boldsymbol{\beta}_1\| = 3$, $\|\boldsymbol{\beta}_2\| = 2$, and $\|\boldsymbol{\beta}_3\| = 4$. The true values for $E\{Y^(1)\}$, $E\{Y^(2)\}$, and $E\{Y^(3)\}$ were 0.61, 0.71, and 0.68 for the 'sparse' scenario, 0.69, 0.77, and 0.76 for the 'moderately sparse' scenario, and 0.76, 0.83, and 0.83 for the 'dense' scenario. To examine the performance as the sample size increases, we let the sample size n be 500, 1000, and 2000 for the 'sparse' scenario.

Our second setting assumed that models were 'sparse' ($|\mathcal{C}| = |\mathcal{Z}| = |\mathcal{Y}| = 5$) and considered a set of association equations for the treatment assignment that varied in degrees of nonlinearity and nonadditivity. In this case, $\boldsymbol{V}_i$ in (3.4) may contain some transformation of covariates in $(\boldsymbol{X}_{\mathcal{C}}, \boldsymbol{X}_{\mathcal{Y}})$ and/or interaction effects. The structure of $\boldsymbol{V}_i$ are shown in Table C.1. Specifically, we considered scenarios with nonlinear main effects and no interactions (NL), linear main and interaction effects (L-L), nonlinear main effects and linear interactions (NL-L), and nonlinear main and interaction effects (NL-NL). All confounders were continuous in this setting, and the outcomes were generated using the same model as was used in the first setting except that $(\beta_{01}, \beta_{02}, \beta_{03}) = (0, -0.6, 0.4)$.

Our third setting also assumed that the models were sparse and let censoring come into play. Instead of sampling binary outcomes directly from a binomial distribution, we first generated time to event T_i from a logistic distribution with mean function

$$\beta_{01}I(Z_i = 1) + \beta_{02}I(Z_i = 2) + \beta_{03}I(Z_i = 3) + \boldsymbol{W}_i^T\boldsymbol{\beta}$$

and scale parameter $s = 6$, where $\boldsymbol{W}_i = (\boldsymbol{X}_{\mathcal{C}}^T, \boldsymbol{X}_{\mathcal{Y}}^T)^T$, $(\beta_{01}, \beta_{02}, \beta_{03}) = (120, 100, 115)$ and $\boldsymbol{\beta} = (5, \cdots, 5)^T$. Then we obtained the outcome such that $Y_i = I\{T_i < 130\}$. We generated the censoring time C using inverse transform sampling [79] as a function of $\boldsymbol{U}_i = (1, \boldsymbol{X}_{\mathcal{C}})^T$.

Specifically, we assumed a Cox proportional hazards model with the baseline hazard following a Weibull distribution,

$$C_i^{(j)} = \{\lambda^{-1} \exp(\boldsymbol{U}_i \boldsymbol{\gamma}_j)^{-1} \log u\}^{1/\nu},$$

with scale parameter $\lambda = 0.01$ and shape parameter $\nu = 7$, where u was sampled from a Uniform$(0, 1)$ distribution. Note that $\boldsymbol{U}_i$ was assumed to be known in our simulation settings, which resembles our data example where censoring is believed to only depend a low-dimensional set of covariates that can be identified by human experts. The proportion of subject not being observed for $d = 130$ was around 22%.

We considered the six routes displayed in Figure 3.1 for propensity score estimation. We also present the results for three sets of covariates that are usually unknown in practice: confounders only ($\boldsymbol{X}_{\mathcal{C}}$), treatment predictors ($\boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Z}}$), and outcome predictors ($\boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Y}}$). These results were used to illustrate the impact of different groups of covariates on the treatment effect estimates. The ATE were estimated using the IPW method.

3.5.3 Construction of Confidence Intervals

The confidence intervals (CI) were constructed using bootstrap standard errors based on 200 bootstrap replicates. For the settings without censoring, we considered two possible bootstrap procedures. The first applies variable selection to each bootstrap replicate and refits the models for the treatment and outcome using variables selected in each replicate, and we refer to it as usual bootstrap. The second ignores the variability due to selection for covariates. Instead, for LOGIS based on `All` and `OP+All` and all tree-based methods, OP and propensity scores are directly bootstrapped from the OP and propensity scores obtained in the original sample, without refitting the model. For LOGIS based on `Ysel`, `YZsel`, `OP+Ysel`, and `OP+YZsel`, propensity scores are obtained by refitting the final treatment model using variables selected in the original data set. As a result, the usual bootstrap is much more computationally intensive than the modified bootstrap. A trial simulation study under the scenario of linear sparse models for sample sizes of 500 (Table 3.2) and 1000 (Table C.2) showed that usual bootstrap tended to overestimate the standard errors and produce overly conservative CIs for the tree-based methods. For example, most of the coverage rates for the methods that involved OP were above 98%. Over-coverage was also observed for (unpenalized) LOGIS for usual bootstrap. On the other hand, LOGIS based on `OP+All` had close to nominal coverage using usual bootstrap at the expense of high computational burden, and slight over-coverage using modified bootstrap. For the tree-based methods, standard errors based on modified bootstrap were close to their corresponding Monte Carlo standard deviation. In this case, modified bootstrap remedied the overestimation of the usual bootstrap by dropping the

variability due to variable selection. Therefore, we proceed with the modified bootstrap technique for our simulation studies, which directly samples the estimated propensity scores and OP from the original simulated data set for each bootstrap replicate. For the third setting where censoring existed, we again used modified bootstrap, but with the censoring model refitted for each bootstrap sample. Metrics that were used to compare the various propensity score estimation methods included bias, Monte Carlo standard deviations, standard errors, root mean squared error (RMSE), and coverage rate of 95% CIs. True values were determined using 5×10^5 replicates.

	Empirical SD			SE (usual)			Coverage (%; usual)			SE (modified)			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
OAL	54	48	51	57	51	53	96.1	96.0	95.7	55	50	52	94.1	95.6	94.2
LOGIS															
All	52	51	50	56	54	53	59.4	73.2	93.4	55	55	52	59.8	75.9	93.3
Ysel	57	49	54	94	82	85	100	99.7	99.9	58	51	55	95.4	96.0	94.8
YZsel	58	52	54	80	70	74	99.7	99.1	99.4	58	53	55	95.0	95.5	95.0
OP + All	51	46	50	57	52	54	95.8	96.9	95.5	58	56	56	96.3	97.8	96.9
OP + Ysel	57	49	54	94	82	85	100	99.7	99.9	58	51	55	95.4	96.0	94.8
OP + YZsel	55	48	53	80	69	73	99.8	99.4	99.3	55	49	53	95.2	96.0	94.3
CART															
All	79	81	76	95	96	89	78.1	87.4	97.0	78	80	74	63.2	75.9	92.0
Ysel	72	70	67	87	86	82	90.9	94.4	97.5	72	72	69	81.6	88.1	94.0
YZsel	69	66	64	83	83	79	93.1	96.4	99.0	70	68	67	86.2	90.8	94.8
OP + All	80	74	79	101	93	102	98.8	98.6	98.9	79	74	79	93.4	93.8	94.2
OP + Ysel	67	61	65	87	81	88	99.4	99.0	98.7	66	63	66	94.2	94.4	94.8
OP + YZsel	62	58	62	82	77	84	99.3	99.5	99.4	63	59	62	94.3	94.7	94.8
Pruned CART															
All	62	63	56	92	92	86	75.3	90.4	99.1	61	61	56	40.1	60.4	90.4
Ysel	63	60	55	85	84	79	87.0	93.2	99.0	60	60	56	61.0	75.5	92.4
YZsel	64	61	55	82	81	76	87.2	93.5	99.3	61	60	57	68.6	80.7	93.0
OP + All	58	53	56	95	87	96	99.8	99.5	99.9	58	52	57	94.3	94.5	94.6
OP + Ysel	56	49	54	83	77	84	100	99.5	99.7	55	50	54	94.2	94.8	94.6
OP + YZsel	54	50	53	78	73	80	99.8	99.6	99.6	55	50	54	95.7	94.9	94.7
Bagged CART															
All	54	53	51	45	45	42	28.9	46.9	85.3	56	57	53	44.1	65.5	91.2
Ysel	71	63	63	54	51	49	87.3	88.7	86.2	71	66	67	93.4	95.6	95.4
YZsel	91	82	81	63	59	57	80.0	82.1	80.9	86	79	81	91.6	92.5	94.8
OP + All	53	46	52	86	84	90	100	99.9	99.9	52	47	51	94.2	95.4	94.4
OP + Ysel	50	45	49	80	80	85	100	99.9	100	50	45	50	94.6	95.2	95.0
OP + YZsel	50	45	49	77	77	82	100	99.8	99.9	50	46	50	94.6	95.0	94.6
Random Forests															
All	49	50	48	40	40	38	13.4	30.9	80.4	52	53	50	27.1	51.2	90.0
Ysel	57	52	51	41	41	38	68.4	76.2	84.7	59	57	57	84.9	91.2	95.9
YZsel	94	78	79	46	44	41	66.8	75.7	72.0	79	70	75	91.0	91.2	94.5
OP + All	55	47	53	61	56	59	97.5	98.4	97.2	53	48	52	94.8	94.9	94.0
OP + Ysel	51	45	50	70	66	71	99.9	99.5	99.4	51	46	50	94.5	94.8	94.4
OP + YZsel	50	46	50	79	76	82	99.9	99.7	99.9	51	46	50	94.8	95.0	94.7

Table 3.2: Standard errors (SE) and coverage of 95% confidence intervals estimated by usual bootstrap and modified bootstrap for sample size of 500. The scenario with sparse treatment models was considered. Results were obtained based on 1000 simulated datasets. For each dataset, 200 bootstrap samples were generated.

3.5.4 Simulation Results

We present the box plots of bias for the IPW estimates across the simulation settings in Figures 3.2-3.4. The numerical results for all evaluation metrics are reported in the Supplementary Materials.

3.5.4.1 Bias

The numerical results for the setting of linear treatment models are presented in Tables C.3-C.9 in the Supplementary Materials. LOGIS that adjusted for confounders ($\boldsymbol{X}_{\mathcal{C}}$), treatment predictors ($\boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Z}}$), and outcome predictors ($\boldsymbol{X}_{\mathcal{C}} \cup \boldsymbol{X}_{\mathcal{Y}}$) had close to zero empirical bias as expected (e.g., Table C.3). For the CART family methods, using confounders only generally yielded smaller bias than using treatment predictors, outcome predictors, or `All`. Methods based on pre-selected predictors (`Ysel` and `YZsel`) tended to yield smaller empirical bias than methods using all the available predictors as input (Figure 3.2), which indicates that excluding noise variables before fitting the final treatment model can help remove the bias, though the absolute bias was still greater than zero for the tree-based methods. One exception was that random forests based on `YZsel` resulted in substantial bias in the 'sparse' scenario, possibly because there were too few true predictors at each split for random forests to choose from. The OAL method had similar performance in terms of bias to LOGIS based on `OP+Ysel` and `OP+YZsel` in the 'sparse' scenario (Tables C.3-C.5), while the latter outperformed the former in the 'moderately sparse' (Table C.6) and 'dense' scenarios (Tables C.8 and C.9).

When the outcome information was not taken into account, LOGIS in general produced less biased estimates than the tree-based methods for each of the routes [1]-[3] in Figure 3.1. Bias for `All`, `Ysel`, and `YZsel` were getting closer as the model became 'denser' (Figure 3.2). As sample size increased from 500 to 2000 for the 'sparse' scenario (Figure 3.3), the nonzero empirical bias persisted across the methods considered, which illustrates the slow convergence rate of lasso. The inclusion of OP greatly reduced the bias of the estimates in the case where methods based on `All`, `Ysel`, or `YZsel` resulted in larger than zero absolute bias across the simulation settings, which highlights the robustness of the treatment effect estimator provided by incorporating OP into the estimation process.

With nonlinearity and nonadditivity in the treatment model, the performance of LOGIS was not inferior to the tree-based methods in terms of bias (Figure 3.4 and Tables C.10-C.17), possibly because in this case, LOGIS approximated nonlinear functions reasonably well. The good approximation of linear methods to nonlinear functions was also observed in Tu [87], where multivariable linear regression yielded smaller bias than bagged CART and random forests in some cases with nonlinear and nonadditive associations in the treatment models.

When censoring existed, LOGIS based on `Ysel`, `YZsel`, `OP+Ysel`, and `OP+YZsel` pro-

duced much lower bias than their CART family counterparts. One exception was `YZsel`-based bagged CART, which had close to zero bias (Table C.18).

3.5.4.2 Statistical Efficiency and RMSE

In general, estimates produced by routes [4], [5], and [6] had smaller variability than those resulting from routes [1], [2] and [3], respectively, for each method across all the settings, which illustrates the advantage of leveraging the information about the outcome model in terms of statistical efficiency.

Methods based on `OP+All`, `OP+Ysel`, and `OP+YZsel` had smaller variabilities and RMSE of the effect estimates across the simulation settings compared to methods based on `All`, `Ysel`, and `YZsel`, respectively (Figures B.1-B.3). For example, for the scenario with moderately sparse models, the percentage of reduction in RMSE ranged from 1.6% to 20.6% for LOGIS based on `OP+YZsel` compared to `YZsel`, and from 26.3% to 41.6% for random forests based on `OP+YZsel` compared to `YZsel` (Table C.6). The variabilities for `OP+All`, `OP+Ysel`, and `OP+YZsel` were close to one another, and one was not uniformly smaller than the other two. In general, regardless of whether OP was incorporated, CART had the largest variability and RMSE among the methods for all scenarios considered.

In the presence of censoring, OAL and LOGIS based on `OP+YZsel` had the smallest RMSE among the methods under comparison (Table C.18). The inclusion of OP reduced the variability of the effect estimates for all method considered. The bias for LOGIS, CART, and pruned CART decreased when the OP was taken into account, with `OP+YZsel` resulting in the smallest bias for each method. However, we still observed residual bias, especially for the CART and pruned-CART.

3.5.4.3 Coverage of 95% CI

For the tree-based methods and unpenalized LOGIS, standard errors obtained using the modified bootstrap were close to the corresponding Monte Carlo standard deviations in the scenarios with 'sparse' models (regardless of whether the treatment models were linear and/or additive). The modified bootstrap tended to slightly overestimate the variability for LOGIS with lasso penalty. For example, the ratio of standard errors to Monte Carlo standard deviations ranged from 1.12 to 1.21 for `OP+All` in the scenario with linear associations and 'sparse' representation of the models for sample size of 500. The tree-based methods achieved close to nominal coverage of 95% for `OP+All`, `OP+Ysel`, and `OP+YZsel` for most of the scenarios considered. In the case where the coverage fell below the nominal level, for example the 'dense' scenario (Table C.8), the under-coverage was mainly caused by the empirical bias rather than the underestimation of the standard errors.

3.6 Data Analysis

We applied the algorithms considered in Figure 3.1 and the methods considered in Section 3.4 to the same data set of prostate cancer patients from the Optum Clinformatics Data Mart in Chapter 2 to compare the adverse effects of the four drugs for mCRPC. The inclusion and exclusion criteria used to identify our analytic cohort is described in Chapter 2. In the previous analysis, the working model for the outcome was specified as

$$\log \frac{pr(Y_i = 1|\boldsymbol{A}_i, \boldsymbol{B}_i)}{pr(Y_i = 0|\boldsymbol{A}_i, \boldsymbol{B}_i)} = \boldsymbol{A}_i^T \boldsymbol{\beta}_A + \boldsymbol{B}_i^T \boldsymbol{\beta}_B,$$

where $\boldsymbol{A}_i$ contained the sociodemographic factors and other relevant covariates, and $\boldsymbol{B}_i$ contained five pre-existing comorbid conditions, including diabetes, hypertension, arrhythmia, congestive heart failure, and osteoporosis. Specifically, $\boldsymbol{A}_i$ included age, race, education level, household income, geographic region, insurance product type, whether the insurance plan is administrative services only (ASO), metastatic status of cancer, year of first prescription, and provider type [58]. In this analysis, we increased the granularity of the comorbid conditions and considered a list of phenotype codes (phecodes), which are aggregations of the International Classification of Diseases (ICD) codes that represent clinically meaningful phenotypes [98]. We matched the ICD codes in the claims data to the list of phecodes and identified 1042 phecodes as the original reservoir of predictors. These 1042 phecodes represent 16 broad categories of diseases (circulatory system, congenital anomalies, dermatological diseases, endocrine/metabolic diseases, genitourinary diseases, hematopoietic diseases, infectious diseases, injuries and poisonings, mental disorders, musculoskeletal diseases, neoplasms, neurological diseases, respiratory diseases, sense organs, and symptoms). The working model for the outcome became

$$\log \frac{pr(Y_i = 1|\boldsymbol{A}_i, \boldsymbol{M}_i)}{pr(Y_i = 0|\boldsymbol{A}_i, \boldsymbol{M}_i)} = \boldsymbol{A}_i^T \boldsymbol{\beta}_A + \boldsymbol{M}_i^T \boldsymbol{\beta}_M,$$

where $\boldsymbol{M}_i$ denotes the list of phecodes of dimension 1042. In the pre-selection step, a lasso was fitted to select the phecodes that are predictive of the outcome, with the coefficients $\boldsymbol{\beta}_A$ unpenalized. Covariates predictive of the treatment were selected in a similar manner using a multinomial logistic regression with lasso-type penalty. In other words, covariates in $\boldsymbol{A}_i$ (such as age, race, and household income) are always adjusted for in the treatment and outcome models. The six routes in Figure 3.1 were followed to obtain six different sets of estimated propensity scores. Note that standard multinomial logistic regression models adjusting for `Ysel` and `YZsel` yielded substantial standard errors for the estimates of ATE (results not shown). Instead, we fitted the models using the lasso-type penalty (with $\boldsymbol{\beta}_A$ not being penalized) to reduce the variability of the estimates. For

censoring, we fitted a Cox model adjusting for $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$, which was a low-dimensional set of covariates. All covariates were coded binary, and the covariates with more than two levels were represented by dummy variables. The standard errors and CIs were constructed using the modified bootstrap procedure.

3.6.1 Data Analysis Results

The sample sizes for the ER visit cohort and hospitalization cohort were 7678 and 7709, respectively. The descriptive statistics of the data and the proportions of patients being censored for each treatment group are summarized in Chapter 2. The overall proportions of patients being censored within 180 days and 360 days were 20.8% and 32.6%, respectively, for ER visits, and 24.6% and 40.6%, respectively, for hospitalization. Sipuleucel-T group had larger percentage of censored patients than the other three groups.

Propensity scores estimated by different routes in Figure 3.1 were highly correlated for each treatment level (results not shown). In general, we observed larger correlations among propensity scores estimated by LOGIS than those estimated by the tree-based methods. For example, correlations between LOGIS based on `YZsel` and `OP+YZsel` ranged from 0.97 to 1, while the correlations between random forests based on `YZsel` and `OP+YZsel` ranged from 0.93 to 0.99.

Numbers of phecodes in each disease group selected by the outcome and/or the treatment model in the pre-selection step for each of the two endpoints are reported in Tables C.19-C.22. For example, 54 phecodes were selected by the outcome model and 69 were selected by the treatment model, with 12 lying in the intersection for ER visits within 180 days. Among the 12 phecodes predictive of both treatment and outcome, 4 were associated with neoplasm (cancer of prostate, secondary malignancy of respiratory organs, secondary malignant neoplasm, secondary malignant neoplasm of liver), 2 were associated with circulatory system (congestive heart failure NOS and congestive heart failure nonhypertensive), 2 were associated with mental disorders (delirium dementia and amnestic and other cognitive disorders, tobacco use disorder), 1 was associated with endocrine/metabolic system (type 2 diabetes), 1 was associated with hematopoietic system (lymphadenitis), 1 was associated with respiratory system (abnormal findings examination of lungs), and 1 was associated with symptoms (nausea and vomiting). For hospitalization within 180 days, 63 and 79 phecodes were selected by the outcome and treatment model, respectively, and the intersection contained 10 phecodes. Among the 10 phecodes, 7 were overlapped with those identified for ER visits within 180 days (congestive heart failure NOS, congestive heart failure nonhypertensive, lymphadenitis, cancer of prostate, secondary malignant neoplasm, secondary malignant neoplasm of liver, and nausea and vomiting), with 3 additional phecodes associated with neoplasm (cancer of stomach, malignant neoplasm of head, face, and neck, and secondary malignancy of

respiratory organs). We also note that a number of phecodes for genitourinary system were identified to be related either to the treatment (e.g., acute cystitis, chronic renal failure [CKD], nephritis, nephrosis, renal sclerosis, other disorders of the kidney and ureters, renal failure, retention of urine, and urinary tract infection) or to the outcome (e.g., chronic kidney disease stage IV, functional disorders of bladder, hyperplasia of prostate, lump or mass in breast, other disorders of prostate, and prostatitis), while the intersection was empty, possibly due to the fine granularity of the phecodes.

Figure 3.5 showed results for (penalized) LOGIS, CART, and Bagged CART for the 180-day risks differences in ER visits among the four treatment groups. Docetaxel users exhibited significantly higher 180-day risks of at least one ER visit than the users of the two oral drugs (abiraterone and enzalutamide), a finding that is consistent with previous studies [69, 80]. The 180-day risk differences between abiraterone and enzalutamide users were generally not significant, except that some of the results yielded by random forests indicated significantly lower risk for the enzalutamide group (Figures 3.5 and B.4). For the 360-day time window, enzalutamide users showed significantly lower risk of ER visits in most cases (Figures B.5 and B.6).

Similar directional results among the four treatment groups were observed for 180-day and 360-day risks in hospitalization. In particular, patients who received enzalutamide as their first-line therapy had significantly lower risk of hospitalization than those who received abiraterone, which is consistent with the findings of a study based on a French insurance system database [99].

As was observed in the simulation studies, bagged CART and random forests using `YZsel` as input yielded estimates with large standard errors. In general, incorporating the OP into the treatment model reduced the variability of the estimates and led to narrower 95% CIs. Greater efficiency gains were noted for the CART family methods compared to LOGIS. For example, for the 360-day risk of ER visits, the ratios of CI widths of `OP+All` over `All` ranged from 0.98 to 1 and from 0.61 to 0.66 for LOGIS and bagged CART, respectively. When the OP were not included in the treatment model, the estimates yielded by LOGIS tended to have lower variance than those produced by CART family methods. For example, for the 360-day risk of ER visits, the ratios of CI widths of bagged CART over CI widths of LOGIS using `All` as input ranged from 1.06 to 2.01. Consistent with what was observed in the simulation studies, the point estimates and confidence widths of `OP+Ysel` and `OP+YZsel` were very close across the propensity score estimation methods.

3.7 Discussion

In this paper, we examined the traditional multinomial logistic regression method and a set of machine learning techniques for propensity score estimation. We presented how the outcome-covariate associations can be used for variable selection as well as propensity score estimation for

the methods considered. The idea of using outcome models to improve the effect estimates has been extensively explored in the literature. However, the use of OP to improve the propensity score estimation (and therefore the estimation of average treatment effects) has not been studied to be best of our knowledge. We conducted simulation studies to evaluate their finite-sample performance in estimating the ATE under the scenarios where the ratio of the number of candidate covariates to the sample size was relatively large. Simulation studies show that simultaneous variable selection and propensity score estimation (i.e., methods based on `All`) in a high-dimensional setting led to substantial bias for the LOGIS method, possibly because of the slow convergence rate resulting from the large number of noise variables. Similar pattern was observed for the tree-based methods. We showed that the inclusion of OP can improve the robustness and statistical efficiency of the treatment effect estimators. If the variable selection step had satisfactory performance in terms of identifying the set of important confounders and controlling for the bias, then the benefits of including OP in terms of bias reduction may be minimal. On the other hand, if methods based on `All`, `Ysel`, and `YZsel` produced biased estimates, then further adjusting for OP in the treatment model can help reduce the bias. OP alleviates the bias by adding back the information about the confounders that are potentially missed by the variable selection procedure.

The LOGIS method (both standard and penalized depending on the dimension of the covariates) outperformed the tree-based methods when the association equation for the treatment model was linear, and performed reasonably well under conditions of nonlinearity and/or nonadditivity, especially when the OP was used. When only the treatment-covariate relationship was considered, nonparametric machine learning methods such as bagged CART and random forests can sometimes produce less biased estimates than multinomial logistic regression, which suggests tree-based methods as promising alternatives to multinomial logistic regression in the presence of interactions and/or nonlinearity. The performance of tree-based methods may be improved by optimizing the tuning parameters, such as minimum size of terminal nodes, maximum number of terminal nodes, and number of trees to grow. In addition, standard cross-validation procedure, which focus on out-of-sample performance, is often used to optimize tuning parameters for accurate predictive performance in practice and may not have desired characteristics for selecting tuning parameters for causal inference (i.e., unbiased treatment effect estimates) [100]. The selection criteria for the tuning parameters in the context of treatment effect estimation could be a topic of future work.

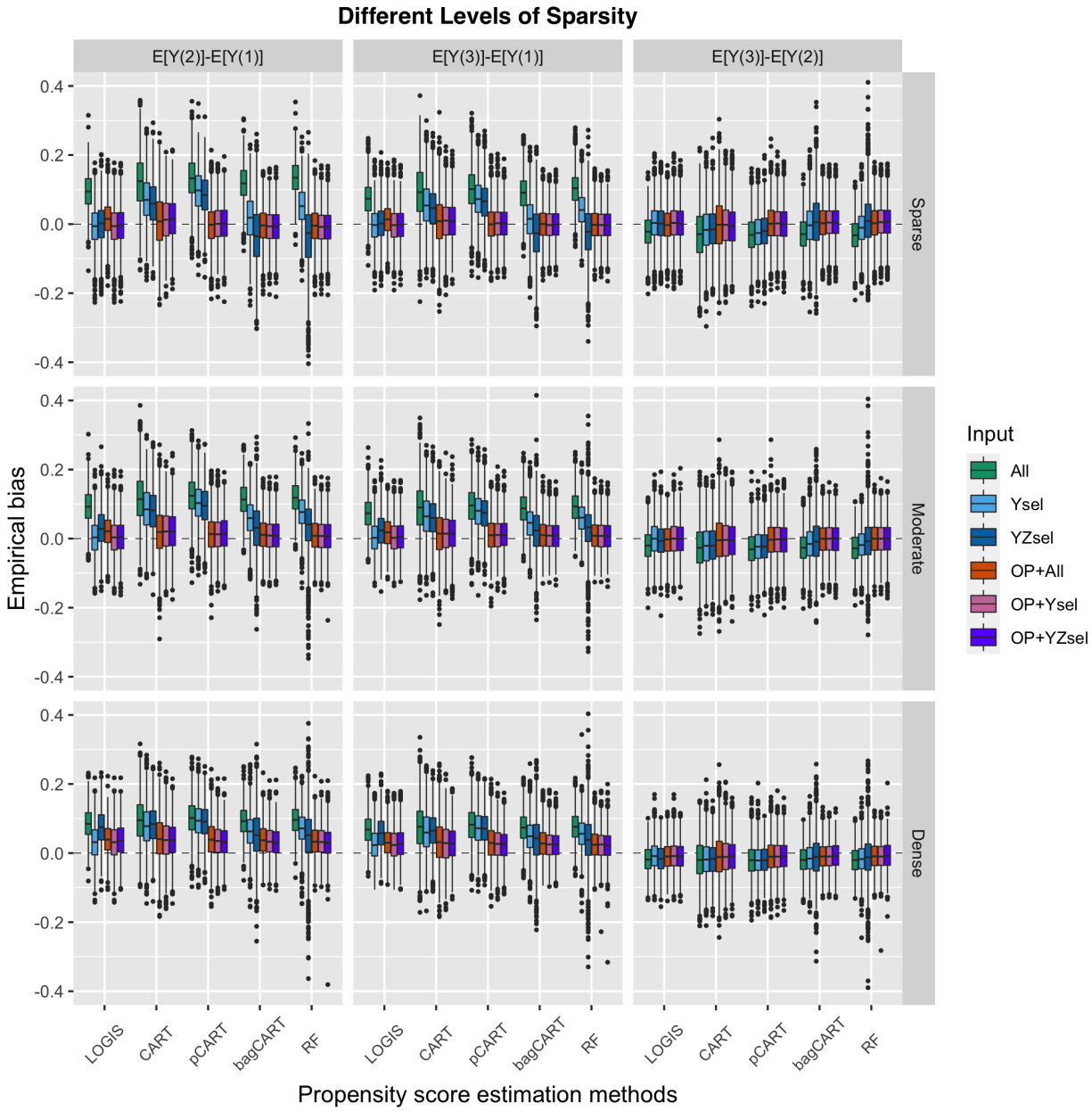

Figure 3.2: Box plots of empirical bias for 2000 inverse probability weighted estimates for the ATE under scenarios with different levels of sparsity. The rows represent scenarios and columns represent treatment pairs. Each simulated dataset contained 500 samples.

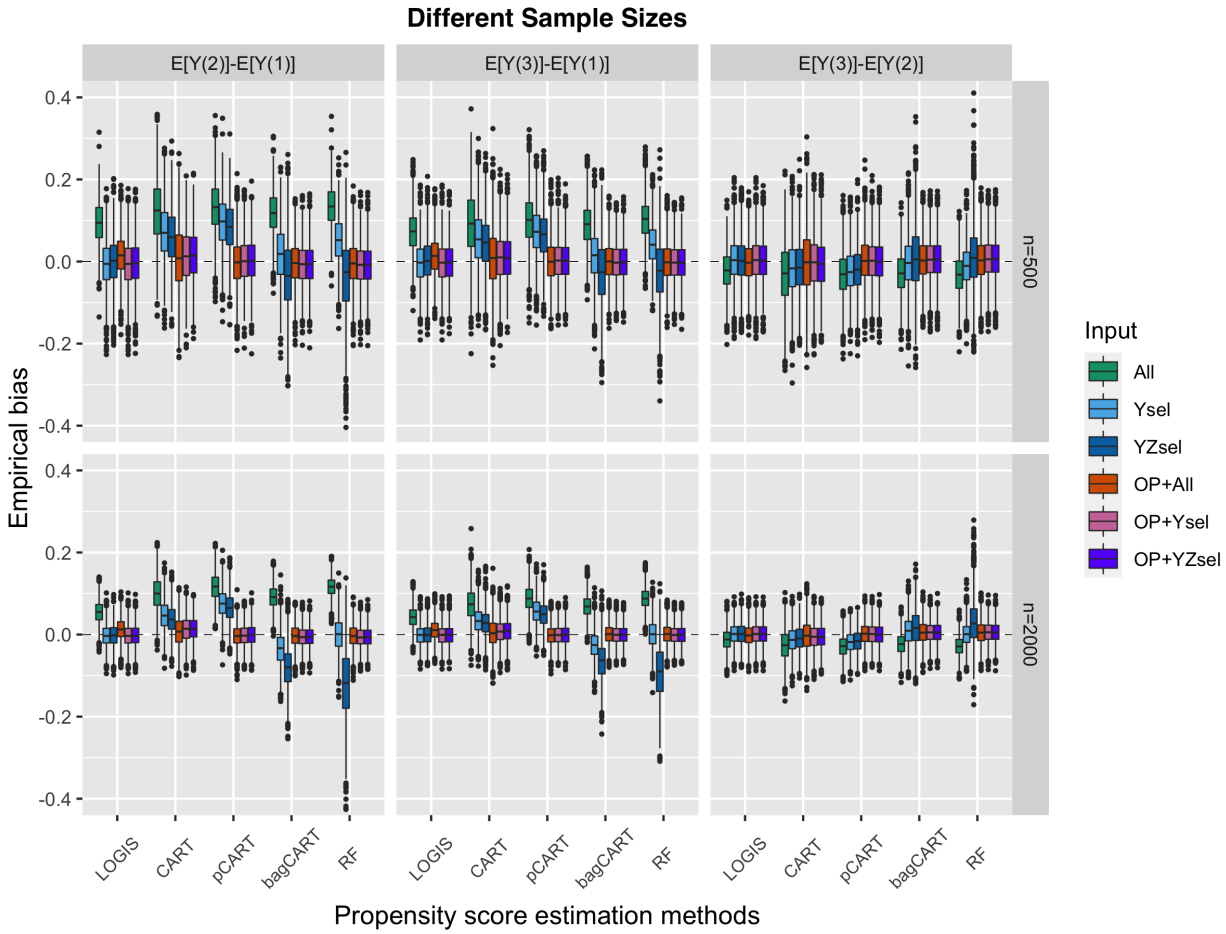

Figure 3.3: Box plots of empirical bias for 2000 inverse probability weighted estimates for the ATE for different sample sizes. The scenario with sparse treatment and outcome models was considered. The rows represent sample sizes and columns represent treatment pairs.

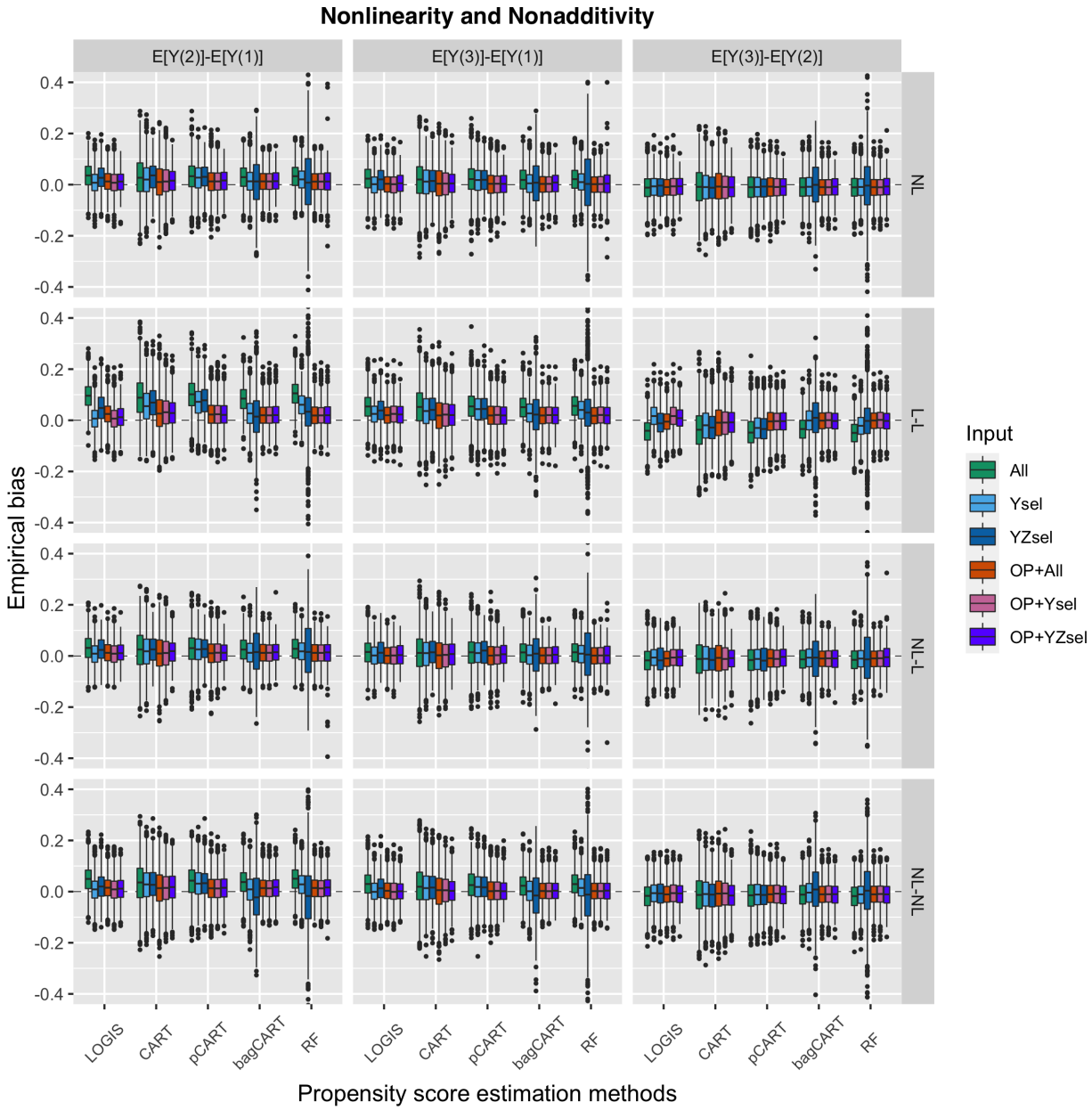

Figure 3.4: Box plots of empirical bias for 2000 inverse probability weighted estimates for the ATE under scenarios with various degrees of nonlinearity and nonadditivity in the treatment generating model. The rows represent scenarios and columns represent treatment pairs. Each simulated dataset contained 500 samples.

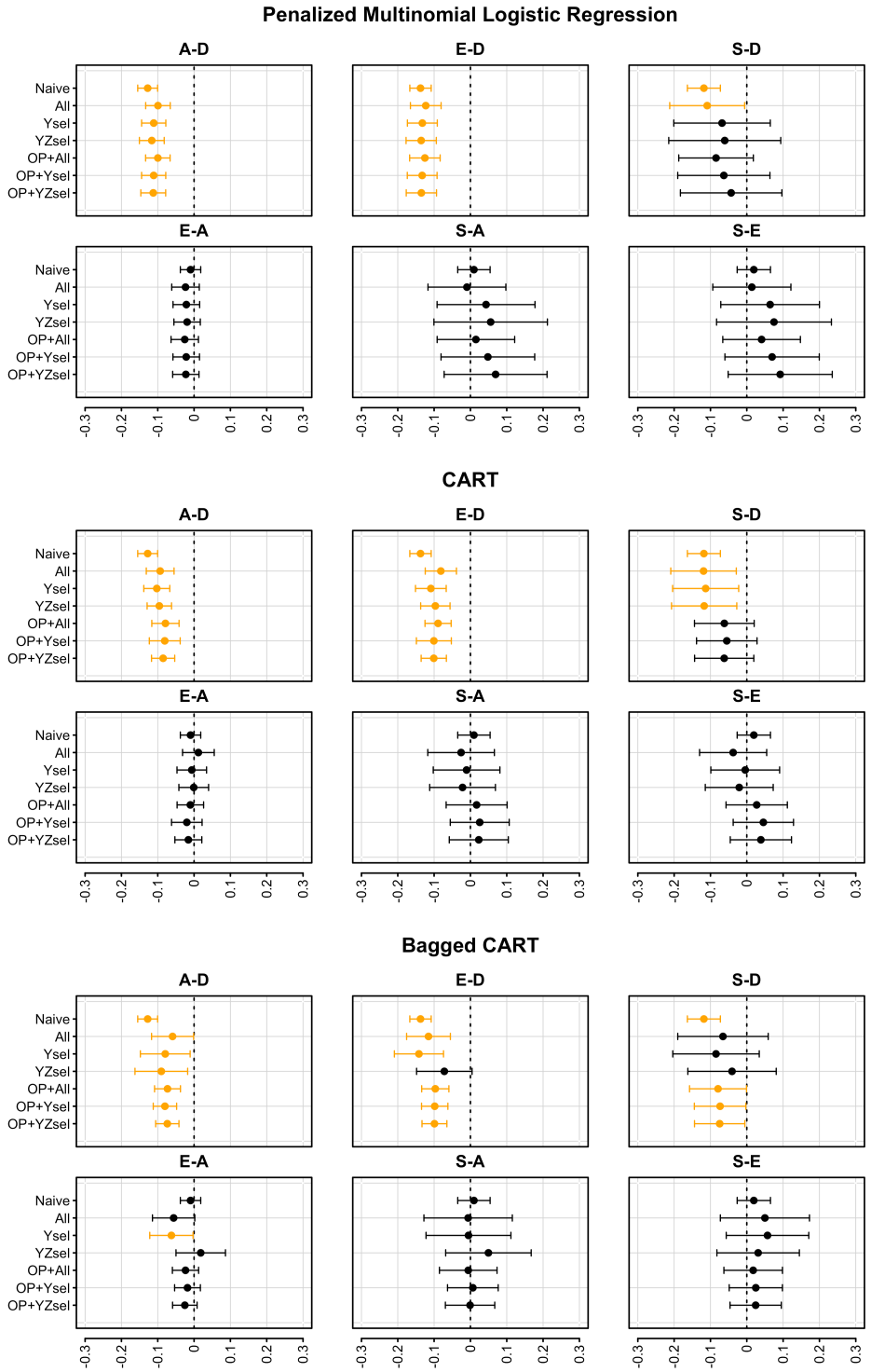

Figure 3.5: Average treatment effects for ER visits within 180 days of treatment initiation for LOGIS, CART, and bagged CART. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7678$ ($N_A = 2757$, $N_D = 2311$, $N_E = 2043$, $N_S = 567$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

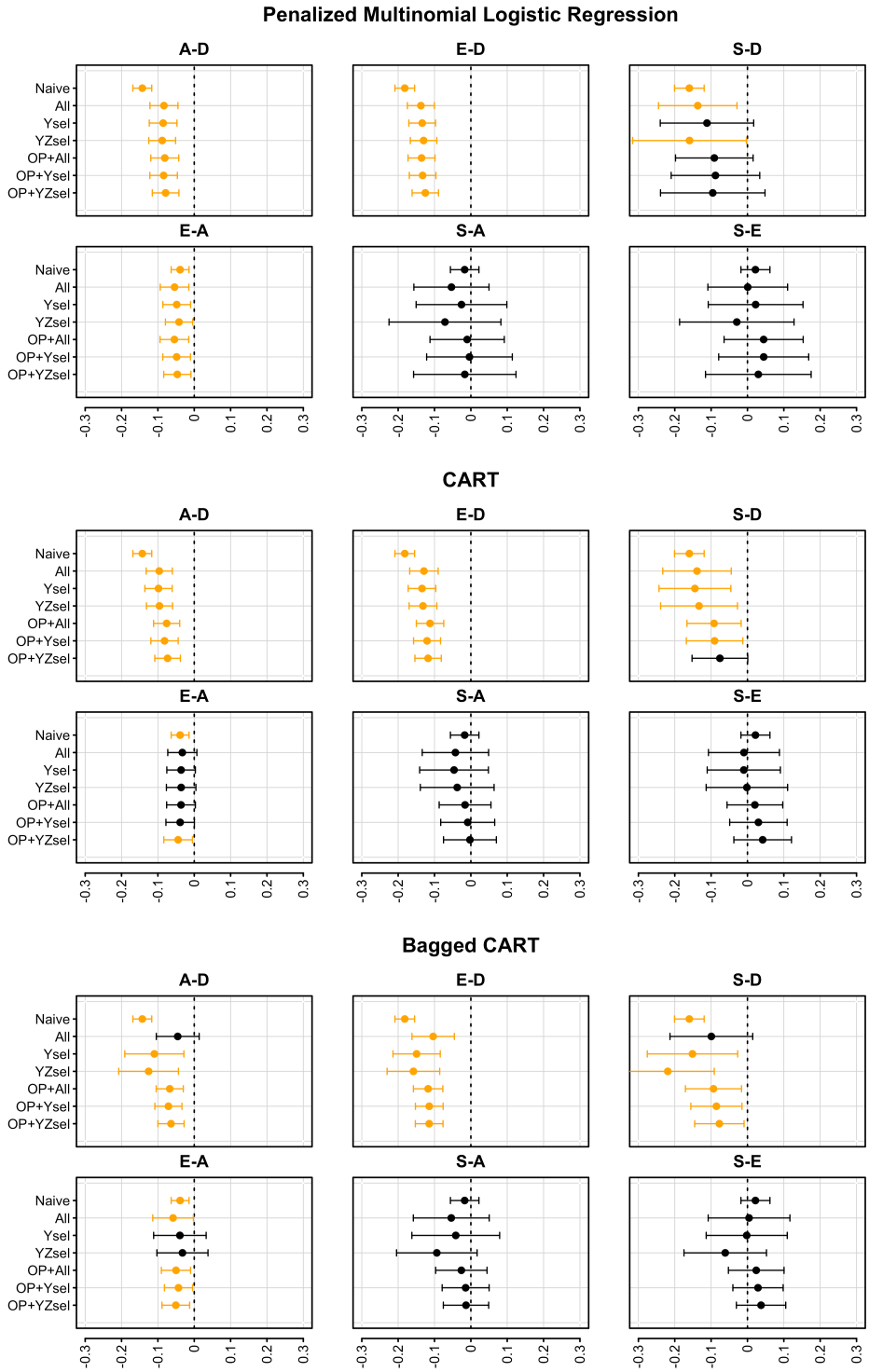

Figure 3.6: Average treatment effects for hospitalization within 180 days of treatment initiation for LOGIS, CART, and bagged CART. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7709$ ($N_A = 2766$, $N_D = 2320$, $N_E = 2051$, $N_S = 572$. Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

APPENDIX A

Supplement for Chapter I

A.1 Covariate Balance Checking

To check for balance in covariates for PEN-GAM in our case with four treatment groups, we fit a generalized linear model for each covariate X including splines of generalized propensity scores and treatment groups as predictors. Then we used a likelihood ratio test to conduct a 3-degree-of-freedom global test on all of the treatment coefficients being zero. Categorical covariates with more than two levels were represented with multiple indicator variables, and statistical tests were conducted on each indicator. P-values before and after adjusting for splines of propensity scores for each covariate are reported in Table A.1.

Let s_z^2 be the variance of covariate X in treatment group z and $\bar{X}_p$ be the mean of X in the target population. To conduct balancing checking for weighting-based and matching-based estimators, we followed Li and Li [26] and inspected the absolute standardized difference in means between each treatment group and the target population, $d = |\bar{X}_z - \bar{X}_p|/s$, where $s^2 = J^{-1}\sum_{z=1}^{J} s_z^2$. For weighting-based estimators (IPW, MW, and OW), $\bar{X}_z$ is the weighted average of X from the zth group, while for matching-based estimators (MCOV, MGPSV, and MGPSS), $\bar{X}_z$ is the group-specific mean of X after imputation. Supplemental Figures 1.1 and 1.2 present the absolute standardized differences for each covariate for weighting-based and matching-based estimators, respectively. The results for the naive estimator indicate the presence of large imbalance for some covariates across treatment groups (e.g., age, provider type, etc.). IPW balanced the covariates well for docetaxel, abiraterone, and enzalutamide, but not for sipuleucel-T. The absolute standardized differences for MW and OW remained small in general for all treatment groups. Improvement of balance was achieved using matching-based estimators for all treatment groups other than sipuleucel-T, with MGPSV and MGPSS performing slightly better than MCOV.

A.2 Supplemental Tables

Covariates		Before adjusting	After adjusting
Age			
	<65	0.038	0.997
	65-74	<0.001	0.997
	≥75	<0.001	0.988
Race			
	White	0.969	0.998
	Black	0.043	0.990
	Other	0.322	0.998
Education level			
	High School Diploma or Less	0.055	0.987
	High School Graduate and Less than Bachelor Degree	0.936	0.998
	Bachelor Degree Plus	0.360	0.991
	Unknown	0.001	0.983
Household income range			
	<50k	<0.001	0.993
	50k-100k	0.351	0.987
	>100k	<0.001	0.956
	Unknown	0.006	0.997
Geographic Region			
	South Atlantic	0.915	0.998
	New England	0.510	0.998
	Middle Atlantic	0.155	0.953
	East North Central	0.693	0.980
	East South Central	0.314	0.996
	West North Central	<0.001	0.916
	West South Central	0.767	0.998
	Mountain	0.266	0.984
	Pacific	<0.001	0.994
Product			
	HMO	0.562	0.993
	PPO	0.422	0.970
	Other	0.423	0.996
Metastatic (Yes)		0.024	0.968
ASO (Yes)		0.686	0.998
Year of First Prescription			
	2014	<0.001	0.994
	2015	0.053	0.999
	2016	0.090	0.999
Diabetes		0.282	0.998
Hypertension		0.462	0.999
Arrhythmia		0.571	0.998
CHF		0.034	0.981
Osteoporosis		0.009	0.950
Provider Type			
	Medical oncologist	<0.001	0.873
	Others	<0.001	0.873

Table A.1: P-values of likelihood ratio test (3-df) on all treatment coefficients being zero before and after adjusting for splines of propensity scores in a model of covariate. Categorical covariates with more than two levels were represented with multiple indicator variables, and statistical tests were conducted on each indicator.

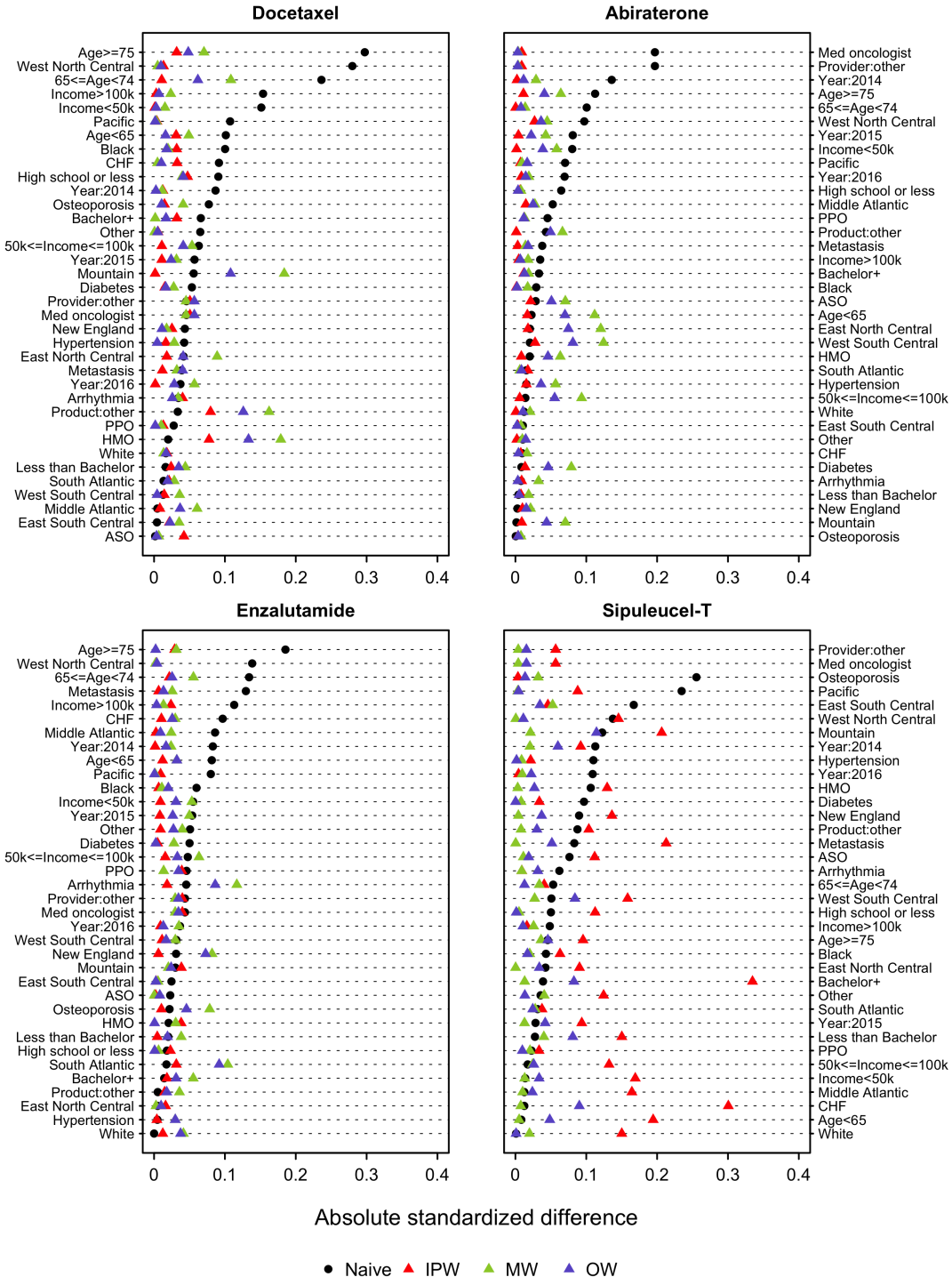

Figure A.1: Group-specific absolute standardized differences for weighting-based methods. Levels of covariates were sorted by the absolute standardized difference for the naive estimator.

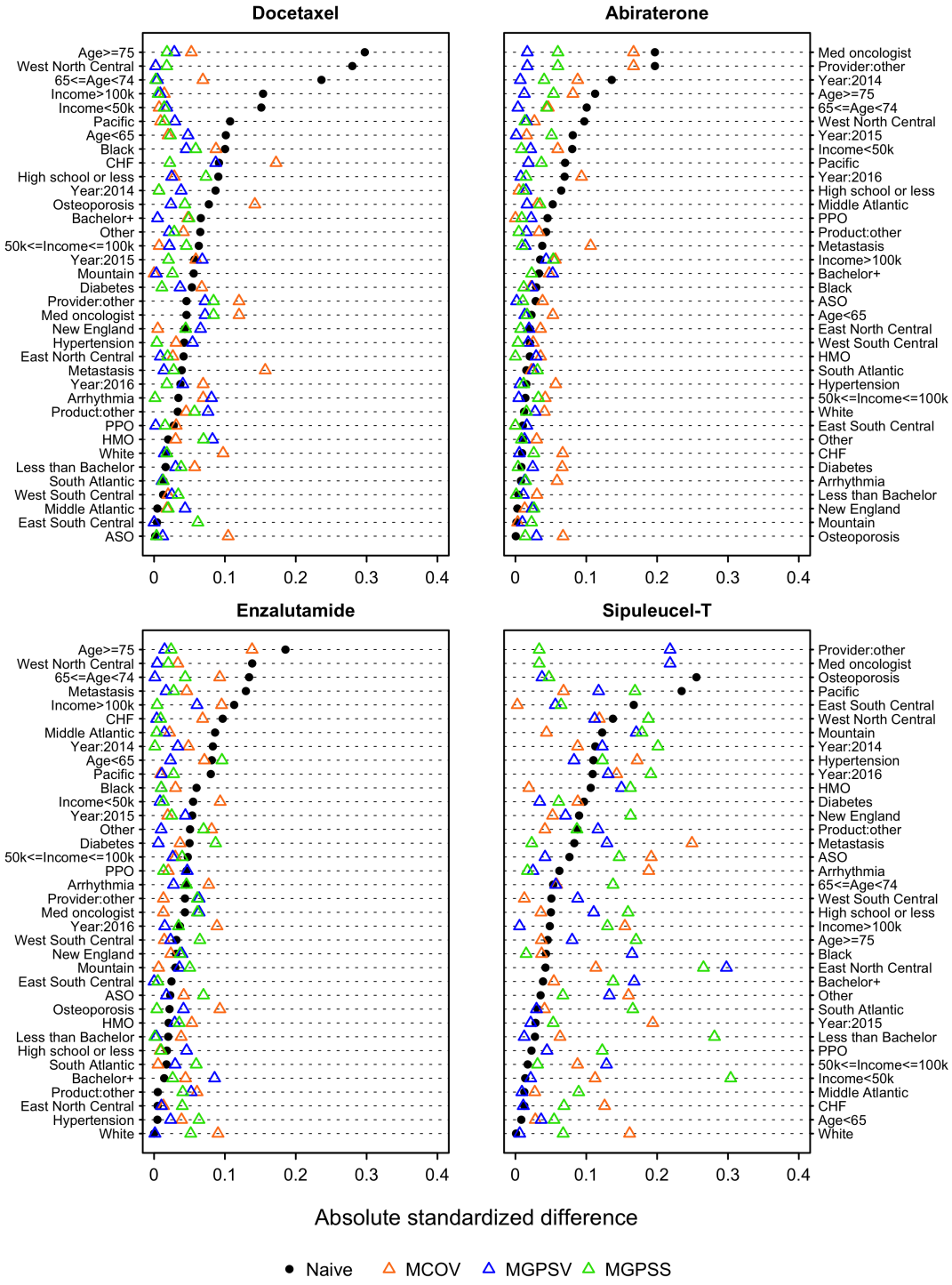

Figure A.2: Group-specific absolute standardized differences for matching-based methods. Levels of covariates were sorted by the absolute standardized difference for the naive estimator.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Covariates in propensity model ($\boldsymbol{X}_i$)	(1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})	(1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})	(1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})	(1, X_{i1}, X_{i2}, X_{i2}^2, X_{i3}, X_{i4}, X_{i5}, X_{i6}, $X_{i1}X_{i3}$)	(1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})
$\boldsymbol{\beta}_1^T$	(0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0, 0)
$\boldsymbol{\beta}_2^T$	(-0.4, 0.1, 0.1, 0.2, 0.1, 0.2, 0.3)	(-0.9, 0.4, 0.4, 0.5, 0.4, 0.4, 0.3)	(-0.9, 0.4, 0.4, 0.5, 0.4, 0.4, 0.3)	(-1.2, 0.6, 0.6, 0.5, 0.5, 0.5, 0.2, 0.4, 0.3)	(-0.9, 0.4, 0.4, 0.5, 0.4, 0.4, 0.3)
$\boldsymbol{\beta}_3^T$	(-0.2, 0.1, 0.2, 0.05, 0.1, 0.1, 0.1)	(-0.4, -0.4, -0.4, 0.5, 0.3, 0.4, -0.3)	(-0.4, -0.4, -0.4, 0.5, 0.3, 0.4, -0.3)	(-0.1, -0.3, -0.5, -0.5, 0.5, 0.5, 0.3, -0.2, 0.2)	(-0.4, -0.4, -0.4, 0.5, 0.3, 0.4, -0.3)
$\boldsymbol{\alpha}_1^T$	$(\log(0.05)$, -0.2, -0.5, -1, 1, 0.5, 0.2)	$(\log(0.05)$, -0.2, -0.5, -1, 1, 0.5, 0.2)	$(\log(0.25)$, -0.8, 0.8, 0.8, 0.8, -1, 0.8)	$(\log(0.05)$, -0.2, -0.5, -1, 1, 0.5, 0.2)	$(\log(0.02)$, -0.2, -0.5, -1, 1, 0.5, 0.2)
$\boldsymbol{\alpha}_2^T$	$(\log(0.2)$, 0.5, 0.3, -0.3, 0.8, -0.3, 0.8)	$(\log(0.2)$, 0.5, 0.3, -0.3, 0.8, -0.3, 0.8)	$(\log(0.15)$, 0.5, 0.8, -0.6, 0.8, 0.6, 0.8)	$(\log(0.2)$, 0.5, 0.3, -0.3, 0.8, -0.3, 0.8)	$(\log(0.01)$, 0.5, 0.3, -0.3, 0.8, -0.3, 0.8)
$\boldsymbol{\alpha}_3^T$	$(\log(0.3)$, -0.5, -0.5, 0.1, 0.2, -0.3, -0.3)	$(\log(0.3)$, -0.5, -0.5, 0.1, 0.2, -0.3, -0.3)	$(\log(0.15)$, 0.6, 0.7, -1, 1, 0.5, 0.6)	$(\log(0.3)$, -0.5, -0.5, 0.1, 0.2, -0.3, -0.3)	$(\log(0.04)$, -0.5, -0.5, 0.1, 0.2, -0.3, -0.3)

Table A.2: Parameter settings for simulation studies in Chapter 1

Methods	Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	44	31	61	157	140	36	252	161	410	170	98	79	66	46	28
OREG (c)	26	25	28	26	24	29	30	32	29	28	23	30	15	15	12
OREG (m)	52	26	46	79	29	66	78	41	100	90	28	75	43	15	37
PEN-GAM (c, c, GLMPS)	25	26	28	30	29	34	33	37	35	32	33	37	17	18	16
PEN-GAM (c, m, GLMPS)	25	25	28	31	29	34	33	37	35	32	33	37	17	18	16
PEN-GAM (m, c, GLMPS)	25	26	28	30	27	32	32	35	34	29	30	35	17	17	15
PEN-GAM (m, m, GLMPS)	51	27	45	92	32	77	93	44	118	29	30	34	50	18	44
IPW (c, GLMPS)	26	26	29	34	31	40	39	53	52	36	45	52	17	17	14
IPW (m, GLMPS)	52	26	46	91	33	77	88	53	117	94	37	116	50	17	45
IPW (c, CBPS)	26	26	29	32	28	34	37	40	37	37	34	45	17	17	14
IPW (m, CBPS)	52	26	46	91	30	79	85	48	111	40	28	45	50	17	45
AIPW (c, c, GLMPS)	26	25	28	29	27	32	32	36	33	34	37	41	17	16	13
AIPW (c, m, GLMPS)	26	25	29	30	29	34	33	42	39	34	37	43	17	16	13
AIPW (m, c, GLMPS)	26	25	28	29	26	32	32	34	32	49	25	49	18	16	15
AIPW (m, m, GLMPS)	52	26	46	94	30	82	91	42	112	65	25	66	51	17	46
AIPW (c, c, CBPS)	26	25	28	29	28	33	32	36	33	34	35	43	17	17	14
AIPW (c, m, CBPS)	26	25	29	30	28	34	33	38	36	37	35	45	17	17	14
AIPW (m, c, CBPS)	26	25	28	29	26	32	32	34	32	31	25	32	18	16	15
AIPW (m, m, CBPS)	52	26	46	93	30	81	91	42	112	31	25	33	51	17	46
MW (c, GLMPS)	30	26	36	50	31	46	53	58	38	50	34	45	35	18	27
MW (m, GLMPS)	52	25	53	65	29	61	55	89	124	33	29	37	36	17	39
MW (c, CBPS)	31	26	38	57	33	51	60	57	43	59	39	52	37	19	28
MW (m, CBPS)	53	25	53	64	29	61	53	91	121	46	31	44	36	18	39
AMW (c, c, GLMPS)	30	26	36	50	30	45	52	57	36	49	34	44	35	18	27
AMW (c, m, GLMPS)	30	26	36	50	30	46	52	57	37	49	34	44	35	18	27
AMW (m, c, GLMPS)	26	27	30	42	29	38	58	65	34	32	37	41	22	19	15
AMW (m, m, GLMPS)	52	25	52	66	29	63	56	88	124	32	38	45	36	17	39
AMW (c, c, CBPS)	30	26	37	53	31	50	54	58	37	52	37	48	36	19	28
AMW (c, m, CBPS)	31	26	37	55	31	50	55	57	38	55	37	50	36	19	28
AMW (m, c, CBPS)	26	27	30	43	29	41	62	70	34	46	29	45	22	19	15
AMW (m, m, CBPS)	52	25	52	65	29	61	54	93	126	43	30	41	36	17	39
OW (c, GLMPS)	27	25	31	42	28	40	42	45	36	53	29	53	30	17	23
OW (m, GLMPS)	52	26	48	72	27	68	67	68	119	35	26	40	39	16	41
OW (c, CBPS)	27	26	32	49	31	43	49	44	44	64	33	61	33	18	25
OW (m, CBPS)	52	26	48	71	28	69	64	69	114	43	28	44	39	16	41
AOW (c, c, GLMPS)	27	25	31	42	29	39	41	43	34	52	29	52	31	18	24
AOW (c, m, GLMPS)	29	26	35	46	29	43	45	45	36	51	30	50	33	18	25
AOW (m, c, GLMPS)	26	25	29	35	28	33	43	47	33	32	31	35	20	18	14
AOW (m, m, GLMPS)	51	26	47	71	28	68	66	67	116	32	32	38	39	16	41
AOW (c, c, CBPS)	27	25	31	45	29	42	43	45	35	56	32	56	33	18	25
AOW (c, m, CBPS)	30	26	37	51	30	47	47	46	38	57	33	56	34	18	26
AOW (m, c, CBPS)	26	25	29	36	28	34	46	50	33	41	27	42	20	18	14
AOW (m, m, CBPS)	51	26	47	71	28	66	64	71	118	42	28	41	38	17	40
MCOV (c)	29	28	32	37	39	37	57	61	100	40	36	38	18	18	15
MCOV (m)	53	28	52	97	41	73	111	65	162	40	35	37	49	19	42
MGPSS (c, GLMPS)	32	31	35	36	33	40	39	45	42	40	43	50	20	20	16
MGPSS (m, GLMPS)	56	31	50	95	36	84	93	49	115	38	37	44	52	20	47
MGPSS (c, CBPS)	31	30	35	37	35	40	40	46	44	40	48	53	20	21	17
MGPSS (m, CBPS)	54	31	49	97	36	84	96	50	117	36	36	42	52	21	47
MGPSV (c, GLMPS)	30	29	32	35	32	38	39	45	44	39	39	45	19	19	15
MGPSV (m, GLMPS)	53	30	49	94	35	81	95	53	127	36	35	41	51	19	45
MGPSV (c, CBPS)	30	29	33	35	33	38	40	45	46	40	42	46	19	20	15
MGPSV (m, CBPS)	53	29	49	95	36	80	96	54	129	36	34	39	51	20	45

Table A.3: Root mean squared error (RMSE) $\times$ 1000 for $n = 1500$. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	35	-19	-54	27	25	29	27	24	28	75	88	53
OREG (c)	0	-1	-1	26	25	28	25	25	28	94	95	94
OREG (m)	45	9	-36	26	25	29	26	24	29	58	93	76
PEN-GAM (c, c, GLMPS)	-1	0	1	25	26	28	26	25	27	96	95	94
PEN-GAM (c, m, GLMPS)	-1	0	0	25	25	28	26	25	27	96	94	95
PEN-GAM (m, c, GLMPS)	-1	-1	0	25	25	28	26	25	27	96	94	95
PEN-GAM (m, m, GLMPS)	44	9	-35	26	25	29	26	25	28	62	92	76
IPW (c, GLMPS)	0	-1	-1	26	26	29	26	25	29	94	94	94
IPW (m, GLMPS)	45	9	-36	26	25	29	26	25	29	58	94	76
IPW (c, CBPS)	-1	-1	0	26	26	29	26	25	29	94	94	94
IPW (m, CBPS)	45	9	-36	26	25	29	26	25	29	58	93	76
AIPW (c, c, GLMPS)	0	-1	-1	26	25	28	26	25	28	95	95	94
AIPW (c, m, GLMPS)	0	-1	-1	26	25	29	26	25	28	95	94	94
AIPW (m, c, GLMPS)	0	-1	-1	26	25	28	25	25	28	94	95	94
AIPW (m, m, GLMPS)	45	9	-36	26	25	29	26	24	29	58	93	76
AIPW (c, c, CBPS)	0	-1	-1	26	25	28	25	25	28	95	94	94
AIPW (c, m, CBPS)	-1	0	0	26	25	29	26	25	28	94	94	94
AIPW (m, c, CBPS)	0	-1	-1	26	25	28	25	25	28	94	95	94
AIPW (m, m, CBPS)	45	9	-36	26	25	29	26	24	29	58	93	76
MW (c, GLMPS)	-15	5	20	26	26	30	26	25	31	95	94	95
MW (m, GLMPS)	45	3	-43	27	25	31	27	25	30	38	94	46
MW (c, CBPS)	-16	5	21	27	26	31	27	25	31	95	94	95
MW (m, CBPS)	45	3	-43	27	25	31	27	25	30	39	94	46
AMW (c, c, GLMPS)	-15	5	20	26	25	30	26	25	30	95	95	95
AMW (c, m, GLMPS)	-15	5	20	26	25	30	26	25	30	95	95	95
AMW (m, c, GLMPS)	0	-7	-7	26	26	30	26	25	29	91	91	85
AMW (m, m, GLMPS)	45	3	-43	27	25	31	27	25	30	39	94	45
AMW (c, c, CBPS)	-16	5	21	26	25	30	26	25	30	95	95	95
AMW (c, m, CBPS)	-16	5	21	26	25	31	26	25	31	95	95	95
AMW (m, c, CBPS)	0	-7	-7	26	26	30	26	25	29	91	91	84
AMW (m, m, CBPS)	45	3	-43	27	25	31	27	25	30	39	94	45
OW (c, GLMPS)	-8	3	11	26	25	29	26	25	29	95	94	94
OW (m, GLMPS)	45	7	-38	26	25	30	26	25	29	49	94	61
OW (c, CBPS)	-9	3	12	26	25	29	26	25	29	95	94	94
OW (m, CBPS)	45	7	-38	26	25	30	26	25	29	49	94	61
AOW (c, c, GLMPS)	-8	3	11	26	25	29	26	25	29	95	95	95
AOW (c, m, GLMPS)	-14	6	20	26	25	29	26	25	29	94	94	94
AOW (m, c, GLMPS)	0	-2	-2	26	25	29	26	25	28	93	94	92
AOW (m, m, GLMPS)	44	7	-37	26	25	29	26	24	29	49	94	61
AOW (c, c, CBPS)	-8	3	11	26	25	29	26	25	29	95	95	95
AOW (c, m, CBPS)	-15	7	22	26	25	30	26	25	30	94	94	93
AOW (m, c, CBPS)	0	-2	-2	26	25	29	26	25	28	93	94	92
AOW (m, m, CBPS)	44	7	-37	26	25	29	26	24	29	49	94	61
MCOV (c)	-2	-2	1	29	28	32	29	28	32	95	95	95
MCOV (m)	44	3	-40	30	28	33	30	28	33	69	94	77
MGPSS (c, GLMPS)	0	0	0	32	31	35	32	30	33	94	94	93
MGPSS (m, GLMPS)	46	9	-36	32	29	34	32	30	34	68	94	82
MGPSS (c, CBPS)	0	-2	-1	31	30	35	32	31	35	95	95	94
MGPSS (m, CBPS)	44	8	-36	31	30	34	32	30	35	72	93	83
MGPSV (c, GLMPS)	0	-1	-1	30	29	32	31	29	33	96	95	95
MGPSV (m, GLMPS)	44	8	-36	30	29	33	31	29	33	69	94	80
MGPSV (c, CBPS)	0	-1	-1	30	29	33	31	29	33	95	95	95
MGPSV (m, CBPS)	44	8	-36	30	28	33	31	29	33	70	94	81

Table A.4: Performance of different causal inference methods in scenario 1 ($n = 1500$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	156	137	-19	62	63	69	62	60	69	28	39	94
OREG (c)	0	0	0	64	58	67	61	57	67	93	94	94
OREG (m)	74	14	-59	69	58	72	66	57	71	79	93	85
PEN-GAM (c, c, GLMPS)	3	5	2	30	28	34	32	30	35	96	96	95
PEN-GAM (c, m, GLMPS)	4	5	1	30	28	34	31	30	35	96	96	95
PEN-GAM (m, c, GLMPS)	6	3	-3	29	27	32	31	28	33	96	96	95
PEN-GAM (m, m, GLMPS)	86	17	-69	33	28	35	34	28	36	28	92	53
IPW (c, GLMPS)	0	2	2	34	31	40	32	30	37	94	94	94
IPW (m, GLMPS)	84	17	-67	35	28	39	35	28	38	32	91	55
IPW (c, CBPS)	-5	-1	4	31	28	34	30	28	33	93	95	94
IPW (m, CBPS)	84	14	-70	34	27	36	33	27	36	29	92	50
AIPW (c, c, GLMPS)	-1	0	1	29	27	32	29	27	31	94	95	94
AIPW (c, m, GLMPS)	0	1	1	30	29	34	30	28	33	94	95	95
AIPW (m, c, GLMPS)	-1	0	1	29	26	32	29	26	32	95	95	94
AIPW (m, m, GLMPS)	87	14	-73	36	27	38	34	26	36	26	92	48
AIPW (c, c, CBPS)	0	0	1	29	28	33	29	27	32	95	95	93
AIPW (c, m, CBPS)	-4	-1	4	30	28	33	30	28	32	94	95	94
AIPW (m, c, CBPS)	-1	0	1	29	26	32	29	26	31	94	95	94
AIPW (m, m, CBPS)	87	14	-72	34	27	36	33	27	35	25	92	47
MW (c, GLMPS)	-38	-10	28	33	29	37	33	29	37	95	95	94
MW (m, GLMPS)	56	6	-50	33	28	36	33	29	36	21	91	44
MW (c, CBPS)	-44	-12	32	36	30	39	35	31	39	94	96	94
MW (m, CBPS)	54	5	-49	34	29	36	34	30	37	25	92	47
AMW (c, c, GLMPS)	-38	-10	28	32	29	36	32	29	37	95	95	94
AMW (c, m, GLMPS)	-38	-10	28	33	29	36	33	29	37	95	95	94
AMW (m, c, GLMPS)	-28	-10	18	30	27	33	31	28	34	93	95	94
AMW (m, m, GLMPS)	57	6	-51	33	28	36	33	29	36	19	92	42
AMW (c, c, CBPS)	-41	-10	31	34	29	39	34	30	39	95	95	94
AMW (c, m, CBPS)	-42	-11	31	35	29	39	34	30	39	94	95	94
AMW (m, c, CBPS)	-31	-9	22	31	28	34	31	29	35	94	95	95
AMW (m, m, CBPS)	56	7	-49	33	28	36	33	29	37	21	91	46
OW (c, GLMPS)	-29	-9	20	31	27	34	31	27	35	95	95	94
OW (m, GLMPS)	65	6	-59	32	27	34	32	27	35	17	92	39
OW (c, CBPS)	-36	-12	23	33	28	36	33	29	36	94	95	94
OW (m, CBPS)	63	4	-60	32	27	35	32	28	35	21	92	40
AOW (c, c, GLMPS)	-28	-9	19	31	28	34	31	28	35	95	95	94
AOW (c, m, GLMPS)	-34	-9	24	32	28	35	32	28	36	94	95	94
AOW (m, c, GLMPS)	-19	-10	8	29	27	32	29	27	32	94	95	94
AOW (m, m, GLMPS)	64	5	-58	32	27	34	32	28	35	18	91	40
AOW (c, c, CBPS)	-31	-9	23	32	28	36	32	29	36	95	95	94
AOW (c, m, CBPS)	-38	-10	28	34	28	37	33	29	38	93	95	93
AOW (m, c, CBPS)	-21	-10	11	30	27	32	30	28	33	94	95	95
AOW (m, m, CBPS)	63	6	-56	32	27	35	32	28	35	20	92	43
MCOV (c)	18	24	6	32	30	36	32	30	36	92	88	95
MCOV (m)	91	28	-63	34	30	38	34	30	38	25	86	61
MGPSS (c, GLMPS)	0	1	1	36	33	40	36	32	38	95	95	94
MGPSS (m, GLMPS)	86	15	-71	41	33	44	40	32	42	43	93	59
MGPSS (c, CBPS)	3	4	1	37	34	40	36	34	39	95	95	94
MGPSS (m, CBPS)	88	15	-73	41	33	43	39	32	42	39	93	58
MGPSV (c, GLMPS)	2	3	1	35	32	38	35	33	38	95	95	95
MGPSV (m, GLMPS)	86	16	-70	39	31	41	38	32	40	38	93	59
MGPSV (c, CBPS)	3	5	2	35	32	37	35	33	38	96	95	95
MGPSV (m, CBPS)	87	18	-69	38	31	41	38	32	40	37	92	59

Table A.5: Performance of different causal inference methods in scenario 2 ($n = 1500$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	250	-158	-409	31	28	29	30	28	28	0	0	0
OREG (c)	-1	-1	0	30	32	29	29	32	29	94	95	94
OREG (m)	71	-24	-94	33	33	32	32	33	32	41	89	16
PEN-GAM (c, c, GLMPS)	5	-6	-12	32	36	34	33	37	37	95	95	97
PEN-GAM (c, m, GLMPS)	6	-6	-12	32	36	33	33	37	37	96	95	97
PEN-GAM (m, c, GLMPS)	7	-5	-12	31	35	32	32	35	36	95	95	96
PEN-GAM (m, m, GLMPS)	86	-26	-112	35	36	36	35	36	39	32	90	14
IPW (c, GLMPS)	0	-4	-4	39	53	52	36	48	46	93	92	93
IPW (m, GLMPS)	79	-29	-108	40	44	46	38	43	43	45	87	30
IPW (c, CBPS)	-8	0	8	36	40	36	34	40	35	92	94	92
IPW (m, CBPS)	76	-28	-104	39	39	39	37	39	37	46	88	19
AIPW (c, c, GLMPS)	-1	-1	0	32	36	33	31	35	32	95	94	94
AIPW (c, m, GLMPS)	-1	-1	-1	33	42	39	32	39	36	94	95	94
AIPW (m, c, GLMPS)	-1	-1	0	32	34	32	31	34	31	94	94	93
AIPW (m, m, GLMPS)	83	-23	-105	38	35	38	36	35	36	37	90	16
AIPW (c, c, CBPS)	-1	-1	0	32	36	33	31	35	32	95	94	94
AIPW (c, m, CBPS)	-3	4	7	33	38	35	32	36	34	94	93	92
AIPW (m, c, CBPS)	-1	-1	0	32	34	32	31	34	31	94	94	93
AIPW (m, m, CBPS)	83	-23	-105	37	35	37	35	35	36	36	90	15
MW (c, GLMPS)	-36	-42	-6	40	40	37	39	40	37	94	95	94
MW (m, GLMPS)	39	-80	-119	39	39	37	39	40	37	50	85	15
MW (c, CBPS)	-41	-35	6	44	44	42	43	44	41	94	94	93
MW (m, CBPS)	33	-81	-114	42	41	39	41	42	39	61	85	22
AMW (c, c, GLMPS)	-35	-42	-6	38	38	36	37	38	35	95	95	94
AMW (c, m, GLMPS)	-35	-42	-7	39	39	36	38	39	36	95	95	94
AMW (m, c, GLMPS)	-45	-53	-8	36	37	33	35	37	33	94	94	94
AMW (m, m, GLMPS)	40	-79	-119	38	38	36	38	38	36	48	84	13
AMW (c, c, CBPS)	-36	-42	-6	40	41	36	40	41	36	94	95	94
AMW (c, m, CBPS)	-37	-38	-1	41	43	38	40	43	37	95	94	93
AMW (m, c, CBPS)	-49	-58	-8	37	38	33	37	39	33	93	93	95
AMW (m, m, CBPS)	37	-84	-120	40	40	36	39	40	36	54	82	12
OW (c, GLMPS)	-21	-24	-3	36	37	36	35	38	36	94	95	94
OW (m, GLMPS)	56	-58	-114	37	37	36	36	38	36	44	86	14
OW (c, CBPS)	-28	-16	12	40	41	42	39	41	40	93	94	91
OW (m, CBPS)	51	-57	-108	39	39	39	38	39	38	53	87	23
AOW (c, c, GLMPS)	-20	-24	-4	35	36	34	35	36	34	94	95	94
AOW (c, m, GLMPS)	-25	-26	0	37	37	36	36	37	35	94	95	94
AOW (m, c, GLMPS)	-27	-31	-5	34	35	32	33	35	32	94	94	94
AOW (m, m, GLMPS)	55	-56	-111	36	36	35	36	36	35	45	86	14
AOW (c, c, CBPS)	-22	-25	-4	37	37	35	36	38	35	95	95	94
AOW (c, m, CBPS)	-27	-22	6	38	40	38	38	40	37	94	94	93
AOW (m, c, CBPS)	-30	-35	-5	35	36	32	34	36	33	94	94	94
AOW (m, m, CBPS)	52	-60	-112	37	37	35	37	38	35	49	84	14
MCOV (c)	44	-48	-93	35	37	37	35	37	37	76	74	29
MCOV (m)	104	-53	-157	38	38	40	38	39	39	20	72	2
MGPSS (c, GLMPS)	1	-1	-2	39	45	42	38	44	41	94	94	94
MGPSS (m, GLMPS)	82	-23	-105	44	43	46	42	43	45	52	92	36
MGPSS (c, CBPS)	7	0	-6	39	46	43	40	47	44	95	95	95
MGPSS (m, CBPS)	85	-23	-108	45	45	47	43	44	46	50	92	34
MGPSV (c, GLMPS)	6	-11	-17	39	44	41	39	46	42	95	95	95
MGPSV (m, GLMPS)	86	-34	-119	42	41	44	41	43	44	46	90	22
MGPSV (c, CBPS)	9	-12	-21	38	44	41	39	46	42	94	95	94
MGPSV (m, CBPS)	87	-34	-121	42	42	44	41	43	44	45	89	20

Table A.6: Performance of different causal inference methods in scenario 3 ($n = 1500$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	167	95	-73	28	26	31	27	27	31	0	6	34
OREG (c)	0	1	1	28	23	30	27	24	29	94	95	94
OREG (m)	85	16	-68	30	23	32	29	24	31	19	89	41
PEN-GAM (c, c, GLMPS)	1	9	8	32	31	36	33	35	39	96	97	96
PEN-GAM (c, m, GLMPS)	2	9	6	32	31	36	33	35	40	96	97	97
PEN-GAM (m, c, GLMPS)	0	8	8	29	29	34	30	33	36	96	97	96
PEN-GAM (m, m, GLMPS)	-1	7	8	29	29	34	30	32	35	96	97	96
IPW (c, GLMPS)	2	10	8	36	44	51	35	35	43	94	93	92
IPW (m, GLMPS)	-78	25	104	51	27	53	45	28	47	55	85	39
IPW (c, CBPS)	-12	2	15	35	34	43	34	32	40	91	95	92
IPW (m, CBPS)	-25	6	31	31	27	33	30	28	32	82	94	81
AIPW (c, c, GLMPS)	1	1	0	34	37	41	32	33	38	94	96	94
AIPW (c, m, GLMPS)	1	2	0	34	37	43	33	33	40	94	96	94
AIPW (m, c, GLMPS)	3	1	-1	49	25	49	37	25	38	93	95	93
AIPW (m, m, GLMPS)	-38	2	40	53	25	53	48	26	49	86	95	85
AIPW (c, c, CBPS)	1	1	1	34	35	43	33	33	40	94	95	94
AIPW (c, m, CBPS)	-12	0	12	35	35	43	34	33	41	92	95	94
AIPW (m, c, CBPS)	1	1	0	31	25	32	30	25	31	94	95	94
AIPW (m, m, CBPS)	3	-1	-4	31	25	32	29	26	31	94	95	94
MW (c, GLMPS)	-34	-16	18	37	30	41	36	31	40	94	95	94
MW (m, GLMPS)	-4	-9	-5	33	28	37	32	29	36	85	94	90
MW (c, CBPS)	-43	-17	26	41	35	45	41	36	44	94	95	93
MW (m, CBPS)	-30	-8	22	35	30	39	35	30	38	94	94	94
AMW (c, c, GLMPS)	-34	-16	17	36	30	40	35	31	39	94	95	94
AMW (c, m, GLMPS)	-33	-16	17	37	30	41	36	31	40	93	95	95
AMW (m, c, GLMPS)	-7	-26	-19	32	27	36	32	28	35	86	94	81
AMW (m, m, GLMPS)	-1	-27	-26	32	27	36	32	28	36	82	94	76
AMW (c, c, CBPS)	-36	-17	19	38	33	44	38	34	43	94	95	94
AMW (c, m, CBPS)	-39	-17	22	39	33	45	39	34	44	94	96	94
AMW (m, c, CBPS)	-32	-9	23	33	28	38	33	28	37	94	94	94
AMW (m, m, CBPS)	-27	-11	16	34	28	38	34	29	37	94	95	94
OW (c, GLMPS)	-40	-4	35	35	28	39	34	29	38	93	95	94
OW (m, GLMPS)	-16	2	18	31	26	35	31	27	34	88	94	92
OW (c, CBPS)	-51	-8	43	39	32	43	38	33	42	93	95	93
OW (m, CBPS)	-28	-3	25	33	28	36	32	28	35	93	95	93
AOW (c, c, GLMPS)	-39	-4	35	35	29	39	34	30	38	93	95	95
AOW (c, m, GLMPS)	-36	-7	29	36	29	40	35	30	39	93	95	94
AOW (m, c, GLMPS)	-9	-17	-9	31	26	34	30	27	33	82	92	73
AOW (m, m, GLMPS)	-5	-19	-14	32	26	35	31	27	35	80	92	69
AOW (c, c, CBPS)	-42	-5	37	37	32	43	37	33	42	94	96	94
AOW (c, m, CBPS)	-42	-9	33	39	32	45	39	33	43	94	96	94
AOW (m, c, CBPS)	-27	-4	22	31	26	35	31	27	35	93	95	93
AOW (m, m, CBPS)	-26	-7	19	33	27	37	33	27	36	92	95	92
MCOV (c)	20	20	0	34	29	38	34	31	38	91	92	95
MCOV (m)	22	20	-1	34	29	37	33	30	37	90	91	95
MGPSS (c, GLMPS)	1	3	1	40	42	50	39	38	44	95	95	94
MGPSS (m, GLMPS)	9	0	-8	37	37	44	36	34	41	94	94	92
MGPSS (c, CBPS)	6	14	7	39	46	52	40	44	50	95	95	94
MGPSS (m, CBPS)	-6	-4	2	36	35	42	35	35	40	94	94	94
MGPSV (c, GLMPS)	6	5	-1	39	39	45	38	38	44	94	95	94
MGPSV (m, GLMPS)	5	2	-3	36	35	41	36	35	40	94	96	94
MGPSV (c, CBPS)	11	14	4	38	39	46	39	39	45	94	95	95
MGPSV (m, CBPS)	-8	-4	3	36	33	39	35	34	38	94	95	95

Table A.7: Performance of different causal inference methods in scenario 4 ($n = 1500$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	63	42	-21	19	17	20	19	18	20	9	33	84
OREG (c)	0	0	0	15	15	12	15	15	13	95	94	97
OREG (m)	38	5	-34	19	15	17	19	15	17	46	93	49
PEN-GAM (c, c, GLMPS)	1	3	3	17	18	16	18	20	18	97	97	98
PEN-GAM (c, m, GLMPS)	1	3	2	17	18	16	18	19	18	97	97	98
PEN-GAM (m, c, GLMPS)	2	2	0	17	17	15	18	18	17	97	97	98
PEN-GAM (m, m, GLMPS)	45	7	-39	22	17	21	23	18	23	50	96	58
IPW (c, GLMPS)	1	1	0	17	17	14	17	16	14	95	95	97
IPW (m, GLMPS)	45	5	-39	23	16	22	22	16	21	49	93	50
IPW (c, CBPS)	-1	0	1	17	17	14	17	16	14	95	95	97
IPW (m, CBPS)	45	4	-40	23	16	21	22	16	20	47	93	46
AIPW (c, c, GLMPS)	1	1	0	17	16	13	16	16	13	95	95	97
AIPW (c, m, GLMPS)	1	1	0	17	16	13	16	16	14	95	95	97
AIPW (m, c, GLMPS)	0	0	0	18	16	15	17	16	15	95	95	97
AIPW (m, m, GLMPS)	45	5	-41	24	16	22	22	16	21	48	93	47
AIPW (c, c, CBPS)	1	1	0	17	17	14	16	16	14	95	95	98
AIPW (c, m, CBPS)	0	0	1	17	17	14	16	16	14	95	95	97
AIPW (m, c, CBPS)	0	0	0	18	16	15	17	16	14	94	95	97
AIPW (m, m, CBPS)	45	5	-41	23	16	21	22	16	20	46	93	46
MW (c, GLMPS)	-31	-8	23	17	17	14	17	17	14	94	95	95
MW (m, GLMPS)	29	-4	-34	21	17	19	21	17	19	19	93	15
MW (c, CBPS)	-32	-8	24	17	17	14	17	17	14	95	95	95
MW (m, CBPS)	29	-5	-34	21	17	19	22	17	20	20	94	16
AMW (c, c, GLMPS)	-31	-8	23	17	17	14	16	17	14	95	95	95
AMW (c, m, GLMPS)	-30	-8	23	17	17	14	17	17	14	95	95	95
AMW (m, c, GLMPS)	-14	-9	5	17	17	14	17	17	14	81	94	74
AMW (m, m, GLMPS)	30	-4	-34	21	17	19	21	17	19	18	93	14
AMW (c, c, CBPS)	-32	-8	24	17	17	14	17	17	14	95	95	95
AMW (c, m, CBPS)	-32	-8	24	17	17	14	17	17	15	95	95	95
AMW (m, c, CBPS)	-15	-9	6	17	17	14	17	17	14	83	94	77
AMW (m, m, CBPS)	29	-4	-33	21	17	19	21	17	20	20	94	16
OW (c, GLMPS)	-26	-7	19	16	16	13	15	16	13	94	95	95
OW (m, GLMPS)	33	-3	-36	20	16	19	20	16	19	16	93	13
OW (c, CBPS)	-28	-7	21	17	16	14	16	16	14	94	95	95
OW (m, CBPS)	33	-4	-36	21	16	19	21	16	19	18	93	14
AOW (c, c, GLMPS)	-26	-7	19	16	16	14	16	16	14	95	95	96
AOW (c, m, GLMPS)	-28	-7	21	16	16	14	16	16	14	95	95	95
AOW (m, c, GLMPS)	-10	-8	3	16	16	13	16	16	14	82	95	76
AOW (m, m, GLMPS)	33	-3	-36	21	16	19	21	16	19	18	93	14
AOW (c, c, CBPS)	-28	-7	21	17	17	14	17	17	14	95	95	96
AOW (c, m, CBPS)	-30	-7	22	17	17	14	17	17	14	94	95	95
AOW (m, c, CBPS)	-11	-8	3	17	16	14	17	16	14	84	95	79
AOW (m, m, CBPS)	32	-3	-36	21	16	19	21	16	19	19	94	16
MCOV (c)	3	6	3	17	17	15	18	18	16	95	94	97
MCOV (m)	44	7	-37	22	17	21	22	18	21	50	93	58
MGPSS (c, GLMPS)	1	1	0	20	20	16	19	19	16	95	95	97
MGPSS (m, GLMPS)	45	5	-40	26	19	24	26	19	23	58	94	59
MGPSS (c, CBPS)	2	2	0	20	21	17	20	19	16	93	94	96
MGPSS (m, CBPS)	45	5	-40	27	20	25	26	19	24	59	94	60
MGPSV (c, GLMPS)	1	1	1	19	19	15	19	19	15	96	95	97
MGPSV (m, GLMPS)	44	5	-39	25	19	23	25	19	22	57	95	57
MGPSV (c, CBPS)	1	2	1	19	20	15	19	19	16	94	94	97
MGPSV (m, CBPS)	44	5	-39	25	19	23	25	19	22	58	93	59

Table A.8: Performance of different causal inference methods in scenario 5 ($n = 1500$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	104	96	99	85	91	83	83	57	62	79	76	72	117	107	144
OREG (c)	98	97	98	82	82	77	80	67	63	77	67	67	91	90	90
OREG (m)	99	96	101	90	83	83	89	69	69	84	67	73	114	90	122
PEN-GAM (c, c, GLMPS)	101	100	96	98	100	93	90	77	82	95	100	92	111	122	129
PEN-GAM (c, m, GLMPS)	101	100	96	98	100	93	90	77	82	97	100	93	110	119	126
PEN-GAM (m, c, GLMPS)	100	99	96	96	94	89	89	73	78	87	93	84	111	112	123
PEN-GAM (m, m, GLMPS)	101	97	99	105	96	98	98	75	85	87	92	84	140	113	163
IPW (c, GLMPS)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
IPW (m, GLMPS)	100	97	102	107	94	101	104	91	95	128	78	111	134	97	147
IPW (c, CBPS)	99	99	100	93	94	88	95	83	76	97	92	94	100	100	97
IPW (m, CBPS)	100	96	101	103	91	95	102	81	81	85	79	75	132	98	143
AIPW (c, c, GLMPS)	98	98	98	88	91	84	84	72	69	91	87	84	97	98	96
AIPW (c, m, GLMPS)	99	99	99	92	94	89	88	81	78	95	89	89	100	98	98
AIPW (m, c, GLMPS)	98	97	98	90	88	85	85	70	69	106	71	89	103	96	105
AIPW (m, m, GLMPS)	99	96	101	106	89	98	99	73	79	138	73	115	135	96	147
AIPW (c, c, CBPS)	98	98	98	90	92	85	86	73	69	96	94	94	99	100	98
AIPW (c, m, CBPS)	98	98	99	91	93	87	89	76	73	99	95	96	99	100	98
AIPW (m, c, CBPS)	98	97	98	89	89	84	85	70	68	85	72	73	102	97	103
AIPW (m, m, CBPS)	99	96	101	102	89	95	97	73	78	85	73	73	132	97	144
MW (c, GLMPS)	102	100	107	102	99	99	107	84	81	105	89	94	100	102	100
MW (m, GLMPS)	103	98	106	103	98	98	107	83	80	93	82	86	128	102	138
MW (c, CBPS)	103	100	108	109	105	106	119	93	90	117	101	104	104	107	104
MW (m, CBPS)	103	98	106	106	101	100	114	87	85	100	86	89	131	105	141
AMW (c, c, GLMPS)	101	98	105	100	98	98	103	80	77	102	88	92	99	102	100
AMW (c, m, GLMPS)	101	99	106	101	98	99	105	81	78	103	88	93	100	102	100
AMW (m, c, GLMPS)	101	99	103	94	95	91	98	78	73	91	79	83	101	101	100
AMW (m, m, GLMPS)	102	97	105	102	97	97	106	80	78	92	79	84	128	102	138
AMW (c, c, CBPS)	101	99	106	104	101	104	109	85	78	109	96	101	102	104	104
AMW (c, m, CBPS)	102	99	107	106	102	105	111	89	82	112	97	103	103	104	104
AMW (m, c, CBPS)	101	99	103	96	97	94	102	81	73	96	81	88	102	103	103
AMW (m, m, CBPS)	102	97	105	103	99	100	109	83	79	97	81	88	130	103	141
OW (c, GLMPS)	99	98	101	95	92	93	97	79	78	98	82	89	94	95	95
OW (m, GLMPS)	100	97	102	97	92	93	99	79	79	88	77	81	122	96	134
OW (c, CBPS)	100	98	102	101	98	98	107	86	87	109	94	98	97	99	98
OW (m, CBPS)	100	96	102	100	95	95	106	82	83	93	80	82	125	98	136
AOW (c, c, GLMPS)	99	97	100	96	94	93	96	76	74	98	85	89	98	99	99
AOW (c, m, GLMPS)	100	97	102	99	95	96	100	77	77	102	85	92	98	99	99
AOW (m, c, GLMPS)	99	98	99	91	92	87	91	74	71	88	76	79	98	99	98
AOW (m, m, GLMPS)	100	96	101	99	93	93	100	76	76	91	77	82	125	99	136
AOW (c, c, CBPS)	99	97	101	99	97	98	100	79	76	106	93	98	100	102	103
AOW (c, m, CBPS)	101	97	104	103	98	101	105	83	80	111	94	102	100	102	102
AOW (m, c, CBPS)	99	98	99	93	93	89	94	76	71	91	77	81	100	100	101
AOW (m, m, CBPS)	100	96	101	100	95	95	102	78	77	95	78	84	127	100	138
MCOV (c)	111	109	111	99	102	96	97	78	80	98	87	89	107	109	114
MCOV (m)	115	109	114	106	102	101	104	81	85	96	86	87	136	108	150
MGPSS (c, GLMPS)	123	119	116	110	109	102	105	91	90	111	108	104	116	116	111
MGPSS (m, GLMPS)	122	116	119	123	108	112	117	90	99	105	97	96	155	115	168
MGPSS (c, CBPS)	124	120	121	111	114	106	109	98	96	114	124	118	118	119	116
MGPSS (m, CBPS)	123	117	121	121	108	112	118	93	99	100	98	94	155	115	168
MGPSV (c, GLMPS)	118	116	114	108	110	101	107	95	92	111	108	103	115	116	110
MGPSV (m, GLMPS)	118	113	116	118	108	108	114	91	97	105	99	93	149	114	160
MGPSV (c, CBPS)	118	116	114	108	110	101	107	95	93	111	110	105	115	115	111
MGPSV (m, CBPS)	118	113	116	117	108	108	113	91	97	101	96	90	149	114	161

Table A.9: 100 $\times$ Ratio of 95% confidence interval width to 95% confidence interval width of GLMPS based IPW(c) for $n = 1500$. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	69	58	83	168	151	72	261	171	416	179	113	101	75	57	48
OREG (c)	57	57	64	64	58	67	68	75	70	62	55	67	35	35	28
OREG (m)	73	56	74	101	60	94	105	81	123	110	57	99	57	35	52
PEN-GAM (c, c, GLMPS)	57	53	62	65	64	76	75	90	99	68	65	80	38	39	36
PEN-GAM (c, m, GLMPS)	57	53	62	65	64	76	75	88	97	68	65	80	38	39	36
PEN-GAM (m, c, GLMPS)	57	53	61	64	62	74	75	87	98	64	66	78	38	39	37
PEN-GAM (m, m, GLMPS)	65	52	70	102	67	91	119	95	163	64	67	79	63	40	58
IPW (c, GLMPS)	58	58	65	76	70	88	86	113	110	82	78	96	39	39	32
IPW (m, GLMPS)	73	57	75	120	69	113	122	108	156	120	70	141	72	38	66
IPW (c, CBPS)	58	58	65	75	70	82	89	98	86	89	79	104	40	39	32
IPW (m, CBPS)	73	56	75	116	67	112	115	98	138	79	68	87	68	38	63
AIPW (c, c, GLMPS)	58	57	64	73	64	79	76	83	81	76	70	89	38	38	31
AIPW (c, m, GLMPS)	57	57	64	75	65	83	79	94	92	80	74	94	40	38	33
AIPW (m, c, GLMPS)	57	57	64	72	62	77	75	81	80	113	59	115	41	37	35
AIPW (m, m, GLMPS)	73	56	74	127	64	121	127	86	145	169	63	171	194	37	192
AIPW (c, c, CBPS)	58	57	64	74	70	80	81	88	82	92	86	108	41	40	34
AIPW (c, m, CBPS)	57	57	64	77	71	83	84	92	88	102	88	116	41	40	34
AIPW (m, c, CBPS)	57	57	64	71	65	77	75	82	78	74	63	78	42	38	35
AIPW (m, m, CBPS)	73	56	74	119	66	113	120	88	139	75	64	80	69	38	64
MW (c, GLMPS)	60	58	70	86	71	91	95	99	86	90	74	95	49	39	39
MW (m, GLMPS)	75	56	79	96	69	99	99	118	147	74	66	84	56	39	56
MW (c, CBPS)	62	58	73	101	78	101	116	105	99	113	89	113	53	42	41
MW (m, CBPS)	75	56	79	96	72	102	103	121	137	90	73	95	57	41	57
AMW (c, c, GLMPS)	60	57	70	85	71	90	92	95	83	88	73	93	50	40	39
AMW (c, m, GLMPS)	59	57	69	85	71	90	93	96	83	89	72	94	49	39	39
AMW (m, c, GLMPS)	59	57	66	78	69	82	92	101	79	73	67	84	42	40	32
AMW (m, m, GLMPS)	75	56	78	96	68	99	99	115	146	73	68	85	56	39	56
AMW (c, c, CBPS)	60	57	71	92	74	99	101	104	86	100	84	107	52	41	42
AMW (c, m, CBPS)	61	57	72	94	75	98	103	104	88	103	85	109	52	41	41
AMW (m, c, CBPS)	59	57	66	81	71	87	100	109	80	85	68	92	43	42	34
AMW (m, m, CBPS)	75	56	78	97	71	101	100	124	150	84	69	90	57	40	57
OW (c, GLMPS)	58	57	67	78	66	84	85	90	83	88	68	95	45	37	36
OW (m, GLMPS)	74	56	76	99	65	101	102	103	143	73	62	83	57	36	56
OW (c, CBPS)	59	57	69	95	74	95	106	97	100	113	84	114	50	39	39
OW (m, CBPS)	74	56	76	97	68	104	102	105	130	86	68	91	57	38	58
AOW (c, c, GLMPS)	58	57	66	79	68	84	84	87	80	87	69	95	47	39	37
AOW (c, m, GLMPS)	58	57	68	82	69	87	87	87	82	89	70	96	47	38	38
AOW (m, c, GLMPS)	58	57	64	73	66	77	82	89	77	71	63	79	40	39	32
AOW (m, m, GLMPS)	73	56	75	99	66	100	101	100	140	72	64	82	57	38	56
AOW (c, c, CBPS)	58	57	67	87	72	93	92	94	83	99	81	108	50	40	40
AOW (c, m, CBPS)	61	57	71	92	72	95	96	95	87	103	82	111	50	40	40
AOW (m, c, CBPS)	58	57	64	76	69	82	89	96	78	80	65	88	41	40	33
AOW (m, m, CBPS)	73	56	75	99	68	101	102	109	144	82	66	89	57	39	57
MCOV (c)	66	64	73	83	84	82	116	114	181	86	76	84	40	41	36
MCOV (m)	81	64	87	127	84	102	155	119	230	87	76	83	65	41	58
MGPSS (c, GLMPS)	71	71	80	85	80	90	93	111	109	89	90	107	46	46	37
MGPSS (m, GLMPS)	85	71	88	128	79	121	134	111	159	81	80	95	74	44	68
MGPSS (c, CBPS)	71	68	78	86	82	96	97	113	114	93	96	111	45	44	40
MGPSS (m, CBPS)	83	69	86	129	80	124	141	111	166	81	83	95	75	44	70
MGPSV (c, GLMPS)	65	66	75	80	75	86	90	103	106	88	82	96	44	44	35
MGPSV (m, GLMPS)	80	65	84	122	76	112	132	108	174	82	77	92	71	43	64
MGPSV (c, CBPS)	66	65	75	83	76	88	95	102	116	90	87	97	43	43	36
MGPSV (m, CBPS)	80	65	84	124	78	112	136	111	180	79	78	90	71	43	64

Table A.10: Root mean squared error (RMSE) $\times$ 1000 for $n = 300$. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	34	-20	-54	60	54	62	60	54	63	91	93	85
OREG (c)	-1	-1	0	57	57	64	58	57	64	95	94	94
OREG (m)	44	8	-36	58	55	65	58	55	65	87	94	91
PEN-GAM (c, c, GLMPS)	-11	-5	6	56	52	62	63	63	66	97	98	97
PEN-GAM (c, m, GLMPS)	-10	-5	6	56	53	62	63	62	66	97	98	96
PEN-GAM (m, c, GLMPS)	-13	-7	5	55	52	61	64	63	67	97	98	97
PEN-GAM (m, m, GLMPS)	31	1	-29	57	52	63	64	62	68	95	98	95
IPW (c, GLMPS)	0	0	-1	58	58	65	59	59	66	95	95	95
IPW (m, GLMPS)	45	8	-37	58	56	66	59	56	66	88	94	91
IPW (c, CBPS)	-1	0	2	58	58	65	58	57	65	95	94	94
IPW (m, CBPS)	45	8	-37	58	56	65	59	56	66	87	94	91
AIPW (c, c, GLMPS)	0	-1	-1	58	57	64	58	58	64	95	95	94
AIPW (c, m, GLMPS)	1	0	-1	57	57	64	59	59	65	95	95	94
AIPW (m, c, GLMPS)	0	-1	0	57	57	64	58	57	64	95	94	94
AIPW (m, m, GLMPS)	44	8	-36	58	55	65	59	56	65	87	94	91
AIPW (c, c, CBPS)	0	-1	-1	58	57	64	58	57	64	95	94	94
AIPW (c, m, CBPS)	-1	1	2	57	57	64	58	57	65	95	94	94
AIPW (m, c, CBPS)	0	-1	0	57	57	64	58	57	64	94	94	94
AIPW (m, m, CBPS)	44	8	-36	58	55	65	59	56	65	87	94	91
MW (c, GLMPS)	-15	4	19	58	57	67	61	59	70	96	95	95
MW (m, GLMPS)	45	4	-42	60	56	67	61	57	69	83	94	85
MW (c, CBPS)	-18	6	24	59	58	69	63	60	72	96	95	95
MW (m, CBPS)	46	4	-41	60	56	67	62	58	70	83	95	85
AMW (c, c, GLMPS)	-14	5	19	58	57	67	60	58	69	95	95	95
AMW (c, m, GLMPS)	-14	5	18	58	57	67	61	58	69	96	95	95
AMW (m, c, GLMPS)	0	-5	-5	59	57	66	60	58	67	95	94	93
AMW (m, m, GLMPS)	45	4	-41	60	56	67	61	57	68	83	95	85
AMW (c, c, CBPS)	-16	6	22	58	57	68	61	58	70	96	95	95
AMW (c, m, CBPS)	-18	5	23	59	57	68	62	59	70	96	95	95
AMW (m, c, CBPS)	0	-6	-5	59	57	66	60	58	67	95	94	93
AMW (m, m, CBPS)	45	4	-41	60	55	67	61	57	69	83	95	85
OW (c, GLMPS)	-9	3	12	57	57	65	59	57	66	95	94	95
OW (m, GLMPS)	45	6	-39	58	56	65	59	56	66	85	94	88
OW (c, CBPS)	-13	4	17	58	57	67	60	57	68	95	94	95
OW (m, CBPS)	45	7	-39	58	55	66	59	56	66	85	95	88
AOW (c, c, GLMPS)	-9	3	12	57	56	65	59	57	66	95	95	94
AOW (c, m, GLMPS)	-13	6	19	57	56	66	60	57	67	95	95	95
AOW (m, c, GLMPS)	-1	-3	-2	58	57	64	59	57	65	95	95	94
AOW (m, m, GLMPS)	44	6	-38	59	55	65	60	56	66	86	95	88
AOW (c, c, CBPS)	-10	4	14	57	56	66	59	57	67	95	95	94
AOW (c, m, CBPS)	-18	6	24	58	57	67	61	58	69	95	95	95
AOW (m, c, CBPS)	-1	-3	-2	58	57	64	59	57	65	95	95	94
AOW (m, m, CBPS)	44	6	-38	59	55	65	60	56	66	86	95	88
MCOV (c)	-4	-5	-1	66	63	73	66	63	72	95	94	94
MCOV (m)	42	-4	-46	69	64	74	68	63	74	90	95	89
MGPSS (c, GLMPS)	2	0	-1	71	71	80	69	66	73	94	92	92
MGPSS (m, GLMPS)	45	7	-37	72	70	80	70	66	75	88	93	91
MGPSS (c, CBPS)	0	-1	-1	71	68	78	73	71	79	95	95	95
MGPSS (m, CBPS)	44	8	-36	71	69	78	72	69	80	90	94	92
MGPSV (c, GLMPS)	0	-1	-1	65	66	75	69	66	75	96	95	94
MGPSV (m, GLMPS)	44	5	-38	67	65	75	69	65	76	91	95	92
MGPSV (c, CBPS)	0	-3	-2	66	65	74	69	66	75	96	95	95
MGPSV (m, CBPS)	44	5	-39	67	65	75	69	65	76	90	95	92

Table A.11: Performance of different causal inference methods in scenario 1 ($n = 300$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	156	137	-19	62	63	69	62	60	69	28	39	94
OREG (c)	0	0	0	64	58	67	61	57	67	93	94	94
OREG (m)	74	14	-59	69	58	72	66	57	71	79	93	85
PEN-GAM (c, c, GLMPS)	1	18	16	65	62	74	78	80	91	98	99	98
PEN-GAM (c, m, GLMPS)	3	15	12	65	62	75	78	79	90	97	98	98
PEN-GAM (m, c, GLMPS)	6	13	7	64	61	73	78	76	88	98	99	97
PEN-GAM (m, m, GLMPS)	75	25	-50	68	62	76	79	76	89	88	98	94
IPW (c, GLMPS)	6	7	1	76	70	88	73	67	82	93	94	93
IPW (m, GLMPS)	87	18	-68	82	67	90	77	64	84	76	93	83
IPW (c, CBPS)	-11	-4	7	74	70	82	73	68	80	93	94	94
IPW (m, CBPS)	83	12	-71	81	66	86	77	65	82	79	94	83
AIPW (c, c, GLMPS)	2	0	-2	73	64	79	68	66	77	93	95	95
AIPW (c, m, GLMPS)	4	1	-3	75	65	83	76	68	85	94	95	95
AIPW (m, c, GLMPS)	0	1	0	72	62	77	69	62	75	93	95	94
AIPW (m, m, GLMPS)	88	14	-75	91	62	95	80	61	86	80	94	84
AIPW (c, c, CBPS)	2	0	-1	74	70	80	76	72	82	95	96	96
AIPW (c, m, CBPS)	-12	-3	9	76	71	82	79	74	86	94	96	96
AIPW (m, c, CBPS)	0	1	1	71	65	77	70	65	76	93	95	94
AIPW (m, m, CBPS)	87	14	-73	81	65	86	78	65	83	79	95	83
MW (c, GLMPS)	-37	-10	27	78	70	87	77	69	86	94	94	95
MW (m, GLMPS)	57	5	-52	77	68	84	77	68	84	76	94	83
MW (c, CBPS)	-54	-17	37	86	76	95	90	80	97	95	96	95
MW (m, CBPS)	52	0	-52	81	72	88	85	74	92	81	96	85
AMW (c, c, GLMPS)	-37	-10	27	77	70	86	76	69	85	93	94	94
AMW (c, m, GLMPS)	-35	-10	26	77	70	86	76	69	85	94	94	94
AMW (m, c, GLMPS)	-27	-10	18	73	68	80	71	67	80	93	94	94
AMW (m, m, GLMPS)	58	5	-53	77	68	84	76	67	84	75	94	82
AMW (c, c, CBPS)	-44	-10	34	81	73	93	84	76	94	95	96	95
AMW (c, m, CBPS)	-47	-13	33	82	74	92	85	77	95	95	96	95
AMW (m, c, CBPS)	-31	-9	22	75	71	84	76	72	86	95	95	95
AMW (m, m, CBPS)	56	6	-51	79	71	87	81	72	89	77	95	84
OW (c, GLMPS)	-28	-8	20	73	66	82	71	64	80	94	94	94
OW (m, GLMPS)	66	6	-60	74	65	81	72	63	80	74	94	81
OW (c, CBPS)	-49	-19	30	81	71	90	83	74	91	94	95	94
OW (m, CBPS)	59	-2	-61	78	68	85	79	69	86	79	95	84
AOW (c, c, GLMPS)	-28	-9	20	74	68	82	73	66	81	94	95	95
AOW (c, m, GLMPS)	-31	-8	23	76	68	84	75	67	83	94	94	95
AOW (m, c, GLMPS)	-18	-9	9	70	66	77	69	65	77	93	94	95
AOW (m, m, GLMPS)	64	5	-59	75	66	81	74	65	81	75	94	83
AOW (c, c, CBPS)	-36	-9	27	79	71	89	81	74	91	95	96	95
AOW (c, m, CBPS)	-44	-12	31	81	71	90	83	75	92	94	96	95
AOW (m, c, CBPS)	-22	-9	13	73	68	81	74	69	82	94	95	95
AOW (m, m, CBPS)	62	5	-57	77	68	84	79	69	86	78	95	85
MCOV (c)	37	44	7	74	71	81	73	69	82	92	90	95
MCOV (m)	100	43	-56	79	72	85	76	70	84	71	89	88
MGPSS (c, GLMPS)	5	3	-2	85	80	90	77	69	83	92	92	93
MGPSS (m, GLMPS)	89	17	-72	92	77	98	84	71	90	79	93	83
MGPSS (c, CBPS)	14	9	-4	85	82	95	87	84	100	95	95	96
MGPSS (m, CBPS)	91	18	-73	91	78	101	90	78	99	81	94	86
MGPSV (c, GLMPS)	9	9	-1	79	75	86	80	73	87	94	94	95
MGPSV (m, GLMPS)	87	22	-66	85	73	91	84	73	90	80	94	87
MGPSV (c, CBPS)	15	15	0	82	75	88	79	74	87	93	94	95
MGPSV (m, CBPS)	89	24	-65	86	74	91	84	73	90	80	94	88

Table A.12: Performance of different causal inference methods in scenario 2 ($n = 300$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	251	-159	-410	67	61	63	67	61	63	4	27	0
OREG (c)	1	-5	-6	67	74	68	67	78	71	95	96	95
OREG (m)	73	-24	-98	73	76	73	74	78	75	83	95	78
PEN-GAM (c, c, GLMPS)	23	-41	-64	71	80	76	83	97	101	97	96	97
PEN-GAM (c, m, GLMPS)	23	-38	-61	71	80	75	82	96	100	97	97	97
PEN-GAM (m, c, GLMPS)	26	-39	-65	70	78	73	81	92	97	96	96	97
PEN-GAM (m, m, GLMPS)	92	-51	-143	76	80	78	84	91	97	84	94	75
IPW (c, GLMPS)	6	-13	-19	84	110	106	85	104	101	95	93	93
IPW (m, GLMPS)	84	-31	-115	88	100	103	86	98	98	83	92	75
IPW (c, CBPS)	-17	-5	12	83	94	86	84	95	84	94	94	93
IPW (m, CBPS)	74	-30	-105	85	89	88	86	90	86	87	93	79
AIPW (c, c, GLMPS)	1	-3	-4	71	81	77	74	89	83	95	97	95
AIPW (c, m, GLMPS)	3	-6	-9	74	90	86	84	107	100	96	96	96
AIPW (m, c, GLMPS)	1	-4	-6	72	79	76	75	83	81	95	96	95
AIPW (m, m, GLMPS)	89	-23	-112	210	81	213	96	84	101	85	95	81
AIPW (c, c, CBPS)	2	-3	-5	77	86	80	82	92	84	96	97	95
AIPW (c, m, CBPS)	-10	5	15	79	89	87	87	96	91	96	96	95
AIPW (m, c, CBPS)	2	-4	-6	73	80	76	75	84	79	95	96	95
AIPW (m, m, CBPS)	85	-23	-108	83	82	85	84	85	86	84	95	79
MW (c, GLMPS)	-32	-37	-5	90	92	85	91	94	87	95	94	95
MW (m, GLMPS)	43	-76	-119	89	90	83	90	91	86	86	93	75
MW (c, CBPS)	-46	-20	26	102	103	97	109	111	102	96	96	94
MW (m, CBPS)	30	-72	-103	96	97	90	102	102	95	91	94	84
AMW (c, c, GLMPS)	-31	-37	-6	88	89	80	87	90	83	95	94	95
AMW (c, m, GLMPS)	-30	-38	-9	88	89	81	89	91	84	96	94	95
AMW (m, c, GLMPS)	-42	-53	-11	83	86	76	83	88	79	95	95	95
AMW (m, m, GLMPS)	44	-75	-119	87	87	81	89	89	83	85	94	73
AMW (c, c, CBPS)	-33	-38	-5	95	96	83	99	101	89	96	96	96
AMW (c, m, CBPS)	-34	-28	6	94	99	86	101	105	91	97	96	96
AMW (m, c, CBPS)	-47	-59	-12	88	90	78	90	94	82	95	95	96
AMW (m, m, CBPS)	41	-82	-123	91	92	83	96	95	86	88	94	74
OW (c, GLMPS)	-20	-23	-4	82	87	82	83	88	83	95	94	94
OW (m, GLMPS)	58	-57	-115	82	86	82	83	86	83	85	93	74
OW (c, CBPS)	-40	-3	37	94	97	95	99	103	96	95	95	93
OW (m, CBPS)	43	-52	-95	90	91	88	94	95	91	90	94	84
AOW (c, c, GLMPS)	-18	-24	-5	82	84	78	82	86	80	95	94	95
AOW (c, m, GLMPS)	-21	-25	-4	84	84	80	85	86	82	95	95	95
AOW (m, c, GLMPS)	-25	-34	-9	78	82	75	78	84	78	95	95	95
AOW (m, m, GLMPS)	57	-56	-113	82	83	79	84	85	82	85	93	74
AOW (c, c, CBPS)	-22	-27	-5	88	90	81	93	95	86	96	96	96
AOW (c, m, CBPS)	-27	-18	10	89	93	85	97	100	90	97	96	96
AOW (m, c, CBPS)	-31	-41	-10	81	86	76	84	89	80	95	95	95
AOW (m, m, CBPS)	53	-63	-117	85	87	81	91	91	85	88	94	75
MCOV (c)	83	-78	-161	80	82	81	80	79	81	81	82	50
MCOV (m)	131	-81	-212	85	84	86	84	82	84	65	82	30
MGPSS (c, GLMPS)	8	-9	-17	91	109	104	84	97	95	92	90	89
MGPSS (m, GLMPS)	88	-31	-119	97	103	105	91	97	101	81	91	80
MGPSS (c, CBPS)	30	-8	-38	95	112	110	98	120	120	95	97	96
MGPSS (m, CBPS)	101	-25	-126	99	106	108	100	110	114	85	95	82
MGPSV (c, GLMPS)	22	-25	-47	87	100	94	90	101	96	94	93	92
MGPSV (m, GLMPS)	96	-47	-143	91	95	96	93	96	98	83	92	71
MGPSV (c, CBPS)	36	-29	-65	88	98	97	89	100	97	93	94	90
MGPSV (m, CBPS)	102	-49	-150	92	92	96	92	96	98	80	92	68

Table A.13: Performance of different causal inference methods in scenario 3 ($n = 300$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	167	93	-74	61	61	70	61	60	68	23	64	80
OREG (c)	-2	1	3	61	55	66	62	56	67	95	95	95
OREG (m)	83	16	-67	66	55	71	67	55	71	77	94	84
PEN-GAM (c, c, GLMPS)	-1	17	18	68	63	78	82	81	95	98	99	98
PEN-GAM (c, m, GLMPS)	0	16	15	68	63	78	83	80	94	98	99	98
PEN-GAM (m, c, GLMPS)	-3	20	24	64	63	75	76	80	90	97	99	98
PEN-GAM (m, m, GLMPS)	-3	19	22	64	64	76	76	80	90	97	98	98
IPW (c, GLMPS)	3	19	16	81	77	99	82	74	94	95	94	94
IPW (m, GLMPS)	-65	27	93	104	67	109	94	64	99	86	92	81
IPW (c, CBPS)	-25	6	31	84	80	100	83	75	93	92	94	91
IPW (m, CBPS)	-30	2	32	72	66	80	73	68	80	92	96	92
AIPW (c, c, GLMPS)	-2	1	3	73	73	89	88	92	102	96	97	96
AIPW (c, m, GLMPS)	0	3	3	79	77	96	101	96	119	97	97	97
AIPW (m, c, GLMPS)	1	3	2	116	65	120	115	65	120	97	96	96
AIPW (m, m, GLMPS)	-45	5	49	161	77	159	205	72	208	99	96	97
AIPW (c, c, CBPS)	-1	2	3	91	89	109	91	83	103	95	95	95
AIPW (c, m, CBPS)	-34	-1	32	94	90	113	96	85	109	93	95	94
AIPW (m, c, CBPS)	1	3	2	72	63	78	77	67	80	96	97	96
AIPW (m, m, CBPS)	-8	0	8	73	64	79	80	69	84	97	97	96
MW (c, GLMPS)	-39	-16	23	83	72	93	87	75	94	95	96	95
MW (m, GLMPS)	-9	-9	0	73	66	85	76	68	85	94	95	94
MW (c, CBPS)	-61	-20	41	97	86	106	107	95	113	96	97	96
MW (m, CBPS)	-42	-14	28	81	71	90	88	77	94	97	96	96
AMW (c, c, GLMPS)	-37	-16	22	81	71	92	84	73	92	95	96	95
AMW (c, m, GLMPS)	-37	-15	22	83	71	93	85	73	93	95	96	95
AMW (m, c, GLMPS)	-10	-24	-14	73	64	83	75	66	83	94	95	93
AMW (m, m, GLMPS)	-4	-25	-21	73	64	83	75	66	83	93	95	92
AMW (c, c, CBPS)	-45	-17	28	91	82	106	98	88	109	96	96	95
AMW (c, m, CBPS)	-49	-18	31	92	83	107	100	89	111	96	96	96
AMW (m, c, CBPS)	-36	-11	24	77	66	89	83	72	92	96	96	95
AMW (m, m, CBPS)	-33	-14	18	78	67	88	84	73	91	96	96	96
OW (c, GLMPS)	-43	-6	38	78	67	89	80	68	88	95	95	95
OW (m, GLMPS)	-19	1	20	70	62	80	71	63	80	95	95	95
OW (c, CBPS)	-70	-17	53	92	81	103	99	88	106	95	96	95
OW (m, CBPS)	-44	-11	33	76	66	85	81	71	88	96	96	96
AOW (c, c, GLMPS)	-42	-6	36	79	69	89	81	71	89	95	95	95
AOW (c, m, GLMPS)	-39	-7	32	82	69	91	84	71	92	95	95	95
AOW (m, c, GLMPS)	-12	-17	-5	70	62	79	72	63	79	94	95	92
AOW (m, m, GLMPS)	-8	-18	-10	72	62	81	74	64	81	93	95	91
AOW (c, c, CBPS)	-49	-9	40	89	80	103	95	86	106	96	96	95
AOW (c, m, CBPS)	-51	-12	39	92	82	106	99	87	109	96	96	95
AOW (m, c, CBPS)	-32	-7	25	74	64	84	80	69	87	96	96	95
AOW (m, m, CBPS)	-32	-10	22	76	65	86	82	70	89	96	96	95
MCOV (c)	39	31	-7	75	68	83	76	68	84	92	92	95
MCOV (m)	41	33	-8	73	69	83	76	68	83	92	92	94
MGPSS (c, GLMPS)	4	8	4	91	92	109	80	72	88	91	90	90
MGPSS (m, GLMPS)	8	4	-4	82	84	96	77	72	87	93	92	92
MGPSS (c, CBPS)	15	18	3	93	97	116	100	100	122	96	96	96
MGPSS (m, CBPS)	3	3	0	80	83	97	85	85	100	96	96	96
MGPSV (c, GLMPS)	16	12	-5	85	82	97	86	80	96	94	94	94
MGPSV (m, GLMPS)	10	5	-4	79	77	88	82	75	90	95	94	95
MGPSV (c, CBPS)	24	27	2	83	83	97	86	81	97	94	94	95
MGPSV (m, CBPS)	4	5	0	77	75	86	80	74	88	95	95	95

Table A.14: Performance of different causal inference methods in scenario 4 ($n = 300$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Bias from ATE × 1000			Empirical SD × 1000			Average SE × 1000			95% Coverage Rate (%)		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	62	41	-21	43	39	44	43	39	45	71	81	94
OREG (c)	-4	-3	1	35	35	28	38	39	32	97	97	98
OREG (m)	37	2	-34	43	35	39	45	37	41	90	96	90
PEN-GAM (c, c, GLMPS)	-1	2	4	37	39	36	55	58	59	100	100	100
PEN-GAM (c, m, GLMPS)	-1	5	7	38	38	35	54	58	58	100	100	100
PEN-GAM (m, c, GLMPS)	-3	2	5	38	39	37	53	57	57	99	100	100
PEN-GAM (m, m, GLMPS)	42	7	-34	48	40	47	60	57	64	96	100	98
IPW (c, GLMPS)	0	1	1	39	39	32	37	37	31	95	94	96
IPW (m, GLMPS)	44	5	-39	57	38	53	48	36	44	87	94	89
IPW (c, CBPS)	-4	-2	2	40	39	32	39	39	31	96	95	97
IPW (m, CBPS)	43	3	-40	53	38	49	49	37	44	87	95	89
AIPW (c, c, GLMPS)	-2	-2	1	38	38	31	40	42	34	97	97	98
AIPW (c, m, GLMPS)	0	-1	-1	40	38	33	46	42	38	97	97	98
AIPW (m, c, GLMPS)	-4	-2	2	41	37	35	43	40	37	96	97	98
AIPW (m, m, GLMPS)	47	3	-44	188	37	187	61	38	57	90	96	91
AIPW (c, c, CBPS)	-2	-1	0	41	40	34	43	44	35	97	98	99
AIPW (c, m, CBPS)	-5	-1	4	41	40	34	44	43	37	97	98	98
AIPW (m, c, CBPS)	-3	-2	2	41	38	35	42	41	36	96	97	98
AIPW (m, m, CBPS)	43	3	-40	54	38	49	51	39	46	89	97	90
MW (c, GLMPS)	-31	-9	22	38	38	32	38	39	32	95	95	98
MW (m, GLMPS)	29	-5	-34	49	38	44	49	38	44	77	95	78
MW (c, CBPS)	-34	-10	24	41	40	33	44	45	35	97	97	99
MW (m, CBPS)	27	-7	-34	50	40	46	53	42	48	81	96	83
AMW (c, c, GLMPS)	-32	-9	22	38	39	32	40	41	34	96	96	98
AMW (c, m, GLMPS)	-30	-9	21	38	38	33	41	40	34	96	96	98
AMW (m, c, GLMPS)	-17	-11	6	39	39	32	41	42	35	94	96	94
AMW (m, m, GLMPS)	28	-6	-34	49	39	44	50	40	45	78	96	78
AMW (c, c, CBPS)	-34	-10	24	40	40	34	44	45	36	97	97	98
AMW (c, m, CBPS)	-33	-10	23	40	40	34	44	44	37	97	97	99
AMW (m, c, CBPS)	-18	-11	7	40	40	33	43	44	37	95	97	95
AMW (m, m, CBPS)	28	-6	-34	50	40	46	52	42	47	81	96	81
OW (c, GLMPS)	-27	-8	19	36	36	30	36	36	30	95	94	97
OW (m, GLMPS)	32	-4	-36	47	36	43	46	36	42	75	94	76
OW (c, CBPS)	-32	-10	22	38	38	32	40	40	33	97	96	98
OW (m, CBPS)	30	-6	-37	48	37	45	50	38	45	80	95	80
AOW (c, c, GLMPS)	-28	-9	19	37	38	32	39	40	33	96	97	98
AOW (c, m, GLMPS)	-28	-8	20	37	37	32	40	39	34	97	96	98
AOW (m, c, GLMPS)	-14	-10	4	38	38	31	40	41	34	94	97	94
AOW (m, m, GLMPS)	31	-5	-36	48	37	43	48	39	44	79	96	79
AOW (c, c, CBPS)	-32	-9	22	39	39	33	43	43	36	97	97	98
AOW (c, m, CBPS)	-31	-9	22	39	39	33	43	43	36	97	97	99
AOW (m, c, CBPS)	-15	-10	5	39	39	33	42	43	36	95	97	95
AOW (m, m, CBPS)	30	-5	-36	49	39	45	51	41	47	82	96	82
MCOV (c)	5	9	4	40	40	35	40	40	37	94	93	96
MCOV (m)	43	9	-34	49	41	46	49	40	46	87	93	89
MGPSS (c, GLMPS)	0	0	0	46	46	37	40	39	32	92	91	96
MGPSS (m, GLMPS)	44	3	-40	59	44	55	52	39	47	86	93	88
MGPSS (c, CBPS)	4	2	-2	45	44	40	46	45	40	95	96	96
MGPSS (m, CBPS)	45	5	-40	60	43	57	57	43	53	88	95	91
MGPSV (c, GLMPS)	0	1	1	44	44	35	42	41	34	94	93	96
MGPSV (m, GLMPS)	43	3	-39	57	43	51	53	41	48	87	94	90
MGPSV (c, CBPS)	2	3	0	43	43	36	42	41	35	94	94	96
MGPSV (m, CBPS)	44	5	-39	56	43	51	53	41	48	87	93	89

Table A.15: Performance of different causal inference methods in scenario 5 ($n = 300$) of simulation studies. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

Methods	Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3	1v2	1v3	2v3
NAIVE	101	93	96	84	90	84	80	59	64	75	83	74	115	106	144
OREG (c)	97	97	97	84	86	81	80	74	71	76	77	72	101	105	104
OREG (m)	98	94	99	91	85	86	87	75	75	82	76	77	120	99	131
PEN-GAM (c, c, GLMPS)	107	108	101	107	120	111	99	95	103	102	113	103	147	160	191
PEN-GAM (c, m, GLMPS)	106	107	100	106	118	109	99	94	102	103	112	102	145	159	189
PEN-GAM (m, c, GLMPS)	108	108	102	106	114	107	98	90	99	94	112	98	144	157	184
PEN-GAM (m, m, GLMPS)	108	106	103	109	114	109	100	88	99	94	111	98	162	155	207
IPW (c, GLMPS)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
IPW (m, GLMPS)	99	96	101	104	96	102	103	94	98	113	88	105	128	96	142
IPW (c, CBPS)	98	98	99	100	102	97	100	92	85	102	103	100	104	105	99
IPW (m, CBPS)	99	95	100	105	97	99	102	87	87	90	94	85	131	100	141
AIPW (c, c, GLMPS)	98	98	98	93	99	93	87	83	80	99	109	103	107	113	109
AIPW (c, m, GLMPS)	100	100	99	104	102	103	97	95	94	115	117	117	122	114	123
AIPW (m, c, GLMPS)	97	97	97	93	92	91	87	79	79	132	85	119	114	108	119
AIPW (m, m, GLMPS)	98	95	99	110	92	104	100	80	88	211	89	186	163	103	182
AIPW (c, c, CBPS)	97	97	97	104	108	99	98	88	84	111	113	110	115	119	114
AIPW (c, m, CBPS)	98	97	98	108	110	105	104	93	91	118	116	116	118	117	118
AIPW (m, c, CBPS)	97	97	97	95	98	92	89	81	79	93	93	85	113	112	116
AIPW (m, m, CBPS)	98	95	99	106	97	101	100	82	86	98	96	90	137	107	149
MW (c, GLMPS)	103	100	106	105	103	104	108	91	88	106	103	102	103	105	102
MW (m, GLMPS)	103	98	105	105	101	102	107	88	86	93	93	91	130	104	142
MW (c, CBPS)	106	102	110	122	120	118	130	108	102	130	131	122	119	121	113
MW (m, CBPS)	104	99	106	116	111	111	121	98	95	108	107	102	142	114	153
AMW (c, c, GLMPS)	101	99	104	103	103	103	104	87	83	102	100	99	108	111	108
AMW (c, m, GLMPS)	102	98	105	104	103	104	106	87	84	104	101	100	108	108	109
AMW (m, c, GLMPS)	101	99	102	97	101	97	98	85	80	91	91	89	109	113	112
AMW (m, m, GLMPS)	103	97	104	104	100	102	105	85	83	92	91	90	133	108	144
AMW (c, c, CBPS)	103	99	106	114	114	115	118	98	89	120	121	117	118	121	116
AMW (c, m, CBPS)	104	100	107	116	115	115	121	102	92	123	123	119	119	120	119
AMW (m, c, CBPS)	101	99	102	104	107	104	107	90	82	101	99	99	116	119	119
AMW (m, m, CBPS)	103	97	104	111	107	108	114	92	87	102	100	99	140	114	152
OW (c, GLMPS)	99	97	101	97	96	97	98	85	84	98	94	95	95	97	96
OW (m, GLMPS)	100	95	101	99	94	97	99	83	84	87	87	86	123	97	135
OW (c, CBPS)	101	98	103	113	110	110	118	100	97	121	120	114	108	110	105
OW (m, CBPS)	100	95	101	108	103	104	111	92	91	99	98	95	132	104	144
AOW (c, c, GLMPS)	99	97	100	99	99	98	97	82	81	99	97	96	105	108	108
AOW (c, m, GLMPS)	100	97	102	102	99	101	100	83	83	102	98	99	106	106	109
AOW (m, c, GLMPS)	99	97	98	94	97	93	92	81	78	88	87	85	106	110	110
AOW (m, m, GLMPS)	100	95	101	101	97	98	100	81	82	90	88	88	130	105	142
AOW (c, c, CBPS)	100	97	102	110	110	110	111	92	86	116	118	114	114	117	115
AOW (c, m, CBPS)	102	98	104	113	112	112	115	96	90	121	120	117	116	116	117
AOW (m, c, CBPS)	99	97	99	100	104	100	100	86	80	97	95	94	112	116	117
AOW (m, m, CBPS)	101	95	101	107	103	105	108	87	85	100	97	96	137	110	150
MCOV (c)	111	107	110	100	103	99	96	77	82	94	94	91	106	108	118
MCOV (m)	114	107	112	103	104	102	100	80	85	93	94	90	132	109	150
MGPSS (c, GLMPS)	117	113	111	105	103	101	100	93	95	97	100	95	107	105	104
MGPSS (m, GLMPS)	118	112	115	115	106	110	108	94	102	94	98	93	140	107	153
MGPSS (c, CBPS)	123	120	120	119	126	121	116	116	120	122	136	130	123	123	127
MGPSS (m, CBPS)	122	117	121	123	117	120	118	107	115	103	117	107	153	117	170
MGPSV (c, GLMPS)	116	113	114	109	110	105	107	99	98	106	111	105	111	111	109
MGPSV (m, GLMPS)	117	111	115	114	109	109	110	93	100	100	103	95	142	111	154
MGPSV (c, CBPS)	116	113	114	108	111	105	106	97	97	105	112	105	112	112	112
MGPSV (m, CBPS)	117	111	115	114	109	109	109	93	99	97	102	94	141	111	154

Table A.16: $100 \times$ Ratio of 95% confidence interval width to 95% confidence interval width of GLMPS based IPW(c) for $n = 300$. For methods that involve both models, the first and second letter in the parentheses correspond to the treatment model and outcome model, respectively.

		Uncensored Subjects ($N = 1955$)	Censored Subjects ($N = 669$)
Variable		Count (%)	Count (%)
Treatment			
	Docetaxel	565 (28.9)	161 (24.1)
	Abiraterone	783 (40.1)	254 (38.0)
	Enzalutamide	476 (24.3)	163 (24.4)
	Sipuleucel-T	131 (6.7)	91 (13.6)
Age			
	<65	255 (13.0)	121 (18.1)
	65-74	657 (33.6)	231 (34.5)
	≥ 75	1043 (53.4)	317 (47.4)
Race			
	White	1310 (67.0)	471 (70.4)
	Black	249 (12.7)	71 (10.6)
	Other	396 (20.3)	127 (19.0)
Education level			
	High School Diploma or Less	581 (29.7)	187 (28.0)
	High School Graduate and Less than Bachelor Degree	970 (49.6)	340 (50.8)
	Bachelor Degree Plus	274 (14.0)	102 (15.2)
	Unknown	130 (6.6)	40 (6.0)
Household income range			
	<50k	669 (34.2)	202 (30.2)
	50k-100k	613 (31.4)	226 (33.8)
	>100k	376 (19.2)	153 (22.9)
	Unknown	297 (15.2)	188 (13.2)
Geographic Region			
	South Atlantic	357 (18.3)	125 (18.7)
	New England	100 (4.9)	29 (4.3)
	Middle Atlantic	197 (10.3)	74 (11.1)
	East North Central	317 (16.2)	119 (17.8)
	East South Central	71 (3.6)	26 (3.9)
	West North Central	181 (9.3)	55 (8.2)
	West South Central	196 (10.0)	63 (9.4)
	Mountain	241 (12.3)	76 (11.4)
	Pacific	295 (15.1)	102 (15.2)
Product			
	HMO	599 (30.6)	171 (25.6)
	PPO	132 (6.8)	42 (6.3)
	Other	1224 (62.6)	456 (68.2)
Metastatic (Yes)		1607 (82.2)	515 (77.0)
ASO (Yes)		238 (12.2)	86 (12.9)
Year of First Prescription			
	2014	812 (41.5)	235 (35.1)
	2015	846 (43.3)	199 (29.7)
	2016	297 (15.2)	235 (35.1)
Diabetes		574 (29.4)	191 (28.6)
Hypertension		1432 (73.2)	475 (71.0)
Arrhythmia		489 (25.0)	148 (22.1)
CHF		234 (12.0)	86 (12.9)
Osteoporosis		160 (8.2)	63 (9.4)
Provider Type			
	Medical oncologist	1196 (61.2)	409 (61.1)
	Others	759 (38.8)	260 (38.9)

Table A.17: Characteristics of censored vs. uncensored subjects.

	Total ($N = 1955$)	Docetaxel ($N = 565$)	Abiraterone ($N = 783$)	Enzalutamide ($N = 476$)	Sipuleucel-T ($N = 131$)
Variable	Count (%)	Count (%)	Count (%)	Count (%)	Count (%)
Age					
<65	255 (13.0)	96 (17.0)	92 (11.7)	49 (10.3)	18 (13.7)
65-74	657 (33.6)	267 (47.3)	216 (27.6)	126 (26.5)	48 (36.6)
≥75	1043 (53.4)	202 (35.8)	475 (60.7)	301 (63.2)	65 (49.6)
Race					
White	1310 (67.0)	390 (60.0)	516 (65.9)	318 (66.8)	86 (65.6)
Black	249 (12.7)	49 (8.7)	106 (13.5)	74 (15.5)	20 (15.3)
Other	396 (20.3)	126 (22.3)	161 (20.6)	84 (17.6)	25 (19.1)
Education level					
High School Diploma or Less	581 (29.7)	144 (25.5)	260 (33.2)	142 (29.8)	35 (26.7)
High School Graduate and Less than Bachelor Degree	970 (49.6)	277 (49.0)	389 (49.7)	241 (50.6)	63 (48.1)
Bachelor Degree Plus	274 (14.0)	87 (15.4)	101 (12.9)	69 (14.5)	17 (13.0)
Unknown	130 (6.6)	57 (10.1)	33 (4.2)	24 (5.0)	16 (12.2)
Household income range					
<50k	669 (34.2)	145 (25.7)	308 (39.3)	174 (36.6)	42 (32.1)
50k-100k	613 (31.4)	161 (28.5)	244 (31.2)	164 (34.5)	44 (33.6)
>100k	376 (19.2)	150 (26.5)	137 (17.5)	66 (13.9)	23 (17.6)
Unknown	297 (15.2)	109 (19.3)	94 (12.0)	72 (15.1)	22 (16.8)
Geographic Region					
South Atlantic	357 (18.3)	101 (17.9)	144 (18.4)	87 (18.3)	25 (19.1)
New England	100 (4.9)	33 (5.8)	42 (5.4)	21 (4.4)	4 (3.1)
Middle Atlantic	197 (10.3)	56 (9.9)	65 (8.3)	61 (12.8)	15 (11.5)
East North Central	317 (16.2)	82 (14.5)	131 (16.7)	79 (16.6)	25 (19.1)
East South Central	71 (3.6)	21 (3.7)	26 (3.3)	15 (3.2)	9 (6.9)
West North Central	181 (9.3)	112 (19.8)	42 (5.4)	20 (4.2)	7 (5.3)
West South Central	196 (10.0)	55 (9.7)	83 (10.6)	43 (9.0)	15 (11.5)
Mountain	241 (12.3)	58 (10.3)	95 (12.1)	64 (13.4)	24 (18.3)
Pacific	295 (15.1)	47 (8.3)	155 (19.8)	86 (18.1)	7 (5.3)
Product					
HMO	599 (30.6)	162 (28.7)	256 (32.7)	148 (31.1)	33 (25.2)
PPO	132 (6.8)	35 (6.2)	62 (7.9)	26 (5.5)	9 (6.9)
Other	1224 (62.6)	368 (65.1)	465 (59.4)	302 (63.4)	89 (67.9)
Metastatic (Yes)	1607 (82.2)	483 (85.5)	644 (82.2)	365 (76.7)	115 (87.8)
ASO (Yes)	238 (12.2)	66 (11.7)	102 (13.0)	56 (11.8)	14 (10.7)
Year of First Prescription					
2014	812 (41.5)	208 (36.8)	383 (48.9)	174 (36.6)	47 (35.9)
2015	846 (43.3)	262 (46.4)	303 (38.7)	222 (46.6)	59 (45.0)
2016	297 (15.2)	95 (16.8)	97 (12.4)	80 (16.8)	25 (19.1)
Diabetes	574 (29.4)	147 (26.0)	228 (29.1)	154 (32.4)	45 (34.4)
Hypertension	1432 (73.2)	402 (71.2)	577 (73.7)	350 (73.5)	103 (78.6)
Arrhythmia	489 (25.0)	128 (22.7)	203 (25.9)	130 (27.3)	28 (21.4)
CHF	234 (12.0)	42 (7.4)	103 (13.2)	75 (15.8)	14 (10.7)
Osteoporosis	160 (8.2)	30 (5.3)	66 (8.4)	43 (9.0)	21 (16.0)
Provider Type					
Medical oncologist	1196 (61.2)	321 (56.8)	565 (72.2)	280 (58.8)	30 (22.9)
Others	759 (38.8)	244 (43.2)	218 (27.8)	196 (41.2)	101 (77.1)

Table A.18: Characteristics of uncensored subjects in the four treatment groups of interest.

Method	A – D	E – D	S – D	E – A	S – A	S – E
NAIVE	*-0.130* *(-0.186, -0.073)*	*-0.177* *(-0.239, -0.115)*	*-0.099* *(-0.197, -0.001)*	-0.047 (-0.105, 0.010)	0.031 (-0.064, 0.126)	0.078 (-0.020, 0.177)
OREG	-0.128 (-0.187, 0.068)	*-0.183* *(-0.248, -0.118)*	-0.108 (-0.231, 0.036)	-0.055 (-0.115, 0.004)	0.020 (-0.121, 0.161)	0.075 (-0.069, 0.219)
PEN-GAM	*-0.120* *(-0.182, -0.058)*	*-0.174* *(-0.241, -0.107)*	-0.169 (-0.396, 0.058)	-0.054 (-0.115, 0.007)	-0.049 (-0.277, 0.178)	0.005 (-0.225, 0.235)
IPW	*-0.126* *(-0.187, -0.066)*	*-0.176* *(-0.243, -0.108)*	-0.058 (-0.259, 0.144)	-0.05 (-0.110, 0.011)	0.068 (-0.129, 0.266)	0.118 (-0.083, 0.319)
AIPW	*-0.125* *(-0.185, -0.064)*	*-0.179* *(-0.246, -0.113)*	-0.095 (-0.263, 0.074)	-0.055 (-0.115, 0.005)	0.03 (-0.137, 0.196)	0.085 (-0.084, 0.253)
MW	*-0.103* *(-0.199, -0.007)*	*-0.179* *(-0.272, -0.085)*	-0.091 (-0.206, 0.024)	-0.076 (-0.166, 0.014)	0.011 (-0.101, 0.124)	0.087 (-0.024, 0.199)
AMW	-0.085 (-0.181, 0.012)	*-0.161* *(-0.256, -0.065)*	-0.079 (-0.189, 0.032)	-0.076 (-0.166, 0.014)	0.006 (-0.099, 0.111)	0.082 (-0.024, 0.188)
OW	*-0.111* *(-0.189, -0.033)*	*-0.174* *(-0.252, -0.095)*	-0.076 (-0.187, 0.035)	-0.063 (-0.137, 0.011)	0.035 (-0.073, 0.143)	0.098 (-0.012, 0.207)
AOW	*-0.099* *(-0.178, -0.020)*	*-0.163* *(-0.244, -0.082)*	-0.084 (-0.191, 0.024)	-0.064 (-0.139, 0.011)	0.015 (-0.087, 0.118)	0.080 (-0.026, 0.185)
MCOV	*-0.144* *(-0.211, -0.076)*	*-0.202* *(-0.272, -0.131)*	*-0.149* *(-0.291, -0.008)*	-0.058 (-0.121, 0.005)	-0.006 (-0.143, 0.132)	0.052 (-0.087, 0.192)
MGPSS	*-0.131* *(-0.197, -0.064)*	*-0.195* *(-0.272, -0.119)*	-0.019 (-0.232, 0.195)	-0.065 (-0.135, 0.005)	0.112 (-0.100, 0.324)	0.177 (-0.037, 0.391)
MGPSV	*-0.125* *(-0.197, -0.053)*	*-0.164* *(-0.240, -0.087)*	-0.125 (-0.301, 0.051)	-0.039 (-0.107, 0.029)	0 (-0.173, 0.173)	0.039 (-0.136, 0.040)

Table A.19: Difference (95% confidence interval) in probability of at least one emergency room visit within 180 days of first prescription across four treatment groups ($N = 1776$). Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, Sipuleucel-T. 95% confidence intervals that exclude 0 are italicized. The 95% confidence intervals were calculated using: (1) bootstrapped standard errors from 50 bootstrap samples for OREG, IPW, AIPW, MW, AMW, OW, AOW, and CBPS-based MGPSS; (2) Wald-type confidence interval based on original data for NAIVE; (3) Abadie and Imbens [31] confidence interval for MCOV and both GLMPS- and CBPS-based MGPSV; (4) Abadie and Imbens [48] confidence interval for GLMPS-based MGPSS; (5) Rubin's imputation rule for PEN-GAM [51]

Methods	Computational time (milliseconds)
Estimation of GPS	
GLMPS	19.2
Just-identified CBPS	1423.3
Over-identified CBPS	2460.9
Estimation of ATE	
OREG	8.7
IPW	19.9
AIPW	29.3
MW	22.8
OW	23.8
MCOV	218.5
MGPSS	514.6
MGPSV	171.9
PENCOMP	1858.2

Table A.20: Average computational time across 100 simulated datasets for the methods under comparison. GLMPS was used for propensity-based methods when estimating ATE.

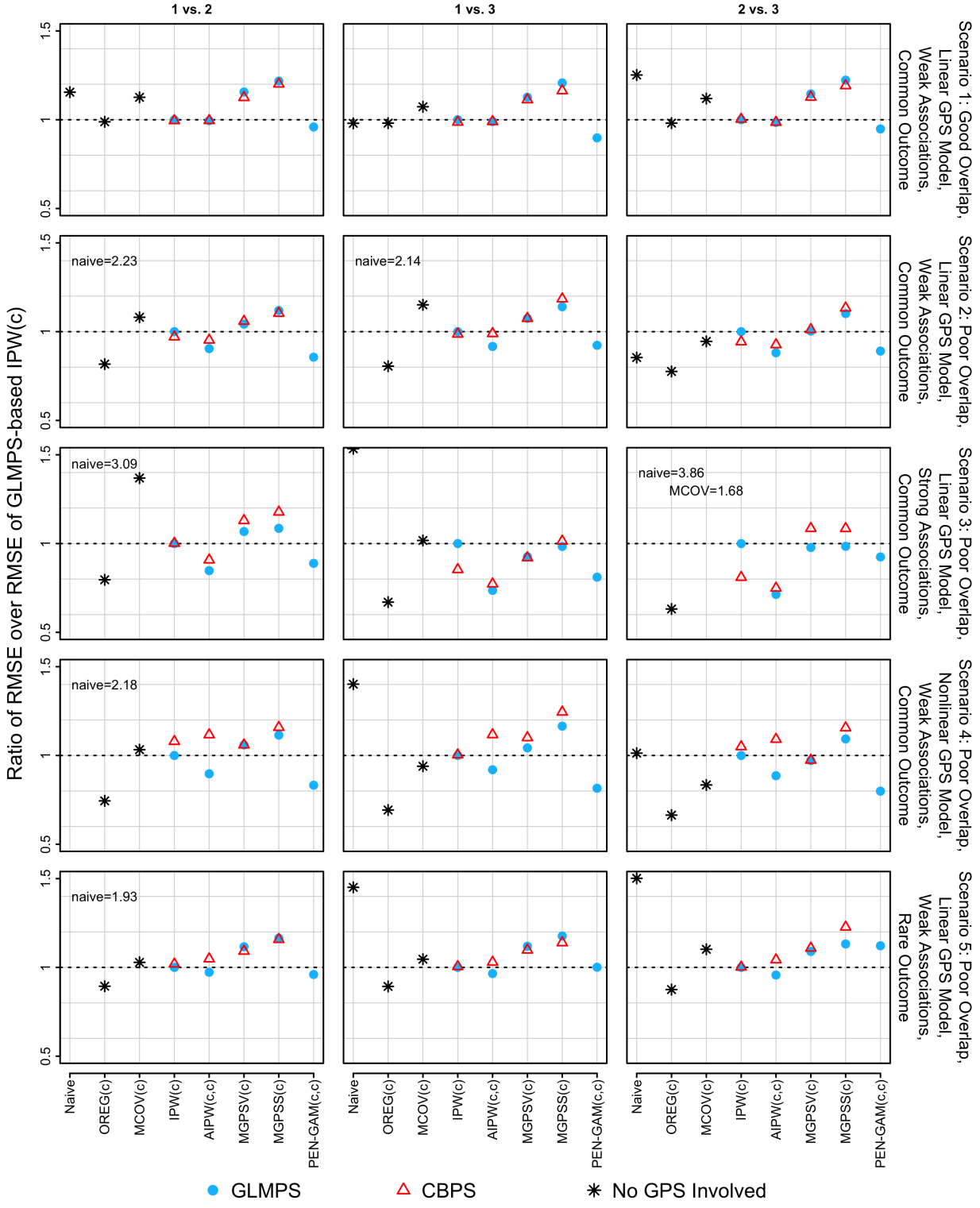

Figure A.3: Ratio of RMSE over RMSE of GLMPS-based IPW(c) for $n = 300$ across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

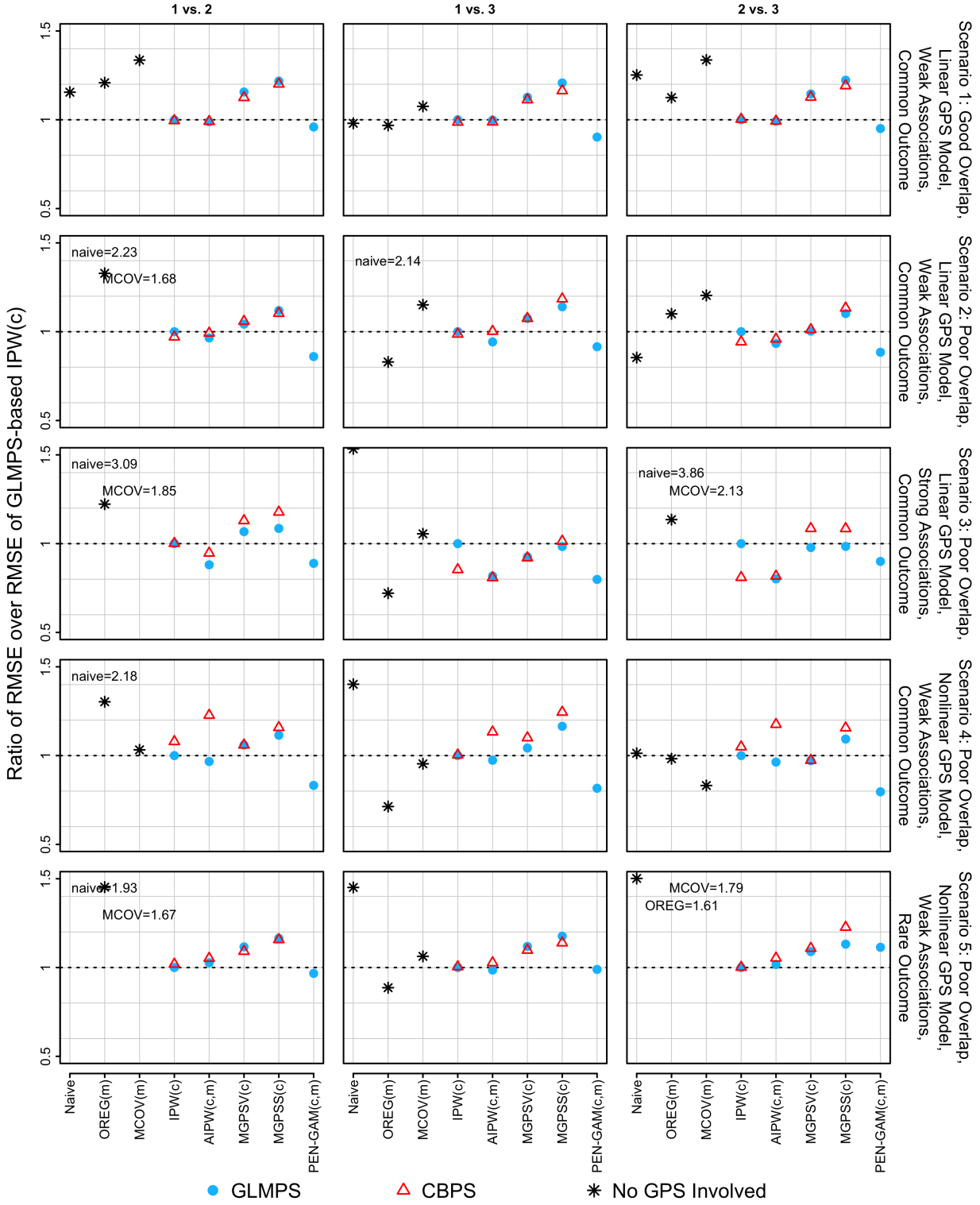

Figure A.4: Ratio of RMSE over RMSE of GLMPS-based IPW(c) for $n = 300$ across methods based on a correctly specified propensity model only. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

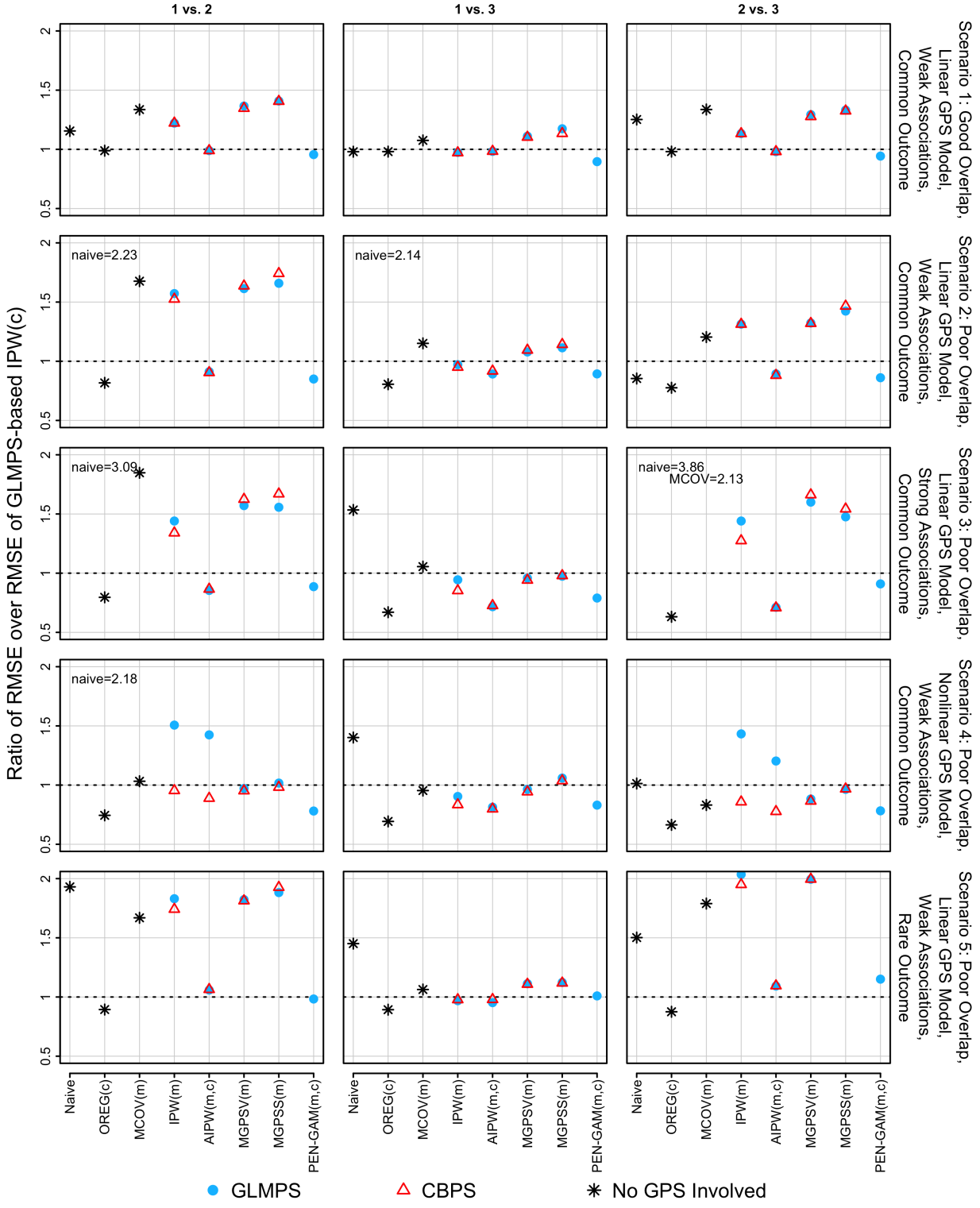

Figure A.5: Ratio of RMSE over RMSE of GLMPS-based IPW(c) for $n = 300$ across methods based on a correctly specified outcome model only. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

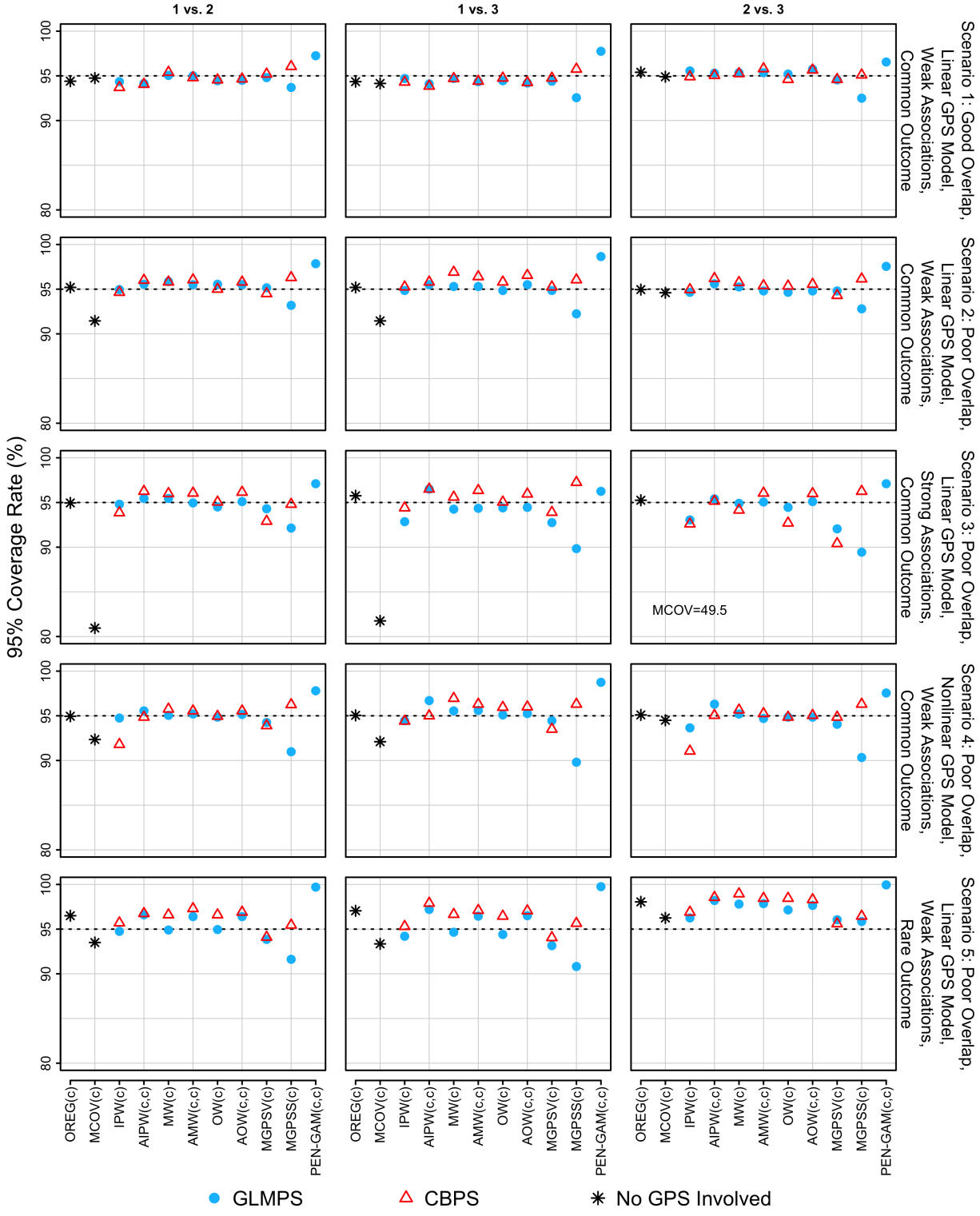

Figure A.6: 95% Coverage probability for $n = 300$ across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

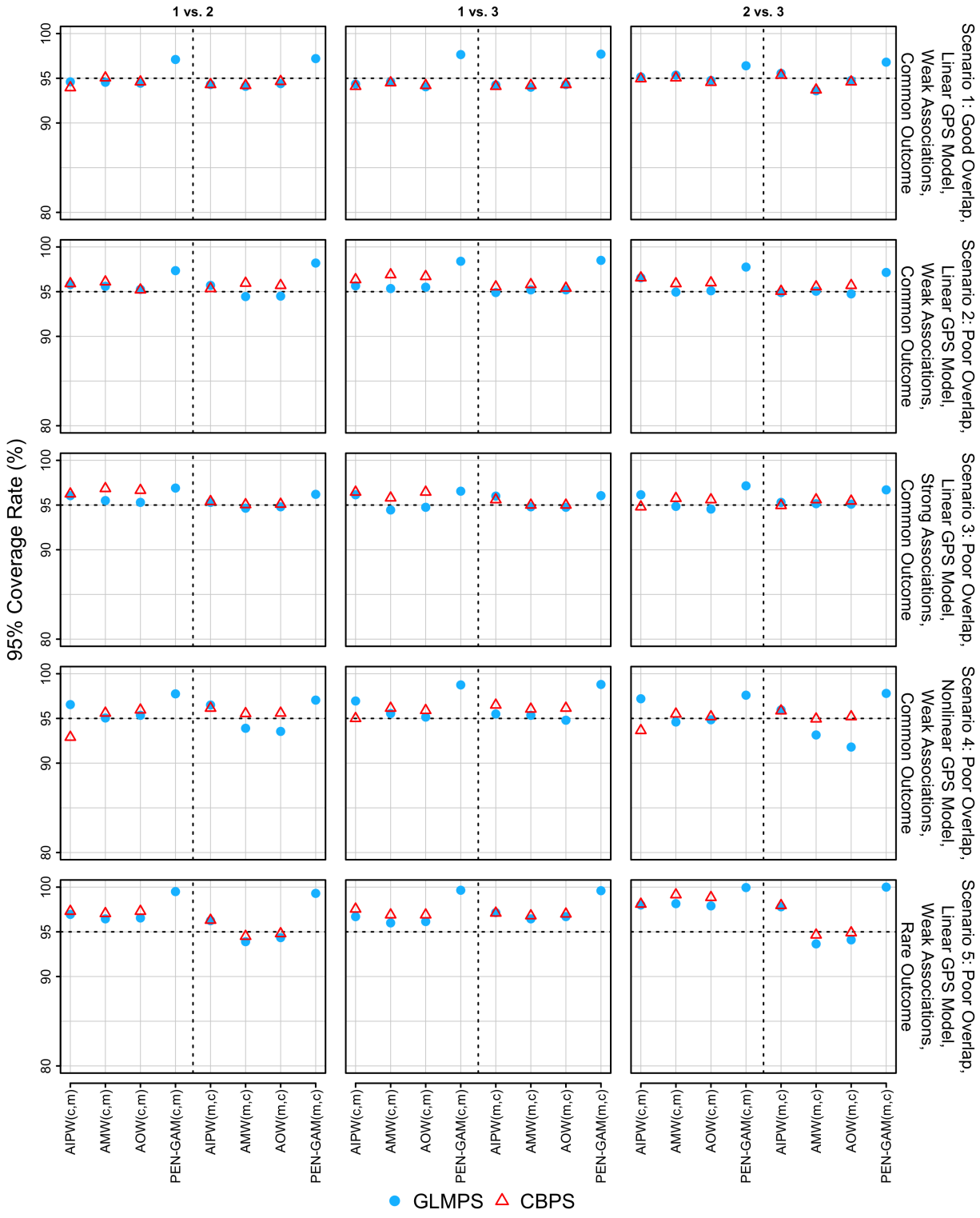

Figure A.7: 95% Coverage probability for $n = 300$ across methods based on a correctly specified propensity score or outcome model. For methods that involve both models, the first and second letter in the parentheses correspond to the propensity and outcome model, respectively. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

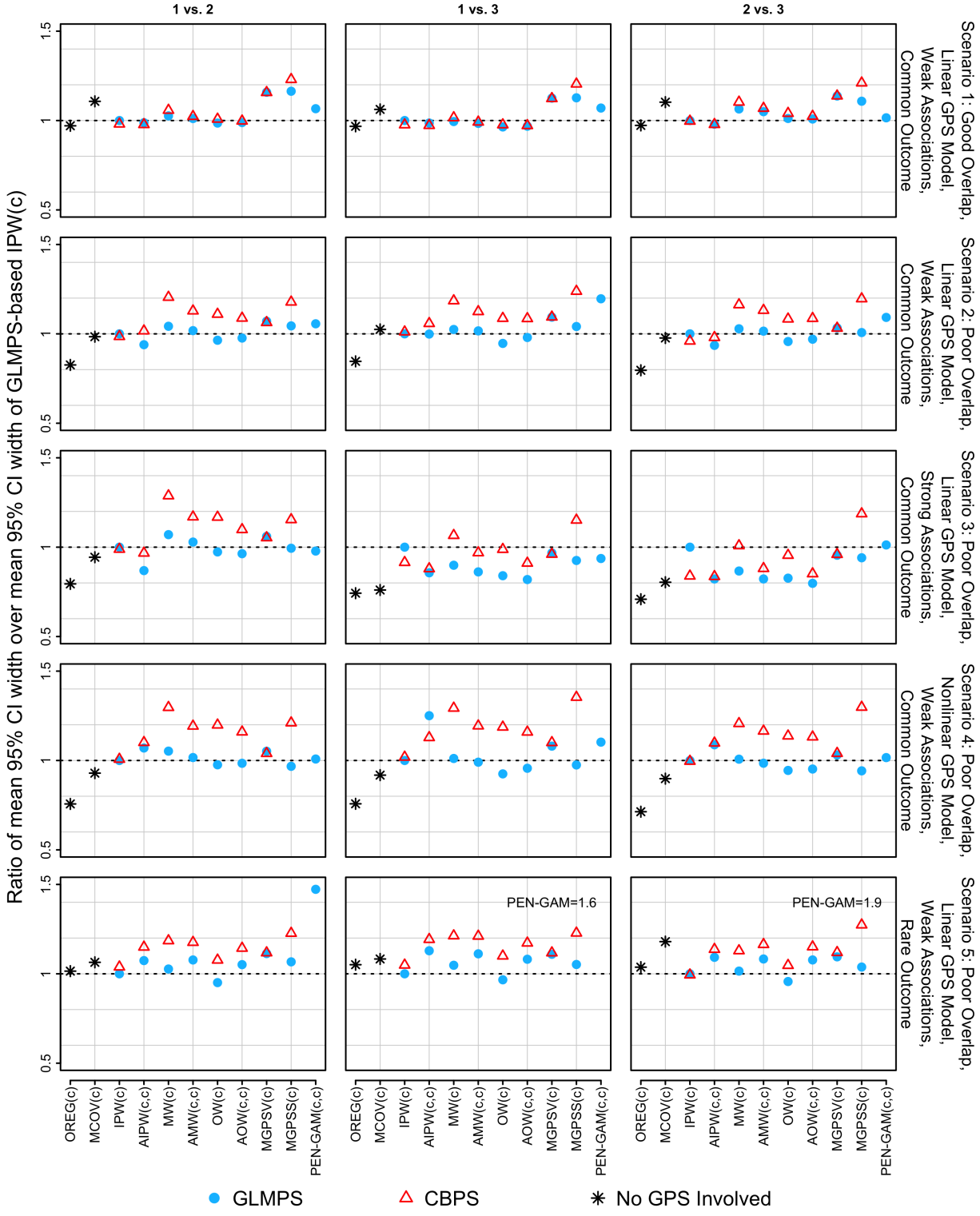

Figure A.8: Ratio of mean 95% CI width over mean 95% CI width of GLMPS-based IPW(c) for $n = 300$ across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

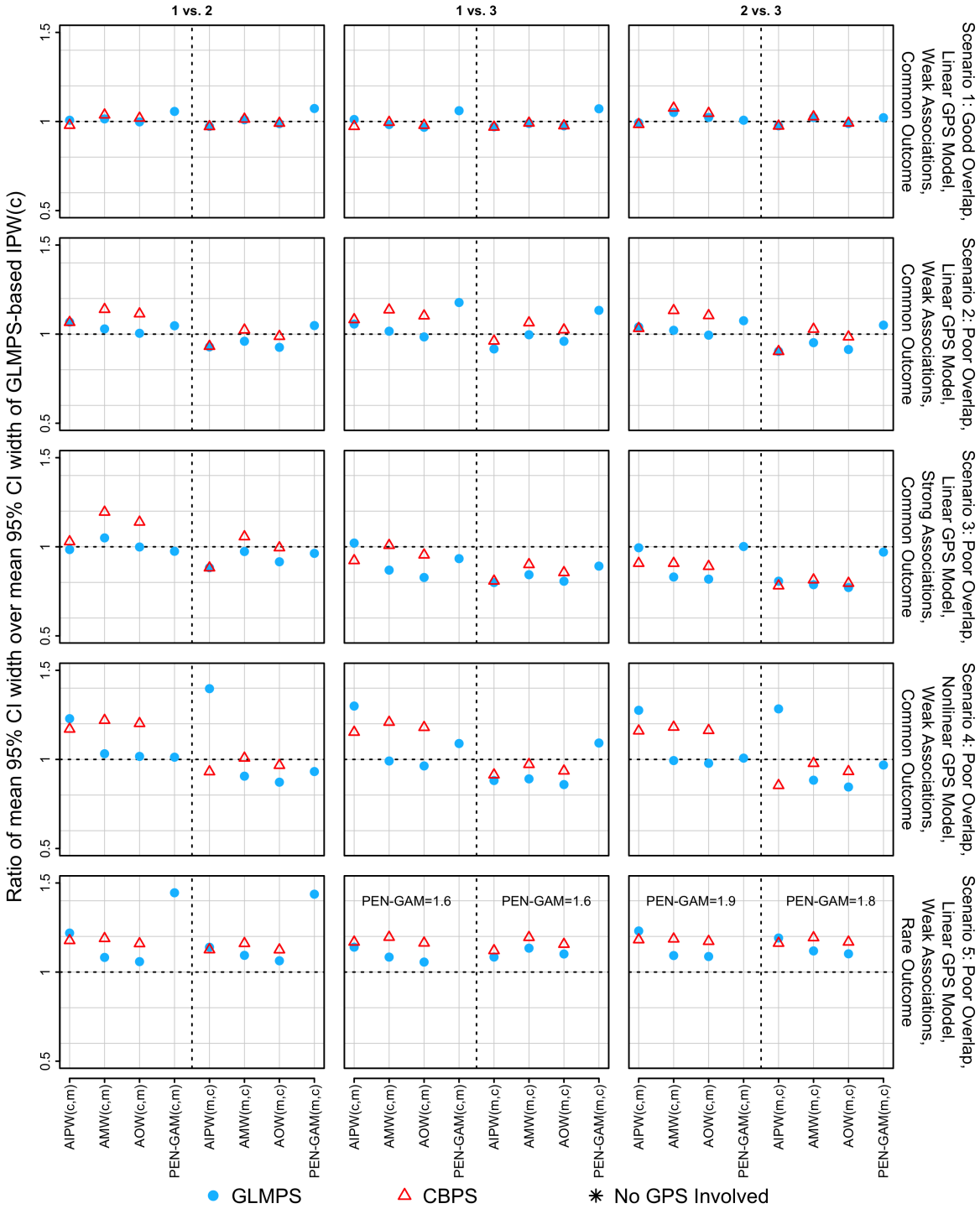

Figure A.9: Ratio of mean 95% CI width over mean 95% CI width of GLMPS-based IPW(c) for $n = 300$ across methods based on a correctly specified propensity score or outcome model. For methods that involve both models, the first and second letter in the parentheses correspond to the propensity and outcome model, respectively. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets.

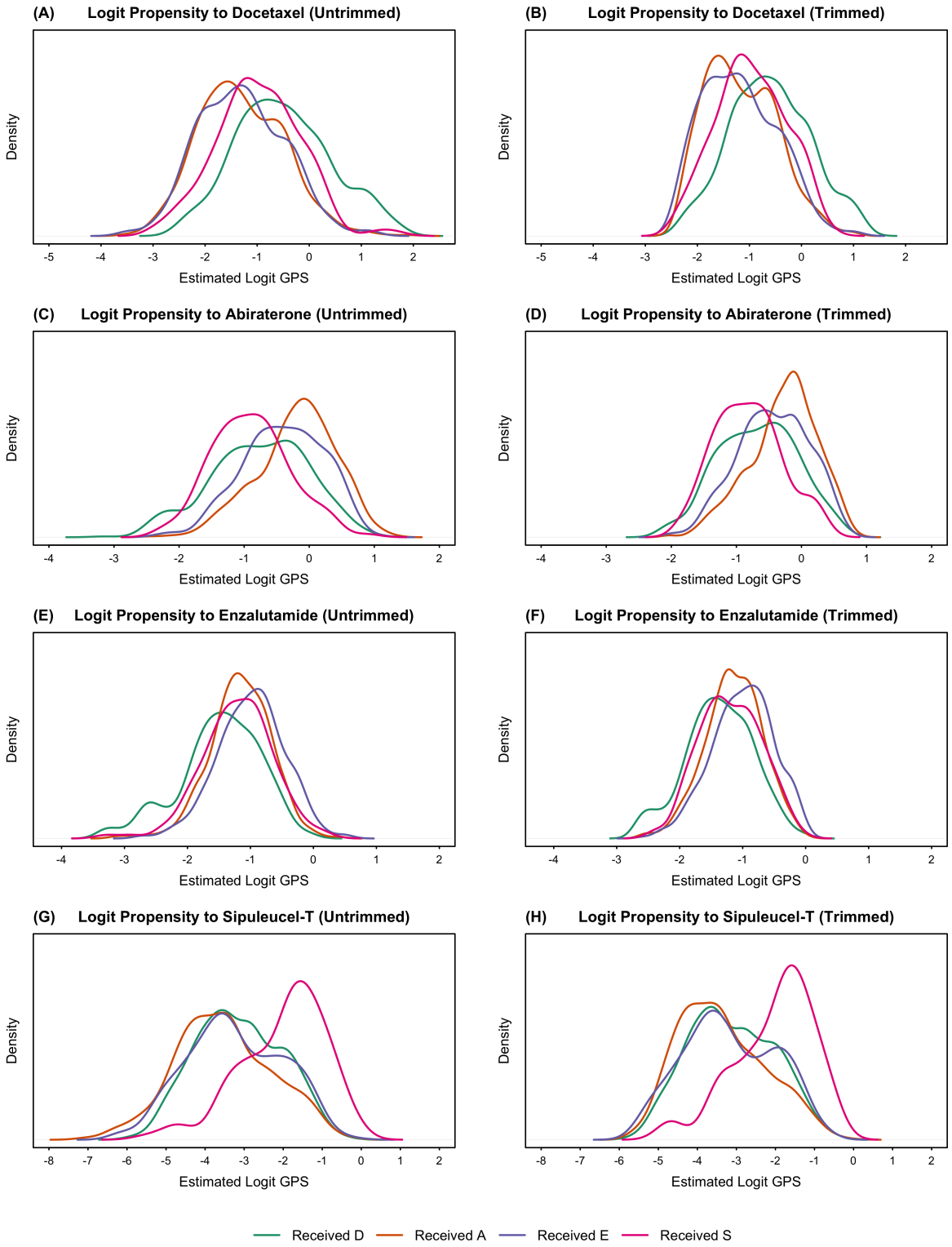

Figure A.10: Distribution of the estimated generalized propensity scores in logit scale for the original ($N = 1955$) and trimmed samples ($N = 1777$).

APPENDIX B

Supplement for Chapter II

B.1 Derivation of the Asymptotic Distribution of $\hat{\mu}_j$

B.1.1 Notations

We first list the notations used in the proof. Let i index the subject and j index the treatment group, with $i = 1, \cdots, n$ and $j = 1, \cdots, J$. Same as the notations in the manuscript, $\tilde{\boldsymbol{X}}_i$, Z_i, T_i, and C_i denote the covariates, treatment received, time to event, and censoring time, respectively. Let $\Delta_i = I(T_i \leq C_i)$, $R_i = I\{C_i \geq \min(T_i, d)\}$, and $L_i = \min(T_i, C_i, d)$, where d is a fixed time point. $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are sets of covariates that are associated with treatment assignment and censoring, respectively. The conditional hazard function of C_i given $\boldsymbol{W}_i$ and $Z_i = j$ is denoted by $\lambda_{ij}(t)$. We define

$$D_{ij} = I(Z_i = j)$$

$$\tilde{Y}_i = I(T_i > d)$$

$$\boldsymbol{S}_j^{(q)}(t;\boldsymbol{\gamma}_j) = n^{-1}\sum_{i=1}^{n} Y_{ij}(t)\boldsymbol{W}_i^{\otimes q}\exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}_j) \qquad \boldsymbol{s}_j^{(q)}(t;\boldsymbol{\gamma}_j) = E\{\boldsymbol{S}_j^{(q)}(t;\boldsymbol{\gamma}_j)\}$$

$$\overline{\boldsymbol{W}}_j(t;\boldsymbol{\gamma}_j) = \frac{\boldsymbol{S}_j^{(1)}(t;\boldsymbol{\gamma}_j)}{S_j^{(0)}(t;\boldsymbol{\gamma}_j)} \qquad \overline{\boldsymbol{w}}_j(t;\boldsymbol{\gamma}_j) = \frac{\boldsymbol{s}_j^{(1)}(t;\boldsymbol{\gamma}_j)}{s_j^{(0)}(t;\boldsymbol{\gamma}_j)}$$

$$m_{ij}(\boldsymbol{\beta}_j) = \text{expit}(\boldsymbol{X}_i^T\boldsymbol{\beta}_j)$$

$$\pi_{ij}(\boldsymbol{\alpha}) = \exp(\boldsymbol{V}_i^T\boldsymbol{\alpha}_j)/\sum_{z=1}^{J}\exp(\boldsymbol{V}_i^T\boldsymbol{\alpha}_z), \quad \text{where } \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_J)^T$$

$$d\Lambda_{0j}^*(t) = \frac{E\{dN_{ij}(t)\}}{s_j^{(0)}(t;\boldsymbol{\gamma}_j^*)}$$

$$d\Lambda_{ij}^*(t) = \exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}_j^*)d\Lambda_{0j}^*(t)$$

$$dM_{ij}^*(t) = dN_{ij}(t) - Y_{ij}(t)d\Lambda_{ij}^*(t)$$

with the counting process defined by $N_{ij}(t) = D_{ij} I\{\min(T_i, C_i) \leq t, \Delta_i = 1\}$ and the at-risk process defined by $Y_{ij}(t) = D_{ij} I\{\min(T_i, C_i) \geq t\}$.

B.1.2 Model for Treatment Assignment

We consider a multinomial logistic regression model for the treatment assignment, specified as

$$\frac{\log P(Z_i = l|\boldsymbol{V}_i)}{\log P(Z_i = J|\boldsymbol{V}_i)} = \boldsymbol{V}_i^T \boldsymbol{\alpha}_l, \quad l = 1, \cdots, J-1.$$

Under some regularity conditions [75], $\hat{\boldsymbol{\alpha}}_l$ converges in probability to a constant vector $\boldsymbol{\alpha}_l^*$, denoted by $\hat{\boldsymbol{\alpha}}_l \xrightarrow{p} \boldsymbol{\alpha}_l^*$, and

$$n^{1/2}(\hat{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_l^*) = \boldsymbol{H}_l^{-1}(\boldsymbol{\alpha}^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{V}_i \left\{ D_{il} - \pi_{il}(\boldsymbol{\alpha}^*) \right\} + o_p(1) \tag{B.1}$$

where

$$\boldsymbol{H}_l(\boldsymbol{\alpha}^*) = E\left\{ \boldsymbol{V}_i \boldsymbol{V}_i^T \pi_{il}(\boldsymbol{\alpha}^*) \left[1 - \pi_{il}(\boldsymbol{\alpha}^*)\right] \right\}.$$

If the model for $P(Z_i = l|\boldsymbol{V}_i)$, where $l = 1, \cdots, J-1$, is correctly specified, $\boldsymbol{\alpha}^*$ equals the truth $\boldsymbol{\alpha}^0$.

B.1.3 Model for Censoring

We assume a Cox proportional hazard model for the censoring time hazard, given by

$$\lambda_{ij}(t) \equiv \lambda(t|Z_i = j, \boldsymbol{W}_i) = \lambda_{0j}(t) \exp\left(\boldsymbol{W}_i^T \boldsymbol{\gamma}_j\right), \quad j = 1, \cdots, J. \tag{B.2}$$

Let δ be the time point that satisfies $P\{\min(T_i, C_i) \geq \delta\} > 0$ for $i = 1, \cdots, n$, which practically is set to the maximum observation time. Lin and Wei [76] showed that under some regularity conditions, $\hat{\boldsymbol{\gamma}}_j$ converges in probability to a constant vector $\boldsymbol{\gamma}_j^*$, and $n^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*)$ is asymptotically normal with

$$n^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*) = \boldsymbol{\Omega}_j^{-1}(\boldsymbol{\gamma}_j^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) + o_p(1), \tag{B.3}$$

where $\boldsymbol{\Omega}_j(\boldsymbol{\gamma}_j^*) = \int_0^\delta \left\{ \frac{\boldsymbol{s}_j^{(2)}(t;\boldsymbol{\gamma}_j^*)}{s_j^{(0)}(t;\boldsymbol{\gamma}_j^*)} - \overline{\boldsymbol{w}}_j(t;\boldsymbol{\gamma}_j)^{\otimes 2} \right\} E\{Y_{ij}(t)\lambda_{ij}(t)\} dt$ and $\boldsymbol{U}_{ij}(\boldsymbol{\gamma}_j^*) = \int_0^\delta \{\boldsymbol{W}_i - \overline{\boldsymbol{w}}(t;\boldsymbol{\gamma}_j^*)\} dM_{ij}^*(t)$.

$\Lambda_{0j}(t)$ can be estimated by the Breslow estimator, specified as

$$\hat{\Lambda}_{0j}(t;\hat{\boldsymbol{\gamma}}_j) = \int_0^t \frac{\sum_{i=1}^n dN_{ij}(u)}{\sum_{i=1}^n Y_{ij}(u)\exp\{\boldsymbol{W}_i^T\hat{\boldsymbol{\gamma}}_j\}} \tag{B.4}$$

We make the following decomposition:

$$n^{1/2}\{\hat{\Lambda}_{ij}(t) - \Lambda^*_{ij}(t)\} = n^{1/2}\{\hat{\Lambda}_{ij}(t;\hat{\boldsymbol{\gamma}}_j) - \hat{\Lambda}_{ij}(t;\boldsymbol{\gamma}^*_j)\} \tag{B.5}$$

$$+ n^{1/2}\{\hat{\Lambda}_{ij}(t;\boldsymbol{\gamma}^*_j) - \Lambda^*_{ij}(t)\} \tag{B.6}$$

Applying a Taylor series expansion about $\boldsymbol{\gamma}^*_j$ to (B.5), we have

$$\begin{aligned}(B.5) &= \int_0^t \{\boldsymbol{W}_i - \overline{\boldsymbol{w}}_j(u;\boldsymbol{\gamma}^*_j)\}d\hat{\Lambda}_{ij}(u;\boldsymbol{\gamma}^*_j)n^{1/2}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}^*_j) + o_p(1)\\ &= \boldsymbol{K}^T_{ij}(t;\boldsymbol{\gamma}^*_j)\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma}^*_j)n^{-1/2}\sum_{i=1}^n \boldsymbol{U}_{ij}(\boldsymbol{\gamma}^*_j) + o_p(1)\end{aligned}$$

where $\boldsymbol{K}_{ij}(t;\boldsymbol{\gamma}^*_j) = \int_0^t\{\boldsymbol{W}_i - \overline{\boldsymbol{w}}_j(u;\boldsymbol{\gamma}^*_j)\}d\Lambda^*_{ij}(u)$.

Plugging (B.4) into (B.6), we can write the second term as

$$\begin{aligned}(B.6) &= \exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}^*_j)n^{1/2}\{\hat{\Lambda}_{0j}(t;\boldsymbol{\gamma}^*_j) - \Lambda^*_{0j}(t)\}\\ &= \exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}^*_j)n^{-1/2}\sum_{i=1}^n\int_0^t \frac{dM^*_{ij}(u)}{s^{(0)}(u;\boldsymbol{\gamma}^*_j)} + o_p(1)\end{aligned}$$

It follows from the above results that

$$\begin{aligned}n^{1/2}\{\hat{\Lambda}_{ij}(t) - \Lambda^*_{ij}(t)\} = \boldsymbol{K}^T_{ij}(t;\boldsymbol{\gamma}^*_j)\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma}^*_j)n^{-1/2}\sum_{i=1}^n \boldsymbol{U}_{ij}(\boldsymbol{\gamma}^*_j)\\ + \exp(\boldsymbol{W}_i^T\boldsymbol{\gamma}^*_j)n^{-1/2}\sum_{i=1}^n\int_0^t \frac{dM^*_{ij}(u)}{s^{(0)}(u;\boldsymbol{\gamma}^*_j)} + o_p(1).\end{aligned}$$

When the censoring model (B.2) is correctly specified, $\boldsymbol{\gamma}^*_j$ and Λ^*_{0j} equals the corresponding truth $\boldsymbol{\gamma}^0_j$ and Λ^0_{0j}, respectively.

B.1.4 Model for Outcome

The assumed model for the outcome is

$$\text{logit}\left\{E(\tilde{Y}_i|\tilde{\boldsymbol{X}}_i, Z_i = j)\right\} = \boldsymbol{X}_i^T\boldsymbol{\beta}_j,$$

where estimator for $\boldsymbol{\beta}_j$, denoted by $\hat{\boldsymbol{\beta}}_j$, can be obtained by solving the set of estimating equations

$$0 = n^{-1} \sum_{i=1}^{n} \frac{D_{ij} R_i \boldsymbol{X}_i \{\tilde{Y}_i - m_{ij}(\hat{\boldsymbol{\beta}}_j)\}}{\pi_{ij}(\hat{\boldsymbol{\alpha}}) \exp\{-\hat{\Lambda}_{ij}(L_i)\}} \equiv G\left(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}(L_i)\right).$$

Under suitable regularity conditions, $\hat{\boldsymbol{\beta}}_j \xrightarrow{p} \boldsymbol{\beta}_j^*$. To obtain the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)$, we make the following decomposition:

$$\begin{aligned}
& n^{1/2} G\left(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}\right) - n^{1/2} G\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_{J-1}^*, \Lambda_{ij}^*\right) \\
= \; & n^{1/2} G\left(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}\right) - n^{1/2} G\left(\boldsymbol{\beta}_j^*, \hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}\right) && \text{(B.7)} \\
& + n^{1/2} G\left(\boldsymbol{\beta}_j^*, \hat{\boldsymbol{\alpha}}_1, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}\right) - n^{1/2} G\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_1^*, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}\right) \\
& + \\
& \vdots && \text{(B.8)} \\
& + n^{1/2} G\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_1^*, \cdots, \hat{\boldsymbol{\alpha}}_{J-1}, \hat{\Lambda}_{ij}\right) - n^{1/2} G\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_{J-1}^*, \hat{\Lambda}_{ij}\right) \\
& + n^{1/2} G\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_{J-1}^*, \hat{\Lambda}_{ij}\right) - n^{1/2} G\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}_1^*, \cdots, \boldsymbol{\alpha}_{J-1}^*, \Lambda_{ij}^*\right). && \text{(B.9)}
\end{aligned}$$

Considering (B.7), through a Taylor series expansion of $\hat{\boldsymbol{\beta}}_j$ about $\boldsymbol{\beta}_j^*$,

$$(B.7) = -\boldsymbol{B}_j(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) n^{1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) + o_p(1),$$

where

$$\boldsymbol{B}_j(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) = n^{-1} \sum_{i=1}^{n} \frac{D_{ij} R_i \boldsymbol{X}_i \boldsymbol{X}_i^T m_{ij}(\boldsymbol{\beta}_j^*)\{1 - m_{ij}(\boldsymbol{\beta}_j^*)\}}{\pi_{ij}(\boldsymbol{\alpha}^*) \exp\{-\Lambda_{ij}^*(L_i)\}}.$$

Considering (B.8), using a Taylor series expansion of $\hat{\boldsymbol{\alpha}}_l$ about $\boldsymbol{\alpha}_l^*$ for $l = 1, \cdots, J-1$, and substituting the results of (B.1), we have

$$(B.8) = \boldsymbol{F}_{jl}\left(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*\right) \boldsymbol{H}_l^{-1}(\boldsymbol{\alpha}^*) n^{-1/2} \sum_{i=1}^{n} \boldsymbol{V}_i \{D_{il} - \pi_{il}(\boldsymbol{\alpha}^*)\} + o_p(1),$$

where

$$\boldsymbol{F}_{jl}(\boldsymbol{\beta}_j^*, \boldsymbol{\alpha}^*, \Lambda_{ij}^*) = \begin{cases} n^{-1} \sum_{i=1}^{n} \frac{D_{ij} R_i \boldsymbol{X}_i \boldsymbol{V}_i^T \{1 - \pi_{il}^{-1}(\boldsymbol{\alpha}^*)\}\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j^*)\}}{\exp\{-\Lambda_{ij}^*(L_i)\}} & \text{if} \quad l = j \\ n^{-1} \sum_{i=1}^{n} \frac{D_{ij} R_i \boldsymbol{X}_i \boldsymbol{V}_i^T \exp(\boldsymbol{V}_i^T \boldsymbol{\alpha}_l^*)\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}_j^*)\}}{\{D_{iJ} + (1 - D_{iJ}) \exp(\boldsymbol{V}_i^T \boldsymbol{\alpha}_j^*)\} \exp\{-\Lambda_{ij}^*(L_i)\}} & \text{if} \quad l \neq j \end{cases}$$

Considering (B.9), by Taylor series expansion of $\hat{\Lambda}_{ij}(L_i)$ about $\Lambda^*_{ij}(L_i)$, we have

$$(B.9) = \boldsymbol{P}_j(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\boldsymbol{\Omega}_j^{-1}(\boldsymbol{\gamma}^*_j)n^{-1/2}\sum_{i=1}^{n}\boldsymbol{U}_{ij}(\boldsymbol{\gamma}^*_j) + \boldsymbol{Q}_j(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}, \Lambda^*_{ij})n^{-1/2}\sum_{i=1}^{n}\int_0^{L_i}\frac{dM^*_{ij}(u)}{s^{(0)}(u;\boldsymbol{\gamma}^*_j)} + o_p(1),$$

where

$$\boldsymbol{P}_j(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij}) = n^{-1}\sum_{i=1}^{n}\frac{D_{ij}R_i\boldsymbol{X}_i\boldsymbol{K}^T_{ij}(L_i;\boldsymbol{\gamma}^*_j)\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}^*_j)\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda^*_{ij}(L_i)\}},$$
$$\boldsymbol{Q}_j(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij}) = n^{-1}\sum_{i=1}^{n}\frac{D_{ij}R_i\boldsymbol{X}_i\exp(\boldsymbol{W}^T_i\boldsymbol{\gamma}^*_j)\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}^*_j)\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda^*_{ij}(L_i)\}}.$$

B.1.5 Asymptotic Distribution of $\hat{\mu}_j$

For $j = 1, \cdots, J$, through a Taylor series expansion of $\hat{\mu}_j = n^{-1}\sum_{i=1}^{n} m_{ij}(\hat{\boldsymbol{\beta}}_j)$ about $\boldsymbol{\beta}^*_j$,

$$n^{1/2}(\hat{\mu}_j - \mu^0_j) = n^{-1/2}\sum_{i=1}^{n}\left\{m_{ij}(\boldsymbol{\beta}^*_j) - \mu_j + \boldsymbol{A}_j(\boldsymbol{\beta}^*_j)(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*_j)\right\} + o_p(1),$$

where μ^0_j is the underlying truth, and

$$\boldsymbol{A}_j(\boldsymbol{\beta}^*_j) = n^{-1}\sum_{i=1}^{n}\boldsymbol{X}^T_i m_{ij}(\boldsymbol{\beta}^*_j)\{1 - m_{ij}(\boldsymbol{\beta}^*_j)\}.$$

Combining the above results, we can represent $n^{1/2}(\hat{\mu}_j - \mu^0_j)$ as $n^{-1/2}\sum_{i=1}^{n}\psi_{ij} + o_p(1)$, where

$$\begin{aligned}\psi_{ij} =& m_{ij}(\boldsymbol{\beta}^*_j) - \mu_j + \boldsymbol{A}_j(\boldsymbol{\beta}^*_j)\boldsymbol{B}_j^{-1}(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\frac{D_{ij}R_i\boldsymbol{X}_i\{\tilde{Y}_i - m_{ij}(\boldsymbol{\beta}^*_j)\}}{\pi_{ij}(\boldsymbol{\alpha}^*)\exp\{-\Lambda^*_{ij}(L_i)\}}\\ &+ \boldsymbol{A}_j(\boldsymbol{\beta}^*_j)\boldsymbol{B}_j^{-1}(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\sum_{l=1}^{J-1}\boldsymbol{F}_{jl}\left(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij}\right)\boldsymbol{H}_l^{-1}(\boldsymbol{\alpha}^*)\boldsymbol{V}_i\{D_{il} - \pi_{il}(\boldsymbol{\alpha}^*)\}\\ &+ \boldsymbol{A}_j(\boldsymbol{\beta}^*_j)\boldsymbol{B}_j^{-1}(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\boldsymbol{P}_j(\boldsymbol{\beta}_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\boldsymbol{\Omega}_j^{-1}(\boldsymbol{\gamma}^*_j)\boldsymbol{U}_{ij}(\boldsymbol{\gamma}^*_j)\\ &+ \boldsymbol{A}_j(\boldsymbol{\beta}^*_j)\boldsymbol{B}_j^{-1}(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\boldsymbol{Q}_j(\boldsymbol{\beta}^*_j, \boldsymbol{\alpha}^*, \Lambda^*_{ij})\int_0^{L_i}\frac{dM^*_{ij}(u)}{s^{(0)}(u;\boldsymbol{\gamma}^*_j)},\end{aligned}$$

which is commonly referred to as the ith influence function of $\hat{\mu}_j$.

B.2 Supplemental Tables and Figures

	Group 1	Group 2	Group 3
Outcome $(\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j})$			
Weak	(135, 5, 4, 4)	(130, 5, 5, 4)	(125, 5, 6, -5)
Strong	(135, 10, 8, 8)	(130, 10, 10, 8)	(125, 10, 12, -10)
Censoring $(\gamma_{0j}, \gamma_{1j}, \gamma_{2j}, \gamma_{5j})$			
20%	(-31.1, 0.3, 0.3, 0.3)	(-30.8, -0.2, -0.2, -0.3)	(-30.6, 0.4, 0.3, -0.3)
30%	(-30.55, 0.3, 0.3, 0.3)	(-30.3, -0.2, -0.2, -0.3)	(-30.1, 0.4, 0.3, -0.3)
40%	(-30.2, 0.3, 0.3, 0.3)	(-29.9, -0.2, -0.2, -0.3)	(-29.7, 0.4, 0.3, -0.3)
Treatment $(\alpha_{0j}, \alpha_{1j}, \alpha_{2j}, \alpha_{4j})$			
	(0, 0, 0, 0)	(0.1, -0.2, -0.2, -0.2)	(-0.08, -0.3, -0.3, 0.2)

Table B.1: Parameter configurations for Setting I of the simulation studies

	Bias×1000			Empirical SD×1000			RMSE×1000			Coverage		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	65	99	35	37	34	36	74	105	50	46	11	74
IPW (c)	10	15	6	35	35	34	37	38	35	94	92	94
IPW (m)	36	59	23	36	35	35	50	68	42	83	62	90
Pseudo-IPW (c)	-1	0	0	35	35	35	35	35	35	95	95	95
Pseudo-IPW (m)	22	37	16	35	35	36	41	51	39	89	80	92
CAIPW-Wang (c,c)	1	2	1	36	37	36	36	37	36	95	94	94
CAIPW-Wang (c,m)	1	2	1	36	37	36	36	37	36	95	94	95
CAIPW-Wang (m,c)	1	3	2	35	37	37	35	37	37	95	94	94
CIPW (c)	-1	0	0	34	35	35	34	35	35	94	95	95
CIPW (m)	22	38	16	35	35	36	41	51	39	90	81	92
CIPW-ZS (c)	-1	-1	0	34	35	35	34	35	35	94	95	95
CIPW-ZS (m)	21	37	16	35	35	36	41	51	39	90	81	92
CAIPW-ZS (c,c)	-1	-1	0	33	34	34	33	34	34	94	94	94
CAIPW-ZS (c,m)	-1	-1	0	33	35	34	33	35	34	94	95	94
CAIPW-ZS (m,c)	2	3	1	33	34	34	33	34	34	94	94	94
CIPWR (c,c)	-1	0	1	33	34	34	33	34	34	95	96	96
CIPWR (c,m)	-1	0	1	34	35	34	34	35	35	95	96	96
CIPWR (m,c)	-1	0	1	33	34	34	33	34	34	95	95	96

Table B.2: Simulation results for the scenario of random censoring and weak outcome-covariate associations ($n = 1500$) in Setting I. For Pseudo-IPW, (c) denotes a correctly specified propensity model and (m) denotes a misspecified propensity model. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c$*$ to denote the case where the true predictors for the outcome were included in the model. Abbreviations: RMSE, root mean squared error; SD, standard deviation.

	Bias			Empirical SD			RMSE			Coverage		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	37	100	63	33	32	33	50	105	71	76	10	44
IPW (c)	12	23	10	32	33	32	34	40	34	93	89	94
IPW (m)	23	63	40	33	33	33	40	71	52	89	54	78
Pseudo (c)	-19	9	28	32	33	33	37	34	43	92	94	86
Pseudo (m)	4	45	41	32	33	34	32	56	53	95	73	77
CAIPW-Wang (c,c)	23	15	-9	32	34	34	39	37	35	88	92	94
CAIPW-Wang (c,m)	23	15	-9	32	34	34	39	37	35	89	91	94
CAIPW-Wang (m,c)	4	20	16	32	34	34	32	39	37	94	90	92
CIPW (c)	-1	1	1	32	34	34	32	34	34	95	95	94
CIPW (m)	13	43	30	32	33	34	35	54	46	93	75	86
CIPW-ZS (c)	-1	0	0	32	33	34	32	33	34	95	95	94
CIPW-ZS (m)	12	42	30	32	33	34	34	54	45	93	75	85
CAIPW-ZS (c,c)	-1	0	1	31	33	33	31	33	33	94	94	94
CAIPW-ZS (c,m)	-1	0	1	31	33	33	31	33	33	94	94	94
CAIPW-ZS (m,c)	2	4	2	31	32	32	31	33	32	95	94	94
CIPWR (c,c)	-1	0	1	31	33	33	31	33	33	96	95	95
CIPWR (c,m)	-1	1	1	31	33	33	31	33	33	96	96	95
CIPWR (m,c)	2	1	-1	31	33	33	31	33	33	95	95	95

Table B.3: Simulation results for the scenario with 20% censoring and weak outcome-covariate associations ($n = 1500$). For Pseudo, (c) denotes a correctly specified propensity model and (m) denotes a misspecified propensity model. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c$*$ to denote the case where the true predictors for the outcome were included in the model. Abbreviations: RMSE, root mean squared error; SD, standard deviation.

	Bias			Empirical SD			RMSE			Coverage		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	28	108	80	36	34	36	46	113	87	80	7	26
IPW (c)	20	34	13	34	36	35	40	49	37	92	84	93
IPW (m)	23	74	50	35	35	35	42	81	61	91	46	70
Pseudo (c)	-31	13	44	34	36	36	46	38	56	87	92	76
Pseudo (m)	-8	48	56	34	35	36	35	60	67	95	73	65
CAIPW-Wang (c,c)	35	24	-12	35	37	38	49	44	39	82	89	93
CAIPW-Wang (c,m)	36	24	-12	35	37	38	50	44	39	82	89	93
CAIPW-Wang (m,c)	5	30	25	35	37	37	35	48	45	95	87	90
CIPW (c)	-1	2	3	34	38	39	34	39	39	95	94	95
CIPW (m)	6	45	39	34	37	38	35	58	54	95	77	81
CIPW-ZS (c)	-1	0	1	34	37	37	34	37	37	95	94	95
CIPW-ZS (m)	6	44	39	34	36	37	35	57	54	94	76	81
CAIPW-ZS (c,c)	-1	0	1	33	36	36	33	36	36	95	94	94
CAIPW-ZS (c,m)	-1	0	1	33	36	37	33	36	37	95	95	94
CAIPW-ZS (m,c)	1	2	2	33	35	35	33	35	35	95	94	94
CIPWR (c,c)	-1	1	1	33	36	37	33	36	37	96	95	95
CIPWR (c,m)	0	2	2	33	37	37	33	37	37	96	95	96
CIPWR (m,c)	4	3	-1	33	36	37	33	36	37	95	95	95

Table B.4: Simulation results for the scenario with 30% censoring and weak outcome-covariate associations ($n = 1500$). For Pseudo, (c) denotes a correctly specified propensity model and (m) denotes a misspecified propensity model. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c$*$ to denote the case where the true predictors for the outcome were included in the model. Abbreviations: RMSE, root mean squared error; SD, standard deviation.

	Bias			Empirical SD			RMSE			Coverage		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	23	119	96	40	36	37	46	125	103	82	4	15
IPW (c)	31	48	17	37	38	37	48	61	40	87	76	92
IPW (m)	26	87	61	38	37	37	46	95	72	90	36	62
Pseudo (c)	-44	18	61	37	39	39	57	43	73	79	91	65
Pseudo (m)	-20	53	73	37	38	39	42	65	83	92	73	53
CAIPW-Wang (c,c)	47	33	-14	38	42	44	60	53	46	78	87	93
CAIPW-Wang (c,m)	47	33	-14	38	42	44	61	53	46	78	86	93
CAIPW-Wang (m,c)	6	41	36	38	41	43	39	58	56	94	83	87
CIPW (c)	0	4	4	36	47	48	36	47	48	94	94	94
CIPW (m)	0	49	49	37	42	44	37	65	66	95	77	77
CIPW-ZS (c)	-1	1	2	36	42	43	36	42	43	94	95	94
CIPW-ZS (m)	-1	48	49	37	40	42	37	63	64	95	77	77
CAIPW-ZS (c,c)	-1	0	0	35	40	42	35	40	42	94	95	95
CAIPW-ZS (c,m)	-1	0	1	35	41	42	35	41	42	94	95	95
CAIPW-ZS (m,c)	0	1	0	35	39	40	35	39	40	94	95	94
CIPWR (c,c)	0	2	2	36	41	43	36	41	43	95	95	95
CIPWR (c,m)	0	5	5	36	42	44	36	43	44	95	95	94
CIPWR (m,c)	7	6	-1	35	41	42	36	41	42	95	95	95

Table B.5: Simulation results for the scenario with 40% censoring and weak outcome-covariate associations ($n = 1500$). For Pseudo, (c) denotes a correctly specified propensity model and (m) denotes a misspecified propensity model. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c$*$ to denote the case where the true predictors for the outcome were included in the model. Abbreviations: RMSE, root mean squared error; SD, standard deviation.

	Bias			Empirical SD			RMSE			Coverage		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	37	100	63	33	32	33	50	105	71	76	10	44
IPW (c)	12	23	10	32	33	32	34	40	34	93	89	94
IPW (m)	23	63	40	33	33	33	40	71	52	89	54	78
Pseudo (c)	-19	9	28	32	33	33	37	34	43	92	94	86
Pseudo (m)	4	45	41	32	33	34	32	56	53	95	73	77
CAIPW-Wang (c,c)	23	15	-9	32	34	34	39	37	35	88	92	94
CAIPW-Wang (c,m)	23	15	-9	32	34	34	39	37	35	89	91	94
CAIPW-Wang (m,c)	4	20	16	32	34	34	32	39	37	94	90	92
CIPW (c)	-1	1	1	32	34	34	32	34	34	95	95	94
CIPW (m)	13	43	30	32	33	34	35	54	46	93	75	86
CIPW-ZS (c)	-1	0	0	32	33	34	32	33	34	95	95	94
CIPW-ZS (m)	12	42	30	32	33	34	34	54	45	93	75	85
CAIPW-ZS (c,c)	-1	0	1	31	33	33	31	33	33	94	94	94
CAIPW-ZS (c,m)	-1	0	1	31	33	33	31	33	33	94	94	94
CAIPW-ZS (m,c)	2	4	2	31	32	32	31	33	32	95	94	94
CIPWR (c,c)	-1	0	1	31	33	33	31	33	33	96	95	95
CIPWR (c,m)	-1	1	1	31	33	33	31	33	33	96	96	95
CIPWR (m,c)	2	1	-1	31	33	33	31	33	33	95	95	95

Table B.6: Simulation results for the scenario of 30% censoring and strong outcome-covariate associations ($n = 1500$). For Pseudo, (c) denotes a correctly specified propensity model and (m) denotes a misspecified propensity model. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c$*$ to denote the case where the true predictors for the outcome were included in the model. Abbreviations: RMSE, root mean squared error; SD, standard deviation.

	Bias			Empirical SD			RMSE			Coverage		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Scenario 1												
Naive	-12	-66	-54	34	36	38	36	75	66	88	40	56
IPW	-12	-67	-55	34	36	38	36	76	67	93	52	70
Pseudo-IPW	-17	-55	-38	36	38	39	40	67	54	92	65	84
CAIPW-Wang	-9	-38	-28	32	33	34	33	50	44	92	77	86
CIPW	0	-1	0	37	37	38	37	37	38	95	94	95
CIPW-ZS	-1	-2	-1	36	37	38	36	37	38	95	94	94
CAIPW-ZS	-1	-2	-1	33	33	34	33	33	34	94	94	95
CIPWR	1	1	0	33	32	34	33	32	34	94	94	94
Scenario 2												
Naive	-31	-29	3	33	33	34	46	44	34	80	83	93
IPW	-23	-24	-1	32	32	32	40	40	32	88	89	94
Pseudo-IPW	-7	-12	-4	32	32	32	33	34	32	94	94	94
CAIPW-Wang	-5	-4	2	28	28	27	28	28	27	93	95	95
CIPW	-1	-1	0	32	32	32	32	32	32	94	95	94
CIPW-ZS	-1	-2	0	32	32	31	32	32	31	94	95	94
CAIPW-ZS	-1	-1	0	29	29	28	29	29	28	98	98	98
CIPWR	-1	-1	0	28	28	28	28	28	28	94	95	94

Table B.7: Simulation results for the setting of crossed hazard functions (Setting II). In this setting, the models for treatment and censoring were correctly specified. The outcome model was always misspecified. Abbreviations: RMSE, root mean squared error; SD, standard deviation.

	ER visits ($N = 7678$)			All-cause hospitalization ($N = 7709$)		
	Overall	Due to treatment switch	Due to dropout	Overall	Due to treatment switch	Due to dropout
180 days	1595 (20.8%)	716 (44.9%)	879 (55.1%)	1879 (24.6%)	836 (44.1%)	1061 (55.9%)
270 days	2107 (27.4%)	955 (45.3%)	1152 (54.7%)	2585 (33.5%)	1145 (44.3%)	1440 (55.7%)
360 days	2503 (32.6%)	1136 (45.4%)	1367 (54.6%)	3129 (40.6%)	1382 (44.2%)	1747 (55.8%)

Table B.8: Number (%) of patients who were censored by a given time point.

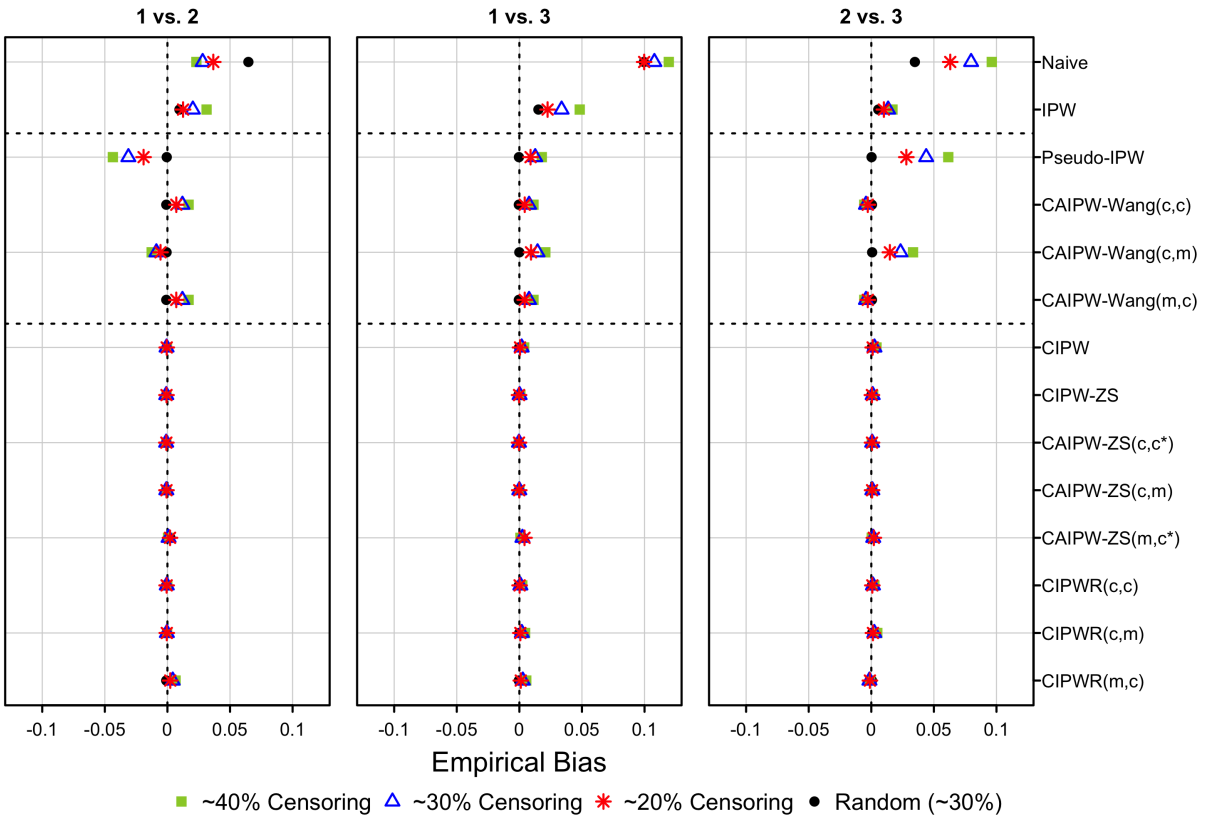

Figure B.1: Empirical bias for different proportions of censoring. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c* to denote the case where the true predictors for the outcome were included in the model. Propensity model is correctly specified for IPW, Pseudo, CIPW, and CIPW-ZS. Sample size was 1500. Results were obtained using 2000 simulated datasets. Sample size was 1500. Results were obtained using 2000 simulated datasets.

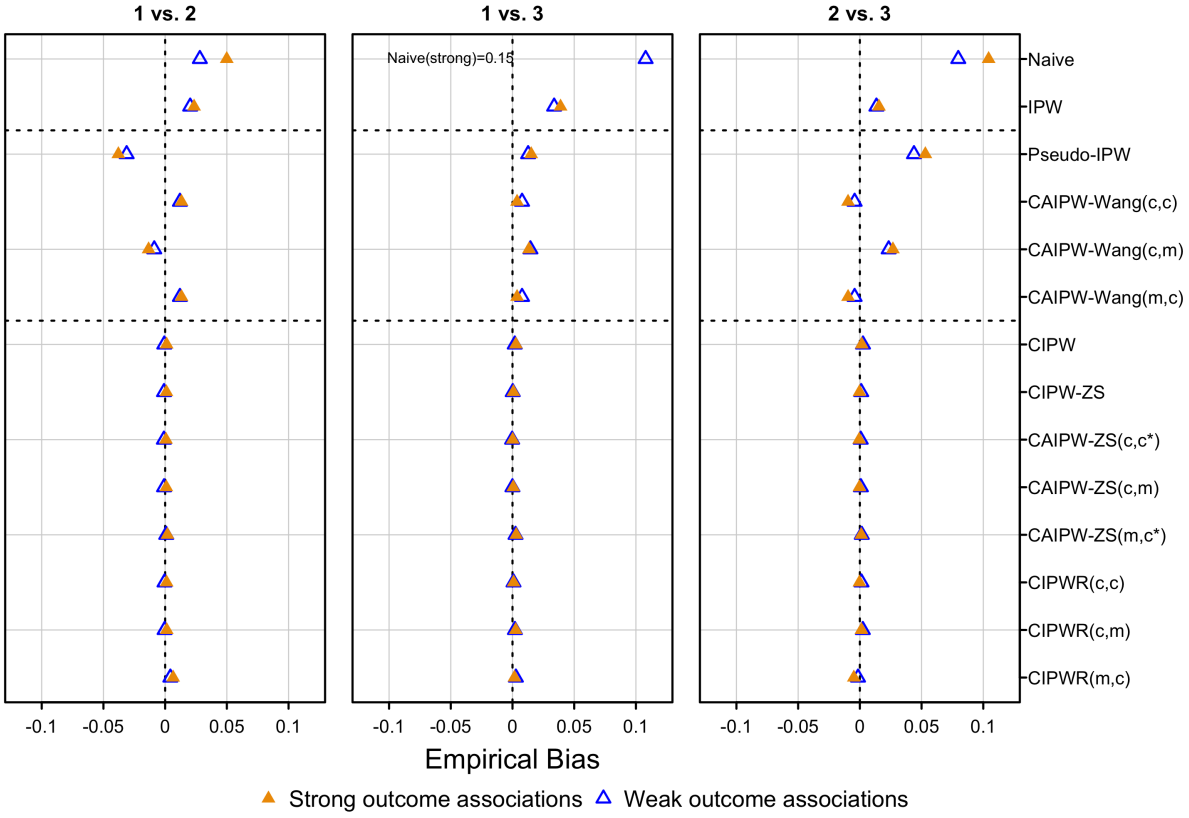

Figure B.2: Empirical bias for different levels of outcome-covariate associations. Censoring depended on covariates and the proportion of censoring at $d = 130$ was 30%. For CAIPW-Wang, the first letter and second letter denote the specification of the propensity and outcome model, respectively. For CIPWR and CAIPW-ZS, the first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. The outcome model in CAIPW-ZS is always misspecified, and we use c* to denote the case where the true predictors for the outcome were included in the model. Propensity model is correctly specified for IPW, Pseudo, CIPW, and CIPW-ZS. Sample size was 1500. Results were obtained using 2000 simulated datasets. Sample size was 1500. Results were obtained using 2000 simulated datasets.

First-line therapy	ER visits ($N = 7678$)			Hospitalization ($N = 7709$)		
	Within 180 days					
	Number of patients	Number (%) of uncensored patients	At least one ER visit (%)*	Number of patients	Number (%) of uncensored patients	At least one hospitalization record (%)*
Docetaxel	2311	1877 (81.2)	1006 (53.6)	2320	1797 (77.5)	738 (41.1)
Abiraterone	2757	2261 (82.0)	923 (40.8)	2766	2177 (78.7)	583 (26.8)
Enzalutamide	2043	1586 (77.6)	632 (39.8)	2051	1503 (73.3)	344 (22.9)
Sipuleucel-T	567	359 (63.3)	150 (41.8)	572	335 (58.6)	84 (25.1)
	Within 270 days					
Docetaxel	2311	1751 (75.8)	1132 (64.6)	2320	1623 (70.0)	849 (52.3)
Abiraterone	2757	2070 (75.1)	1091 (52.7)	2766	1926 (69.6)	703 (36.5)
Enzalutamide	2043	1442 (70.6)	740 (51.3)	2051	1304 (63.6)	407 (31.2)
Sipuleucel-T	567	308 (54.3)	168 (54.5)	572	271 (47.4)	96 (35.4)
	Within 360 days					
Docetaxel	2311	1651 (71.4)	1186 (71.8)	2320	1489 (64.2)	896 (60.2)
Abiraterone	2757	1924 (69.8)	1230 (63.9)	2766	1731 (62.6)	797 (46.0)
Enzalutamide	2043	1320 (64.6)	810 (61.3)	2051	1129 (55.0)	462 (40.9)
Sipuleucel-T	567	280 (49.4)	186 (66.4)	572	231 (40.4)	106 (45.9)

*Percentage was calculated using the number of uncensored patients as the denominator.

Table B.9: Crude risks of emergency room (ER) visits and hospitalization ignoring censored patients.

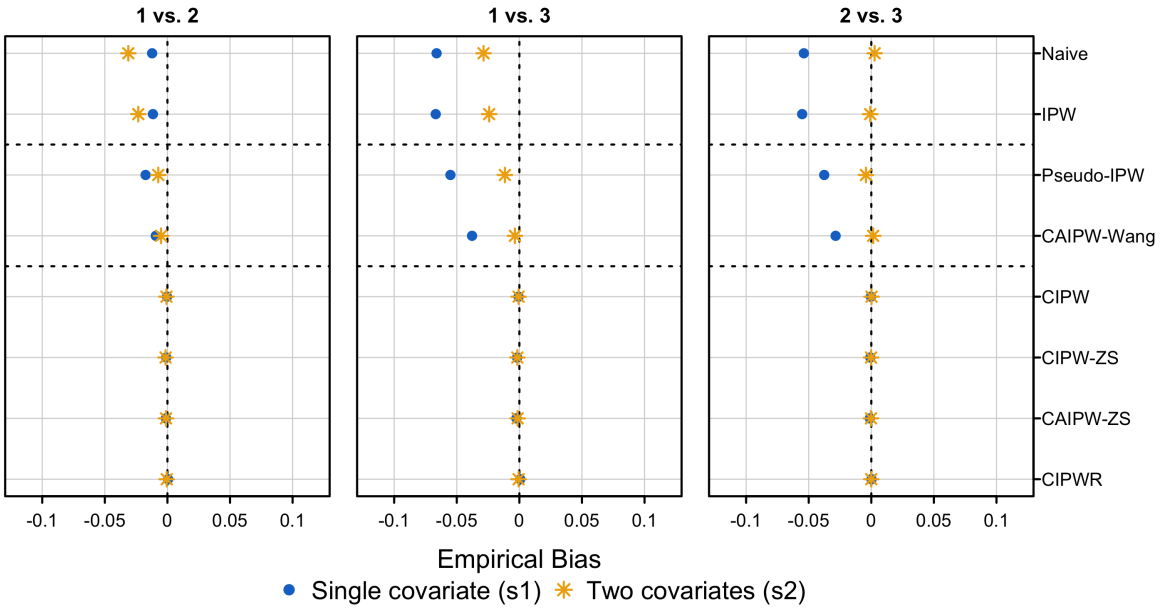

Figure B.3: Empirical bias in the setting of nonproportional hazards (Setting II). The models for treatment assignment and censoring were correctly specified. The logistic regression model and Cox model for the outcome were always misspecified in this setting. Numbers that fall outside the range of x-axis are labeled in the figure. Sample size was 1500. Results were obtained using 2000 simulated datasets.

Variable	Total (N=7678) Count (%)	Docetaxel (N=2311) Count (%)	Abiraterone (N=2757) Count (%)	Enzalutamide (N=2043) Count (%)	Sipuleucel-T (N=567) Count (%)
Age					
<65	1252 (16.3)	618 (26.7)	404 (14.7)	148 (7.2)	82 (14.5)
65-74	2549 (33.2)	926 (40.1)	835 (30.3)	597 (29.2)	191 (33.7)
≥75	3877 (50.5)	767 (33.2)	1518 (55.1)	1298 (63.5)	294 (51.9)
Race					
White	5593 (72.8)	1783 (77.2)	1975 (71.6)	1403 (68.7)	432 (76.2)
Black	1151 (15.0)	294 (12.7)	416 (15.1)	353 (17.3)	88 (15.5)
Other	934 (12.2)	234 (10.1)	366 (13.3)	287 (14.0)	47 (8.3)
Education level					
High School Diploma or Less	2165 (28.2)	650 (28.1)	769 (27.9)	613 (30.0)	133 (23.5)
High School Graduate and Less than Bachelor Degree	4196 (54.6)	1260 (54.5)	1484 (53.8)	1130 (55.3)	322 (56.8)
Bachelor Degree Plus	1317 (17.2)	401 (17.4)	504 (18.3)	300 (14.7)	112 (19.8)
Household income range					
<50k	2443 (31.8)	687 (29.7)	856 (31.0)	746 (36.5)	154 (27.2)
50k-100k	3122 (40.7)	929 (40.2)	1133 (41.1)	830 (40.6)	230 (40.6)
>100k	2113 (27.5)	695 (30.1)	768 (27.9)	467 (22.9)	183 (32.3)
Geographic Region					
South Atlantic	1917 (25.0)	551 (23.8)	681 (24.7)	548 (26.8)	137 (24.2)
New England	333 (4.4)	108 (4.7)	134 (4.9)	82 (4.0)	9 (1.6)
Middle Atlantic	668 (8.7)	194 (8.4)	235 (8.5)	181 (8.9)	58 (10.2)
East North Central	1242 (16.2)	382 (16.5)	455 (16.5)	307 (15.0)	98 (17.3)
East South Central	278 (3.6)	104 (4.5)	83 (3.0)	62 (3.0)	29 (5.1)
West North Central	626 (8.2)	360 (15.6)	133 (4.8)	86 (4.2)	47 (8.3)
West South Central	781 (10.2)	254 (11.0)	269 (9.8)	194 (9.5)	64 (11.3)
Mountain	791 (10.3)	206 (8.9)	267 (9.7)	238 (11.6)	80 (14.1)
Pacific	1042 (13.6)	152 (6.6)	500 (18.1)	345 (16.9)	45 (7.9)
Product					
HMO	10440(13.5)	324 (14.0)	378 (13.7)	309 (14.8)	36 (6.3)
PPO	541 (7.0)	163 (7.1)	205 (7.4)	138 (6.8)	35 (6.2)
Other	6097 (79.4)	1824 (78.9)	2174 (78.9)	1603 (78.5)	496 (87.5)
Metastatic (Yes)	2963 (38.6)	1116 (48.3)	1018 (36.9)	587 (28.7)	242 (42.7)
ASO (Yes)	979 (12.8)	381 (16.5)	364 (13.2)	164 (8.0)	70 (12.3)
Year of First Prescription					
2014	969 (12.6)	319 (13.8)	405 (14.7)	165 (8.1)	80 (14.1)
2015	1000 (13.0)	378 (16.4)	313 (11.4)	233 (11.4)	76 (13.4)
2016	1117 (14.5)	419 (18.1)	317 (11.5)	290 (14.2)	91 (16.0)
2017	1461 (19.0)	425 (18.4)	612 (22.2)	318 (15.6)	106 (18.7)
2018	1762 (22.9)	374 (16.2)	805 (29.2)	464 (22.7)	119 (21.0)
2019	1369 (17.8)	396 (17.1)	305 (11.1)	573 (28.0)	95 (16.8)
Diabetes	2248 (29.3)	579 (25.1)	770 (27.9)	740 (36.2)	159 (28.0)
Hypertension	5490 (71.5)	1573 (68.1)	1948 (70.7)	1557 (76.2)	412 (72.7)
Arrhythmia	1754 (22.8)	452 (19.6)	652 (23.6)	545 (26.7)	105 (18.5)
CHF	908 (11.8)	182 (7.9)	334 (12.1)	346 (16.9)	46 (8.1)
Osteoporosis	393 (5.1)	63 (2.7)	144 (5.2)	129 (6.3)	57 (10.1)
Provider Type					
Medical oncologist	4707 (61.3)	1389 (60.1)	2017 (73.2)	1177 (57.6)	124 (21.9)
Others	2971 (38.7)	922 (39.9)	740 (26.8)	866 (42.4)	443 (78.1)

Table B.10: Characteristics of patients in the four treatment groups of interest. Abbreviations: ASO, Administrative Service Only; REF, reference group; HR, hazard ratio; HMO, Health Maintenance Organization; PPO, Preferred Provider Organization; CHF, congestive heart failure.

Variable	Count (%)					
	180 days		270 days		360 days	
	Uncensored (N=6083)	Censored (N=1595)	Uncensored (N=5571)	Censored (N=2107)	Uncensored (N=5175)	Censored (N=2503)
Treatment						
Docetaxel	1877 (30.9)	434 (27.2)	1751 (31.4)	560 (26.6)	1651 (31.9)	660 (26.4)
Abiraterone	2261 (37.2)	496 (31.1)	2070 (37.2)	687 (32.6)	1924 (37.2)	833 (33.3)
Enzalutamide	1586 (26.1)	457 (28.7)	1442 (25.9)	601 (28.5)	1320 (25.5)	723 (28.9)
Sipuleucel-T	359 (5.9)	208 (13.0)	308 (5.5)	259 (12.3)	280 (5.4)	287 (11.5)
Age						
<65	922 (15.2)	330 (20.7)	824 (14.8)	428 (20.3)	740 (14.3)	512 (20.5)
65-74	2011 (33.1)	538 (33.7)	1834 (32.9)	715 (33.9)	1699 (32.8)	850 (34.0)
≥ 75	3150 (51.8)	727 (45.6)	2913 (52.3)	964 (45.8)	2736 (52.9)	1141 (45.6)
Race						
White	4398 (72.3)	1195 (74.9)	4012 (72.0)	1581 (75.0)	3723 (71.9)	1870 (74.7)
Black	950 (15.6)	201 (12.6)	873 (15.7)	278 (13.2)	810 (15.7)	341 (13.6)
Other	735 (12.1)	199 (12.5)	686 (12.3)	248 (11.8)	642 (12.4)	292 (11.7)
Education level						
High School Diploma or Less	1741 (28.6)	424 (26.6)	1605 (28.8)	560 (26.6)	1519 (29.4)	646 (25.8)
High School Graduate and Less than Bachelor Degree	3316 (54.5)	880 (55.2)	3048 (54.7)	1148 (54.5)	2808 (54.3)	1388 (55.5)
Bachelor Degree Plus	1026 (16.9)	291 (18.2)	918 (16.5)	399 (18.9)	848 (16.4)	469 (18.7)
Household income range						
$<$50k	1981 (32.6)	462 (29.0)	1838 (33.0)	605 (28.7)	1732 (33.5)	711 (28.4)
50k-100k	2476 (40.7)	646 (40.5)	2271 (40.8)	851 (40.4)	2108 (40.7)	1014 (40.5)
$>$100k	1626 (26.7)	487 (30.5)	1462 (26.2)	651 (30.9)	1335 (25.8)	778 (31.1)
Geographic Region						
South Atlantic	1551 (25.5)	366 (22.9)	1433 (25.7)	484 (23.0)	1324 (25.6)	593 (23.7)
New England	263 (4.3)	70 (4.4)	245 (4.4)	88 (4.2)	222 (4.3)	111 (4.4)
Middle Atlantic	524 (8.6)	144 (9.0)	479 (8.6)	189 (9.0)	444 (8.6)	224 (8.9)
East North Central	991 (16.3)	251 (15.7)	901 (16.2)	341 (16.2)	846 (16.3)	396 (15.8)
East South Central	209 (3.4)	69 (4.3)	184 (3.3)	94 (4.5)	170 (3.3)	108 (4.3)
West North Central	503 (8.3)	123 (7.7)	467 (8.4)	159 (7.5)	438 (8.5)	188 (7.5)
West South Central	588 (9.7)	193 (12.1)	535 (9.6)	246 (11.7)	496 (9.6)	285 (11.4)
Mountain	615 (10.1)	176 (11.0)	561 (10.1)	230 (10.9)	517 (10.0)	274 (10.9)
Pacific	839 (13.8)	203 (12.7)	766 (13.7)	276 (13.7)	718 (13.9)	324 (12.9)
Product						
HMO	820 (13.5)	220 (13.8)	755 (13.6)	285 (13.5)	701 (13.5)	339 (13.5)
PPO	449 (7.4)	92 (5.8)	418 (7.5)	123 (5.8)	392 (7.6)	149 (6.0)
Other	4814 (79.1)	1283 (80.4)	4398 (78.9)	1699 (80.6)	4082 (78.9)	2015 (80.5)
Metastatic (Yes)	2332 (38.3)	631 (39.6)	2139 (38.4)	824 (39.1)	2005 (38.7)	958 (38.3)
ASO (Yes)	747 (12.3)	232 (14.5)	666 (12.0)	313 (14.9)	602 (11.6)	377 (15.1)
Year of First Prescription						
2014	767 (12.6)	202 (12.7)	703 (12.6)	266 (12.6)	658 (12.7)	311 (12.4)
2015	818 (13.4)	182 (11.4)	766 (13.7)	234 (11.1)	726 (14.0)	274 (10.9)
2016	935 (15.4)	182 (11.4)	871 (15.6)	246 (11.7)	831 (16.1)	286 (11.4)
2017	1222 (20.1)	239 (15.0)	1139 (20.4)	322 (15.3)	1082 (20.9)	379 (15.1)
2018	1532 (25.2)	230 (14.4)	1447 (26.0)	315 (15.0)	1376 (26.6)	386 (15.4)
2019	809 (13.3)	560 (35.1)	645 (11.6)	724 (34.4)	502 (9.7)	867 (34.6)
Diabetes	1823 (30.0)	425 (26.6)	1689 (30.3)	559 (26.5)	1586 (30.6)	662 (26.4)
Hypertension	4394 (72.2)	1096 (68.7)	4042 (72.6)	1448 (68.7)	3781 (73.1)	1709 (68.3)
Arrhythmia	1441 (23.7)	313 (19.6)	1346 (24.2)	408 (19.4)	1283 (24.8)	471 (18.8)
CHF	754 (12.4)	154 (9.7)	712 (12.8)	196 (9.3)	684 (13.2)	224 (8.9)
Osteoporosis	307 (5.0)	86 (5.4)	280 (5.0)	113 (5.4)	271 (5.2)	122 (4.9)
Provider Type						
Medical oncologist	3788 (62.3)	919 (57.6)	3472 (62.3)	1235 (57.6)	3227 (62.4)	1480 (59.1)
Others	2295 (37.7)	676 (42.4)	2099 (37.7)	872 (41.4)	1948 (37.6)	1023 (40.9)

Table B.11: Characteristics of patients who were censored vs. who were not censored within different time windows for ER visits. Abbreviations: ASO, Administrative Service Only; REF, reference group; HR, hazard ratio; HMO, Health Maintenance Organization; PPO, Preferred Provider Organization; CHF, congestive heart failure.

	Docetaxel		Abiraterone		Enzalutamide		Sipuleucel-T	
Variable	Log HR	p-value	Log HR	p-value	Log HR	p-value	Log HR	p-value
Age (REF: <65)								
65-74	-0.29	<0.01	-0.19	0.03	-0.36	<0.01	0.01	0.96
≥75	-0.39	<0.01	-0.21	0.02	-0.49	<0.01	-0.23	0.23
Race (REF: White)								
Black	0.18	0.10	-0.11	0.24	-0.29	<0.01	0.17	0.30
Other	0.04	0.74	-0.21	0.04	0.02	0.88	-0.42	0.05
Education level (REF: High School Diploma or Less)								
High School Graduate and Less than Bachelor Degree	0.11	0.22	0.04	0.64	0.09	0.27	-0.05	0.74
Bachelor Degree Plus	0.02	0.86	-0.02	0.86	0.19	0.12	0.31	0.13
Household income range								
50k-100k	-0.12	0.19	0.03	0.68	-0.04	0.58	-0.04	0.78
>100k	-0.23	0.03	0.05	0.60	-0.05	0.65	-0.08	0.64
Geographic Region (REF: South Atlantic)								
New England	-0.05	0.78	0.36	0.05	-0.09	0.68	0.00	1.00
Middle Atlantic	0.13	0.33	0.22	0.06	0.08	0.52	-0.15	0.50
East North Central	0.12	0.28	0.07	0.48	0.17	0.12	0.3	0.10
East South Central	0.13	0.39	0.23	0.17	0.17	0.39	-0.17	0.58
West North Central	-0.26	0.04	0.07	0.61	-0.31	0.1	0.26	0.33
West South Central	0.00	0.98	0.01	0.93	0.26	0.03	0.42	0.03
Mountain	-0.14	0.32	-0.02	0.85	0.07	0.55	0.41	0.04
Pacific	-0.06	0.69	-0.01	0.95	-0.19	0.14	0.17	0.47
Product (REF: HMO)								
PPO	-0.24	0.18	-0.20	0.22	-0.04	0.84	0.72	0.04
Other	-0.22	0.05	-0.18	0.11	-0.13	0.29	0.55	0.04
Metastatic (Yes)	0.19	<0.01	0.03	0.67	0.27	<0.01	0.13	0.35
ASO (Yes)	-0.03	0.78	0.20	0.02	0.01	0.97	0.15	0.42
Year of First Prescription (REF: 2014)								
2015	-0.03	0.82	-0.30	0.01	0.26	0.12	-0.39	0.09
2016	-0.01	0.92	-0.26	0.03	-0.07	0.67	-0.24	0.27
2017	0.20	0.12	-0.15	0.16	0.26	0.1	0.33	0.13
2018	0.60	<0.01	0.12	0.23	0.58	<0.01	0.14	0.49
2019	1.72	<0.01	1.27	<0.01	2.10	<0.01	0.8	<0.01
Diabetes	-0.06	0.50	-0.04	0.62	-0.01	0.90	-0.16	0.23
Hypertension	-0.04	0.57	0.02	0.79	0.01	0.88	0.18	0.17
Arrhythmia	0.10	0.32	-0.06	0.50	-0.12	0.17	0.02	0.91
CHF	0.07	0.65	0.24	0.05	-0.06	0.60	-0.13	0.66
Osteoporosis	0.07	0.74	-0.14	0.34	0.05	0.70	-0.24	0.25
Provider Type (REF: Medical oncologist)								
Others	-0.06	0.46	0.04	0.50	0.11	0.11	-0.24	0.10

Table B.12: Treatment-specific log hazard ratios and associated p-values for each covariate from Cox proportional hazard models on censoring time. Abbreviations: ASO, Administrative Service Only; REF, reference group; HR, hazard ratio; HMO, Health Maintenance Organization; PPO, Preferred Provider Organization; CHF, congestive heart failure.

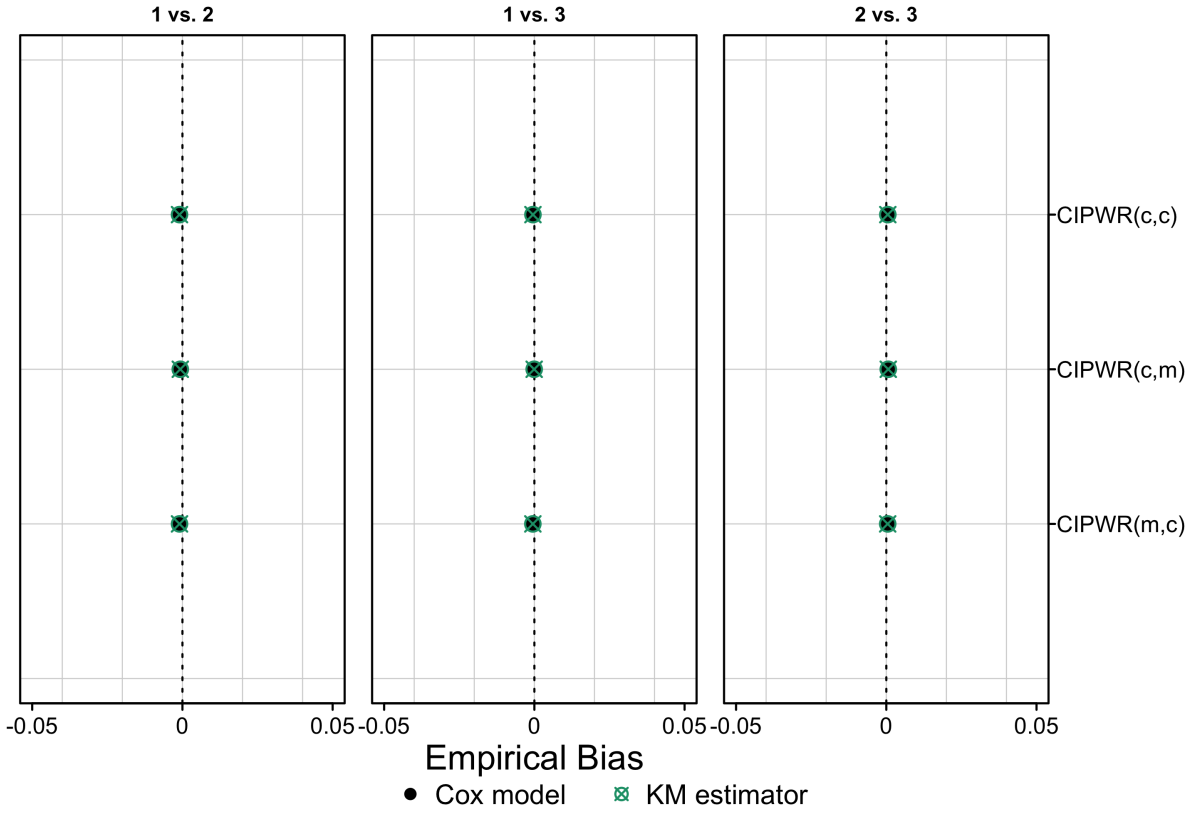

Figure B.4: Empirical bias for CIPWR using Cox model or Kaplan-Meier estimator for estimating censoring probability. The first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. Sample size was 1500. Results were obtained using 2000 simulated datasets.

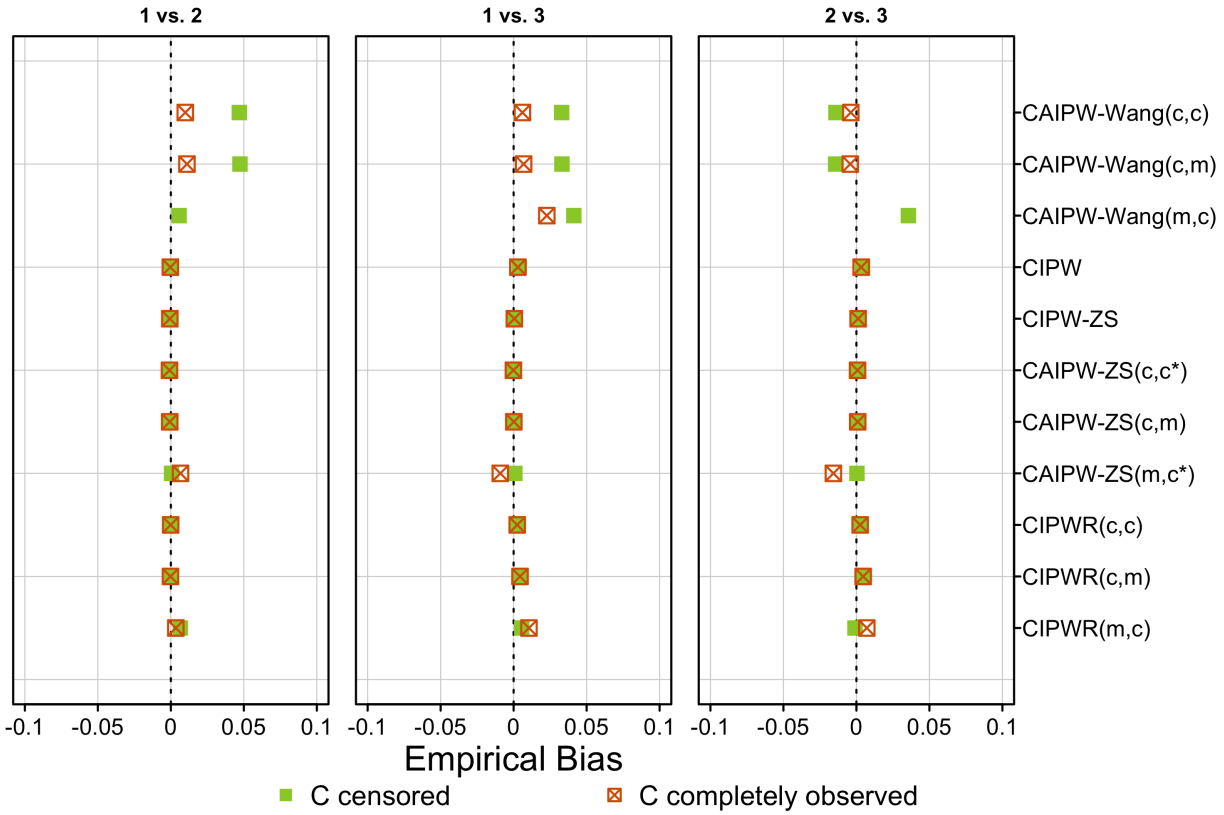

Figure B.5: Empirical bias for Cox model-based CIPWR using observed censoring time or observation time. The first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. Sample size was 1500. Results were obtained using 2000 simulated datasets.

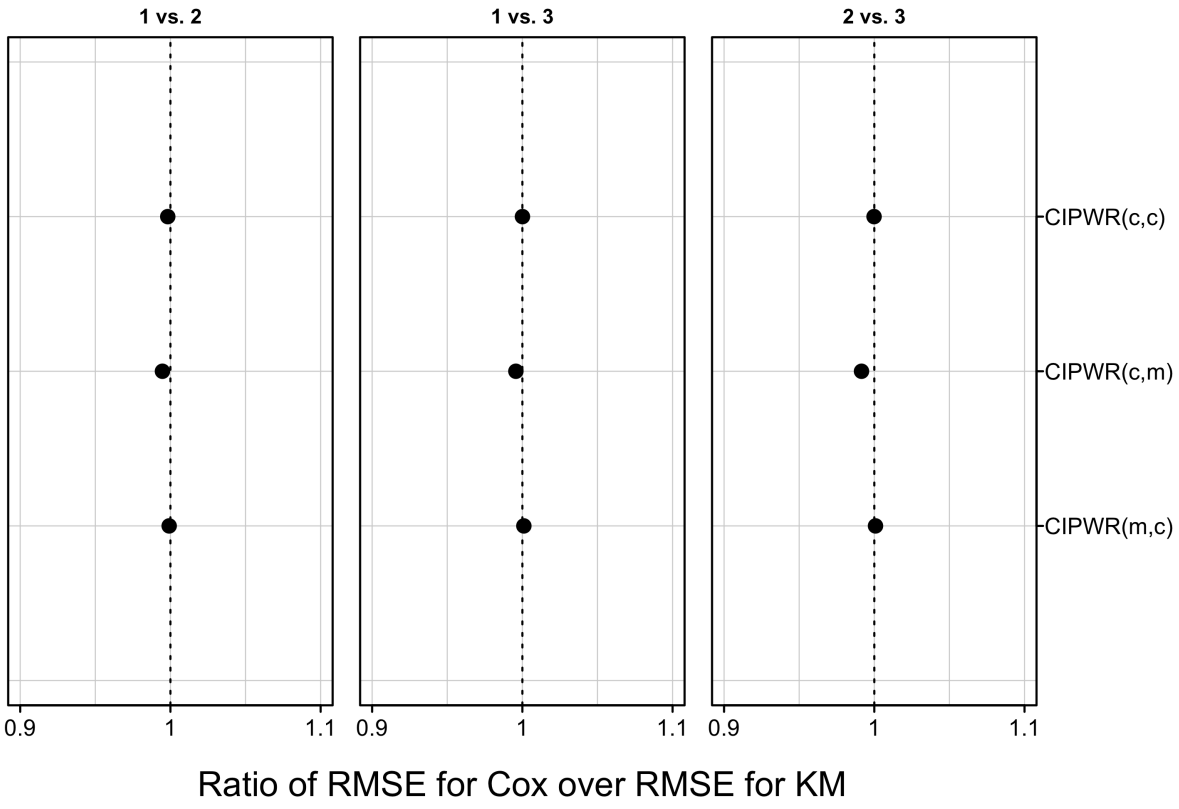

Figure B.6: RMSE for CIPWR using Cox model over RMSE for CIPWR using Kaplan-Meier estimator for estimating censoring probability. The first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. Sample size was 1500. Results were obtained using 2000 simulated datasets.

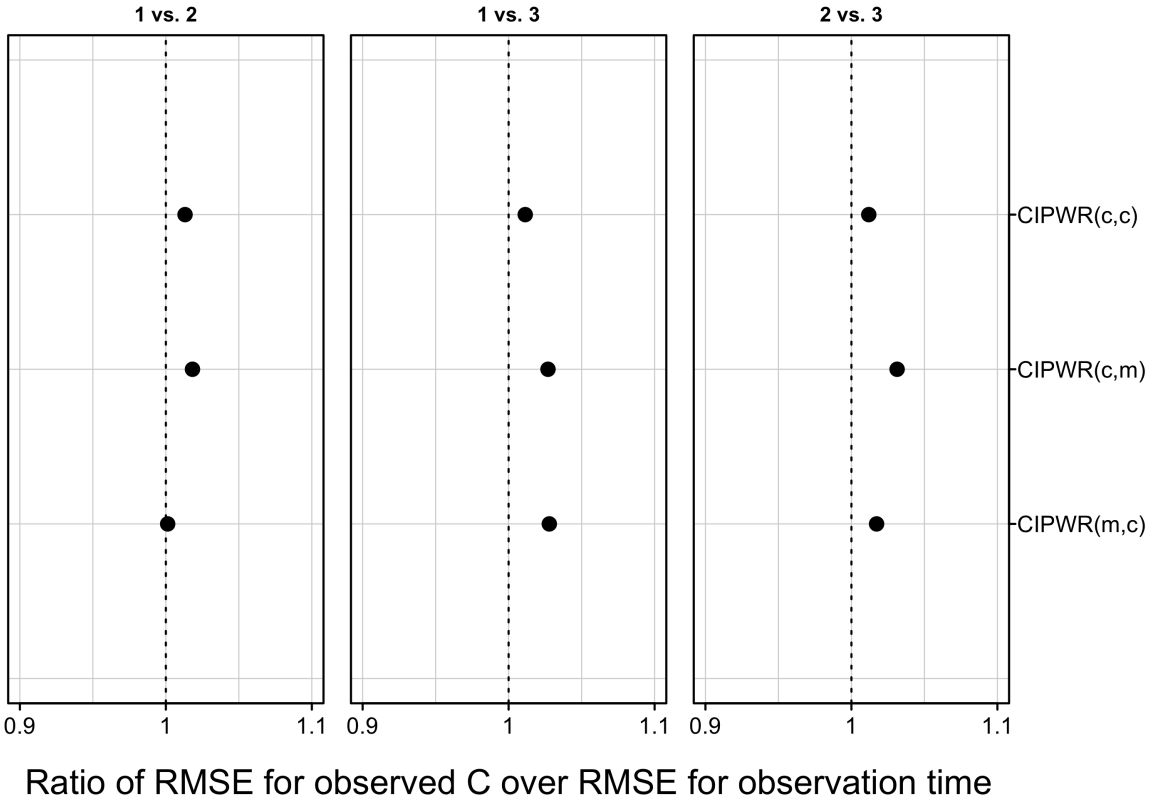

Figure B.7: RMSE for CIPWR using observed censoring time over RMSE for CIPWR using observation time for estimating censoring probability. The first and second letter in the parentheses correspond to the model for coarsening mechanism and outcome, respectively. Sample size was 1500. Results were obtained using 2000 simulated datasets.

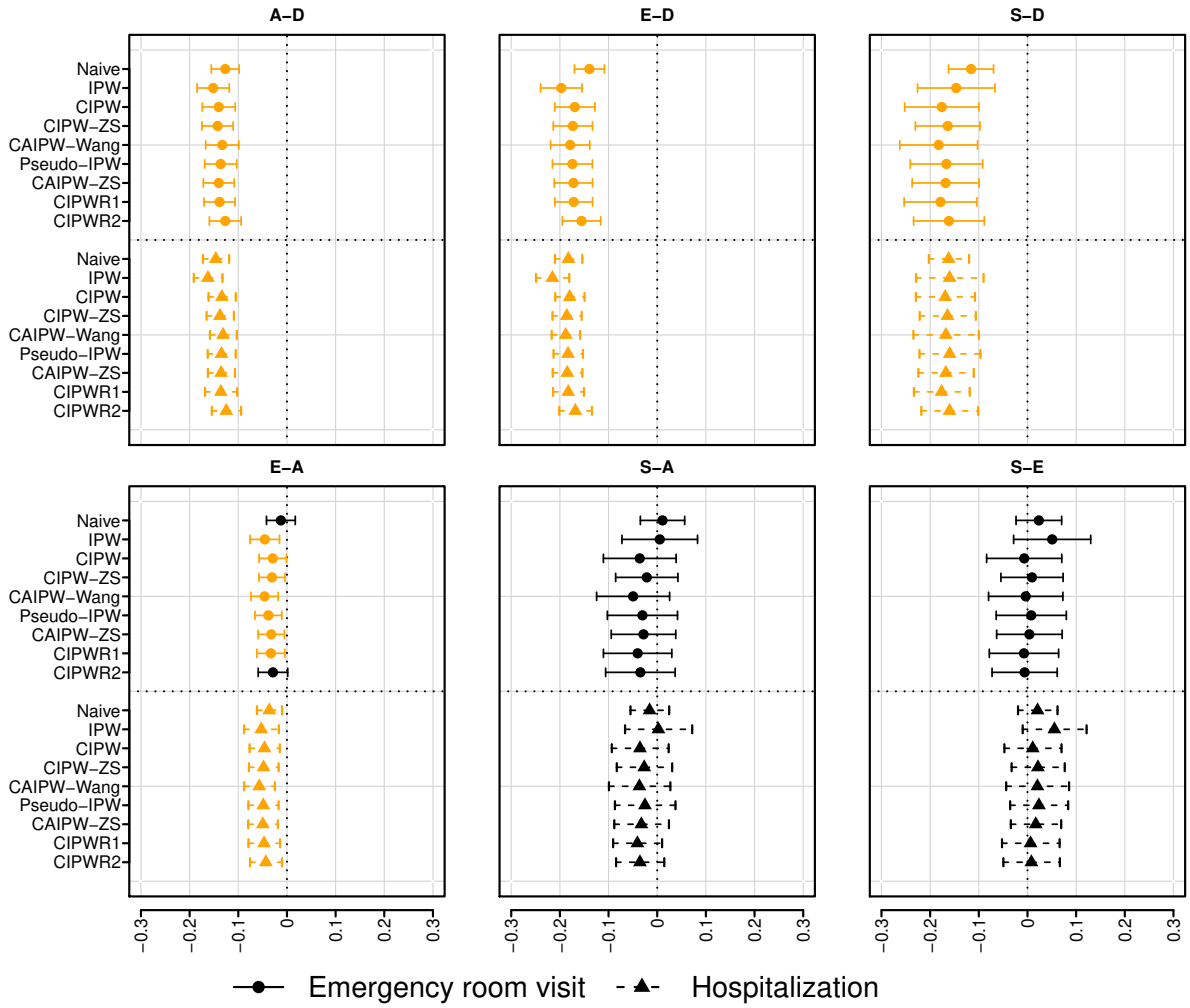

Figure B.8: Average treatment effects for ER visits and hospitalization within 180 days of treatment initiation. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7003$ ($N_A = 2458$, $N_D = 2162$, $N_E = 1833$, $N_S = 550$) for ER visits, and $N = 7045$ ($N_A = 2474$, $N_D = 2172$, $N_E = 1843$, $N_S = 556$) for hospitalization. CIPWR1 is based on observation time, and CIPWR2 is based on observed censoring time. Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T; ER, emergency room visit.

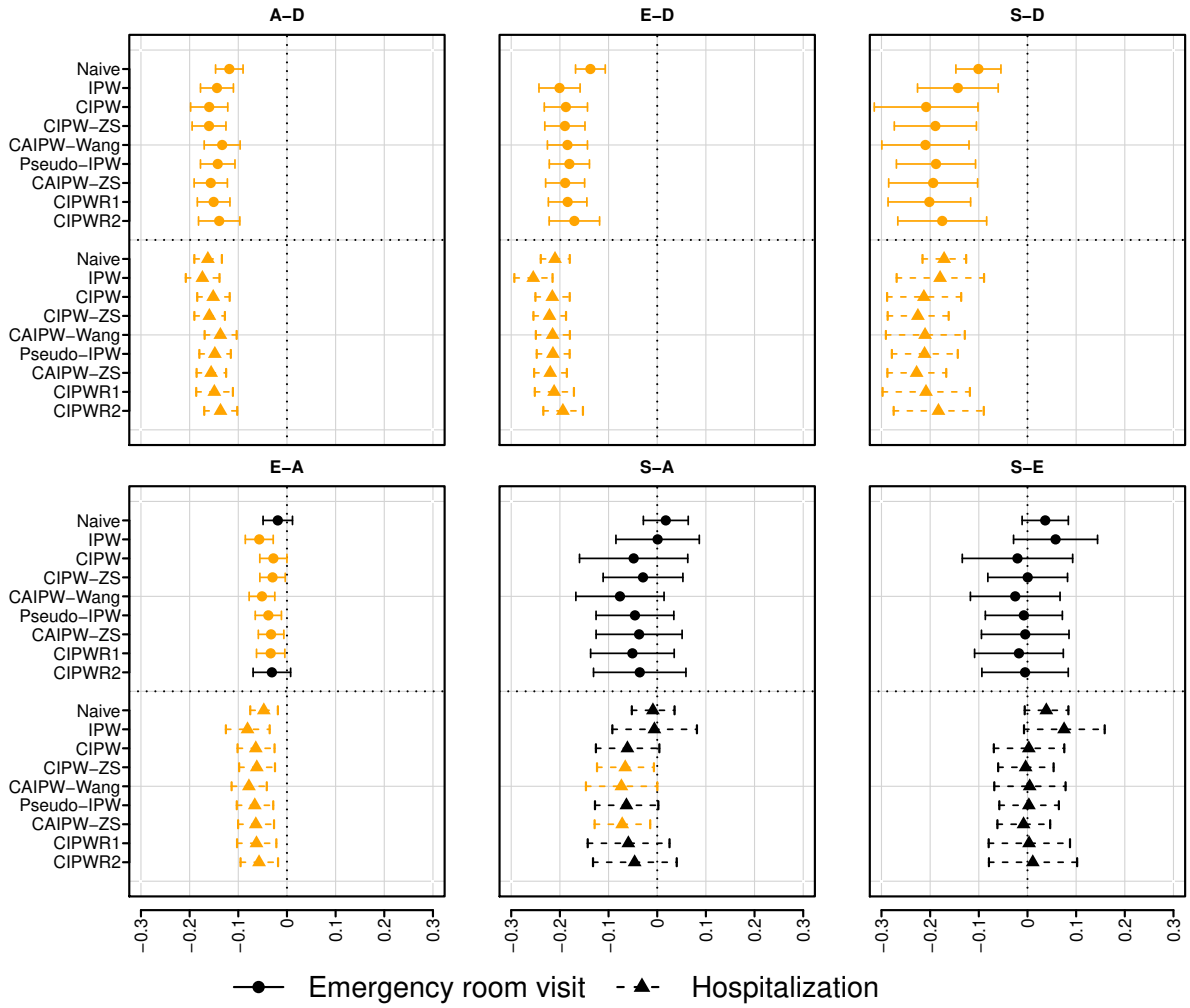

Figure B.9: Average treatment effects for ER visits and hospitalization within 270 days of treatment initiation. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7003$ ($N_A = 2458$, $N_D = 2162$, $N_E = 1833$, $N_S = 550$) for ER visits, and $N = 7045$ ($N_A = 2474$, $N_D = 2172$, $N_E = 1843$, $N_S = 556$) for hospitalization. CIPWR1 is based on observation time, and CIPWR2 is based on observed censoring time. Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T; ER, emergency room visit.

B.3 Generation of Survival Time with Crossing Hazards

We assumed a three-phase model for the time to event T, and the cumulative hazard function is specified as follows:

$$\begin{aligned}\Lambda(t|X_1, X_2) =& \exp(\beta_1 X_1 + \beta_2 X_2) I(t \le a) \\ &+ \Big\{ \frac{\exp(\beta_1 X_1 + \beta_2 X_2) b - \exp(\alpha_1 X_1 + \alpha_2 X_2) a}{b - a} \\ &- \frac{\exp(\beta_1 X_1 + \beta_2 X_2) - \exp(\alpha_1 X_1 + \alpha_2 X_2)}{b - a} t \Big\} I(a < t \le b) \\ &+ \exp(\alpha_1 X_1 + \alpha_2 X_2) I(t > b).\end{aligned}$$

In the first scenario, $b = 0.25, a = 0.2, \beta_1 = 2, \alpha_1 = -2, \beta_2 = 0, \alpha_2 = 0$. In the sceond scenario, $b = 0.25, a = 0.2, \beta_1 = 2, \alpha_1 = 0, \beta_2 = -2, \alpha_2 = -1$. In Sections B.3.1 and B.3.2, we list the equations used to generate the event times.

B.3.1 Scenario 1

Define

$$\begin{aligned}\text{termA} &= \{\exp(2X_1) - \exp(-2X_1)\}/2 \\ \text{termB} &= -\{0.25 \exp(2X_1) - 0.2 \exp(-2X_1)\} \\ \text{termC} &= -0.05 \log u + 0.02\{\exp(2X_1) - \exp(-2X_1)\}\end{aligned}$$

Then let

$$\begin{aligned}\mathcal{I}_1 &= \frac{-\log u}{\exp(2X_1)} \\ \mathcal{I}_2 &= \frac{-\text{termB} + \sqrt{\text{termB}^2 - 4\text{termA} \times \text{termC}}}{2\text{termA}} \\ \mathcal{I}_3 &= \frac{-\log u - 0.225\{\exp(2X_1) - \exp(-2X_1)\}}{\exp(-2X_1)}\end{aligned}$$

$$T' = \mathcal{I}_1 I(\mathcal{I}_1 \le 0.2) + \mathcal{I}_2 I(\mathcal{I}_2 \le 0.25) I(\mathcal{I}_2 > 0.2) + \mathcal{I}_3 I(\mathcal{I}_3 > 0.25)$$

The final event time was obtained using

$$T = T' I(Z = 1) + (T' + 0.1) I(Z = 2) + (T' + 0.2) I(Z = 3)$$

B.3.2 Scenario 2

Define

$$\begin{aligned}
\text{termA} &= \{\exp(2X_1) - \exp(-2X_1 - X_2)\}/2 \\
\text{termB} &= -\{0.25\exp(2X_1) - 0.2\exp(-2X_1 - X_2)\} \\
\text{termC} &= -0.05\log u + 0.02\{\exp(2X_1) - \exp(-2X_1 - X_2)\}
\end{aligned}$$

Then let

$$\begin{aligned}
\mathcal{I}_1 &= \frac{-\log u}{\exp(2X_1)} \\
\mathcal{I}_2 &= \frac{-\text{termB} + \sqrt{\text{termB}^2 - 4\text{termA} \times \text{termC}}}{2\text{termA}} \\
\mathcal{I}_3 &= \frac{-\log u - 0.225\{\exp(2X_1) - \exp(-2X_1 - X_2)\}}{\exp(-2X_1 - X_2)}
\end{aligned}$$

$$T' = \mathcal{I}_1 I(\mathcal{I}_1 \leq 0.2) + \mathcal{I}_2 I(\mathcal{I}_2 \leq 0.25) I(\mathcal{I}_2 > 0.2) + \mathcal{I}_3 I(\mathcal{I}_3 > 0.25)$$

The final event time was obtained using

$$T = T' I(Z = 1) + (T' + 0.1) I(Z = 2) + (T' + 0.15) I(Z = 3)$$

APPENDIX C

Supplement for Chapter III

C.1 Supplementary Tables

Scenario	Predictors for Treatment Generating Model
Linear main effects only (L)*	$(X_{\mathcal{C}1}, X_{\mathcal{C}2}, X_{\mathcal{C}3}, X_{\mathcal{C}4}, X_{\mathcal{C}5}, X_{\mathcal{Z}1}, X_{\mathcal{Z}2}, X_{\mathcal{Z}3}, X_{\mathcal{Z}4}, X_{\mathcal{Z}5})$
Nonlinear main effects only (NL)	$(X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0), \exp\{X_{\mathcal{C}2}\}^{-1/2}, \lvert X_{\mathcal{C}3}\rvert, \log\lvert X_{\mathcal{C}4} + 1\rvert, \lvert X_{\mathcal{C}5}\rvert^{-1/2},$ $X_{\mathcal{Z}1}, X_{\mathcal{Z}2}, X_{\mathcal{Z}3}, X_{\mathcal{Z}4}^2, \log\lvert X_{\mathcal{Z}5}\rvert)$
Linear main effects and linear interactions (L-L)	$(X_{\mathcal{C}1}, X_{\mathcal{C}2}, X_{\mathcal{C}3}, X_{\mathcal{C}4}, X_{\mathcal{C}5},$ $X_{\mathcal{Z}1}, X_{\mathcal{Z}2}, X_{\mathcal{Z}3}, X_{\mathcal{Z}4}, X_{\mathcal{Z}5},$ $X_{\mathcal{Z}1} \times X_{\mathcal{Z}2}, X_{\mathcal{Z}1} \times X_{\mathcal{Z}3}, X_{\mathcal{Z}1} \times X_{\mathcal{Z}4}, X_{\mathcal{Z}1} \times X_{\mathcal{Z}5},$ $X_{\mathcal{Z}2} \times X_{\mathcal{Z}3}, X_{\mathcal{Z}2} \times X_{\mathcal{Z}4}, X_{\mathcal{Z}2} \times X_{\mathcal{Z}5}, X_{\mathcal{Z}3} \times X_{\mathcal{Z}4},$ $X_{\mathcal{Z}3} \times X_{\mathcal{Z}5}, X_{\mathcal{Z}4} \times X_{\mathcal{Z}5}, X_{\mathcal{C}3} \times X_{\mathcal{Z}1})$
Non-linear main effects and linear interactions (NL-L)	$(X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0), \exp\{X_{\mathcal{C}2}\}^{-1/2}, \lvert X_{\mathcal{C}3}\rvert, \log\lvert X_{\mathcal{C}4} + 1\rvert, \lvert X_{\mathcal{C}5}\rvert^{-1/2},$ $X_{\mathcal{Z}1}, X_{\mathcal{Z}2}, X_{\mathcal{Z}3}, X_{\mathcal{Z}4}^2, \log\lvert X_{\mathcal{Z}5}\rvert,$ $X_{\mathcal{Z}1} \times X_{\mathcal{Z}2}, X_{\mathcal{Z}1} \times X_{\mathcal{Z}3}, X_{\mathcal{Z}1} \times X_{\mathcal{Z}4}, X_{\mathcal{Z}1} \times X_{\mathcal{Z}5},$ $X_{\mathcal{Z}2} \times X_{\mathcal{Z}3}, X_{\mathcal{Z}2} \times X_{\mathcal{Z}4}, X_{\mathcal{Z}2} \times X_{\mathcal{Z}5}, X_{\mathcal{Z}3} \times X_{\mathcal{Z}4},$ $X_{\mathcal{Z}3} \times X_{\mathcal{Z}5}, X_{\mathcal{Z}4} \times X_{\mathcal{Z}5}, X_{\mathcal{C}3} \times X_{\mathcal{Z}1})$
Non-linear main effects and non-linear interactions (NL-NL)	$(X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0), \sqrt{\exp\{X_{\mathcal{C}2}\}}, \lvert X_{\mathcal{C}3}\rvert, \log\lvert X_{\mathcal{C}4} + 1\rvert, \sqrt{\lvert X_{\mathcal{C}5}\rvert},$ $X_{\mathcal{Z}1}, X_{\mathcal{Z}2}, X_{\mathcal{Z}3}, X_{\mathcal{Z}4}^2, \log\lvert X_{\mathcal{Z}5}\rvert,$ $X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0) \times \sqrt{\exp\{X_{\mathcal{C}2}\}}, (X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0) \times \lvert X_{\mathcal{C}3}\rvert,$ $X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0) \times \log\lvert X_{\mathcal{C}4} + 1\rvert, X_{\mathcal{C}1}(X_{\mathcal{C}1} > 0) \times \sqrt{\lvert X_{\mathcal{C}5}\rvert},$ $\sqrt{\exp\{X_{\mathcal{C}2}\}} \times \lvert X_{\mathcal{C}3}\rvert, \sqrt{\exp\{X_{\mathcal{C}2}\}} \times \log\lvert X_{\mathcal{C}4} + 1\rvert,$ $\sqrt{\exp\{X_{\mathcal{C}2}\}} \times \sqrt{\lvert X_{\mathcal{C}5}\rvert}, \lvert X_{\mathcal{C}3}\rvert \times \log\lvert X_{\mathcal{C}4} + 1\rvert,$ $\lvert X_{\mathcal{C}3}\rvert \times \sqrt{\lvert X_{\mathcal{C}5}\rvert}, \log\lvert X_{\mathcal{C}4} + 1\rvert \times \sqrt{\lvert X_{\mathcal{C}5}\rvert}$ $\lvert X_{\mathcal{C}3}\rvert \times X_{\mathcal{Z}1}$

* The baseline scenario.

Table C.1: Design matrix for the treatment generating models of various degrees of nonlinearity and/or nonadditivity.

	Empirical SD			SE (usual)			Coverage (%; usual)			SE (modified)			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
OAL	37	34	36	39	35	37	96.0	94.9	94.5	38	34	36	94.9	94.9	94.9
LOGIS															
All	37	37	35	40	37	37	55.5	65.9	92.7	41	40	38	59.8	74.2	94.6
Ysel	38	35	36	48	41	44	98.2	97.3	97.3	39	35	37	95.4	95.0	94.9
YZsel	39	37	38	46	40	43	97.4	96.0	95.9	40	36	38	95.0	95.3	95.4
OP + All	37	35	35	40	36	38	95.3	94.1	95.8	43	40	41	96.6	97.6	97.2
OP + Ysel	38	35	36	48	41	44	98.2	97.3	97.3	39	35	37	95.4	95.0	94.9
OP + YZsel	37	35	36	45	39	42	97.7	96.8	97.0	38	34	37	94.8	95.0	94.8
CART															
All	58	58	55	69	69	65	65.8	81.6	95.7	57	58	55	51.0	70.8	91.9
Ysel	50	50	47	63	62	59	87.6	93.5	98.4	52	51	50	78.1	86.8	94.9
YZsel	48	47	46	60	59	56	91.9	95.3	97.4	50	49	48	84.3	89.4	95.1
OP + All	55	52	56	70	65	71	98.2	98.5	98.3	56	52	56	94.0	94.4	94.4
OP + Ysel	45	44	45	61	57	61	98.8	98.6	99.2	46	44	46	94.0	95.1	95.3
OP + YZsel	43	41	43	57	54	58	98.4	98.4	99.4	44	41	44	93.5	95.0	94.7
Pruned CART															
All	45	43	38	68	68	63	53.0	80.4	99.1	42	42	39	19.4	40.2	88.6
Ysel	44	43	38	62	61	57	78.6	88.6	98.5	42	42	39	48.4	66.0	92.8
YZsel	46	45	39	60	58	55	80.6	90.7	97.8	43	42	40	60.0	73.4	93.9
OP + All	39	36	38	68	62	68	99.9	100	99.9	40	36	39	94.0	95.0	95.9
OP + Ysel	36	35	37	58	54	59	99.9	99.3	99.5	38	34	37	94.6	94.9	95.6
OP + YZsel	37	35	37	55	52	56	99.1	99.5	99.5	38	35	37	94.8	95.4	95.2
Bagged CART															
All	38	38	36	31	31	29	10.1	28.2	79.1	40	40	38	24.4	46.9	91.1
Ysel	52	46	48	39	36	34	84.8	87.0	85.0	52	48	49	94.4	95.8	96.0
YZsel	66	58	59	46	41	39	69.8	74.8	81.0	63	57	60	85.8	86.9	95.2
OP + All	37	34	36	62	62	68	99.9	99.8	100	38	33	37	94.7	95.0	94.7
OP + Ysel	34	32	34	57	60	63	99.9	99.9	100	36	32	35	94.3	95.6	94.7
OP + YZsel	34	32	34	55	57	60	99.3	100	99.9	36	32	35	94.5	95.2	94.6
Random Forests															
All	34	35	34	28	28	26	2.5	13.2	73.1	37	38	35	7.0	27.8	88.2
Ysel	44	39	37	30	29	27	69.3	77.2	84.5	43	41	41	86.2	91.4	96.7
YZsel	86	65	63	34	32	29	48.3	57.3	64.3	62	52	58	73.2	78.6	93.5
OP + All	38	35	37	43	39	42	96.6	96.9	96.4	39	34	38	95.3	95.2	94.0
OP + Ysel	35	32	35	50	47	52	98.6	99.5	98.7	36	32	36	94.7	95.2	94.7
OP + YZsel	35	32	35	56	54	59	99.5	99.8	99.8	36	32	36	94.8	95.6	94.7

Table C.2: Standard errors (SE) and coverage of 95% confidence intervals estimated by usual bootstrap and modified bootstrap for sample size of 1000. The scenario with sparse treatment models was considered. Results were obtained based on 1000 simulated datasets. For each dataset, 200 bootstrap samples were generated.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	165	124	-41	50	52	49	172	135	64	50	52	48	9.7	34.1	85.5
OAL	5	5	0	56	50	52	57	50	52	55	50	52	94.1	95.6	94.2
LOGIS															
Confounder	-1	1	2	58	53	55	58	53	55	57	52	54	94.3	94.3	94.3
Treatment	2	3	1	73	63	65	73	63	65	69	61	62	94.1	94.2	93.4
Outcome	-1	0	2	57	49	53	57	49	53	56	50	53	94.5	96.1	94.1
All	95	73	-22	52	51	50	108	89	55	55	55	52	59.8	75.9	93.9
Ysel	-6	-3	3	57	49	54	57	49	54	58	51	55	95.4	96.0	94.8
YZsel	-6	-3	3	58	51	54	58	52	54	58	53	55	95.0	95.5	95.0
OP + All	16	13	-2	51	46	50	54	48	50	58	56	56	96.3	97.8	96.9
OP + Ysel	-6	-3	3	57	46	54	57	59	54	58	51	55	95.4	96.0	94.8
OP + YZsel	-5	-2	2	55	48	53	56	48	53	55	49	53	95.2	96.0	94.3
CART															
Confounder	55	41	-14	70	66	65	89	78	66	69	68	66	86.7	90.6	94.6
Treatment	94	71	-24	72	72	66	119	101	70	73	72	68	73.8	82.7	93.6
Outcome	68	51	-17	71	67	65	99	84	67	71	70	68	81.0	89.1	94.8
All	123	93	-30	79	81	76	146	123	81	78	80	74	63.2	75.9	92.0
Ysel	72	54	-17	72	71	66	102	89	70	72	72	69	81.6	88.1	94.0
YZsel	60	44	-15	69	66	63	91	79	66	70	68	67	86.2	90.8	94.8
OP + All	10	8	-1	80	74	79	80	76	79	79	74	79	93.4	93.8	94.2
OP + Ysel	12	9	-3	67	61	65	68	62	65	66	63	66	94.2	94.4	94.8
OP + YZsel	15	9	-6	62	58	62	64	59	62	63	59	62	94.3	94.7	94.8
Pruned CART															
Confounder	81	61	-20	64	61	55	103	86	59	61	60	57	69.4	81.4	93.3
Treatment	110	83	-27	66	66	58	128	106	64	65	64	59	58.1	72.4	92.3
Outcome	92	69	-22	63	59	54	111	91	59	60	60	56	63.1	77.0	93.0
All	133	100	-32	62	63	56	147	118	66	61	61	56	40.1	60.4	90.4
Ysel	96	72	-24	63	60	55	115	94	60	60	60	56	61.0	75.5	92.4
YZsel	84	63	-20	64	61	55	105	87	59	61	60	57	68.6	80.7	93.0
OP + All	-2	1	2	58	53	56	58	52	56	58	52	57	94.3	94.5	94.6
OP + Ysel	1	2	1	56	49	54	56	49	54	55	50	54	94.2	94.8	94.6
OP + YZsel	2	2	0	54	50	53	54	50	53	55	50	54	95.7	94.9	94.7
Bagged CART															
Confounder	-50	-37	12	93	83	86	105	91	87	88	80	84	90.1	90.0	93.8
Treatment	43	33	-10	81	76	71	92	83	72	79	74	71	89.0	92.0	93.9
Outcome	-2	0	2	76	65	70	76	65	70	75	70	71	94.2	96.4	95.0
All	119	90	-30	54	53	51	131	104	59	56	57	53	44.1	65.5	91.2
Ysel	17	14	-3	71	63	63	73	65	63	71	66	67	93.4	95.6	95.2
YZsel	-34	-27	7	91	82	81	97	85	82	86	79	81	91.6	92.5	94.8
OP + All	-4	-1	4	53	46	52	53	46	52	52	47	51	94.2	95.4	94.4
OP + Ysel	-7	-2	5	51	45	49	51	45	50	50	45	50	94.6	95.2	95.0
OP + YZsel	-7	-2	5	51	45	49	51	45	49	50	46	50	94.6	95.0	94.6
Random Forests															
Confounder	-22	-13	9	77	67	71	80	68	71	75	68	72	94.7	95.1	95.0
Treatment	62	48	-14	62	58	57	87	75	58	63	61	59	82.6	89.4	94.8
Outcome	35	28	-7	58	52	54	67	59	54	62	59	59	92.3	95.7	96.2
All	135	102	-33	49	51	48	144	114	58	52	53	50	27.1	51.2	90.0
Ysel	53	41	-11	57	52	51	77	66	53	59	57	57	84.9	91.2	94.5
YZsel	-36	-24	13	94	78	79	101	81	82	79	70	75	91.0	91.2	94.0
OP + All	-5	-1	4	55	47	53	55	47	52	53	48	52	94.8	94.9	94.0
OP + Ysel	-9	-2	6	51	45	50	52	45	50	51	46	50	94.5	94.8	94.4
OP + YZsel	-8	-2	5	50	46	50	51	45	50	51	46	50	94.8	95.0	94.7

Table C.3: Simulation results for the scenario with sparse models and sample size of 500. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	165	125	-41	36	37	33	169	130	53	36	37	34	0.5	8.3	78.4
OAL	-2	0	2	38	33	36	38	33	36	38	34	36	94.9	94.9	94.9
LOGIS															
Confounder	-3	0	2	40	36	37	40	36	37	40	36	38	94.5	95.0	95.2
Treatment	-2	0	1	50	43	43	50	43	43	49	43	43	94.2	94.7	94.4
Outcome	-3	0	2	38	33	36	38	33	36	38	34	37	94.8	95.1	94.9
All	71	55	-16	38	36	35	81	66	38	41	40	38	59.8	74.2	94.6
Ysel	-6	-3	3	38	34	36	39	34	36	39	35	37	95.4	95.0	94.9
YZsel	-5	-2	3	40	36	37	40	36	37	40	36	38	95.0	95.3	95.4
OP + All	12	11	-1	37	33	35	39	35	35	43	40	41	96.6	97.6	97.2
OP + Ysel	-6	-3	3	38	34	36	39	34	36	39	35	37	95.4	95.0	94.9
OP + YZsel	-5	-3	3	38	34	36	39	34	36	38	34	37	94.8	95.0	94.8
CART															
Confounder	44	32	-12	49	45	43	66	55	45	50	48	48	85.5	90.8	95.3
Treatment	83	63	-21	54	52	48	99	81	52	53	52	50	64.4	77.0	93.2
Outcome	56	42	-15	50	48	45	76	63	47	51	50	49	79.7	86.8	94.8
All	111	82	-29	58	58	54	125	100	62	57	58	55	51.0	70.8	91.9
Ysel	60	43	-17	51	49	46	79	66	49	52	51	50	78.1	86.8	94.9
YZsel	47	35	-13	49	47	44	69	58	46	50	49	48	84.3	89.4	95.1
OP + All	5	4	-1	56	51	56	56	51	56	56	52	56	94.0	94.4	94.4
OP + Ysel	12	7	-5	46	43	45	48	44	45	46	44	46	94.0	95.1	95.3
OP + YZsel	14	8	-6	44	40	42	46	41	43	44	41	44	93.5	95.0	94.7
Pruned CART															
Confounder	70	52	-18	49	45	37	85	69	41	43	42	40	59.0	73.6	93.8
Treatment	100	76	-24	50	46	40	112	89	47	46	45	42	42.2	58.4	91.2
Outcome	81	61	-20	47	43	37	94	74	42	42	41	39	51.1	67.7	91.9
All	125	94	-30	46	43	38	133	104	48	42	42	39	19.4	40.2	88.6
Ysel	84	63	-21	47	43	37	96	77	42	42	42	39	48.4	66.0	92.8
YZsel	71	53	-17	48	45	38	86	70	42	43	42	40	60.0	73.4	93.9
OP + All	-6	-2	3	40	35	38	40	35	38	40	36	39	94.0	95.0	95.9
OP + Ysel	-3	-2	1	38	34	36	38	34	36	38	34	37	94.6	94.9	95.6
OP + YZsel	-2	-1	0	38	33	36	38	33	36	38	35	37	94.8	95.4	95.2
Bagged CART															
Confounder	-64	-51	12	66	59	62	92	78	63	64	58	62	82.3	84.9	94.0
Treatment	27	19	-8	59	54	52	65	57	52	58	53	52	90.2	94.0	93.8
Outcome	-21	-14	7	54	47	50	58	49	50	55	50	52	94.0	95.7	95.8
All	107	80	-26	39	37	35	114	89	44	40	40	38	24.4	46.9	91.1
Ysel	-4	-3	2	53	46	46	53	46	46	52	48	49	94.4	95.8	96.0
YZsel	-53	-42	12	66	59	59	85	73	60	63	57	60	85.8	86.9	95.2
OP + All	-6	-1	5	38	33	37	38	33	37	38	33	37	94.7	95.0	94.7
OP + Ysel	-9	-3	6	35	31	35	36	31	35	36	32	35	94.3	95.6	94.7
OP + YZsel	-8	-3	6	35	31	35	36	31	35	36	32	35	94.5	95.2	94.6
Random Forests															
Confounder	-43	-31	12	59	49	53	72	58	54	56	49	53	88.3	90.2	95.3
Treatment	50	40	-10	45	41	40	67	57	41	46	43	42	79.0	86.9	95.2
Outcome	19	16	-3	41	37	37	45	40	37	45	42	43	94.1	96.2	96.8
All	125	94	-30	35	35	33	130	101	45	37	38	35	7.0	27.8	88.2
Ysel	31	25	-6	45	39	36	55	46	37	43	41	41	86.2	91.4	96.7
YZsel	-74	-52	22	87	66	64	114	83	68	62	52	58	73.2	78.6	93.5
OP + All	-6	-1	5	39	33	38	39	33	38	39	34	38	95.3	95.2	94.0
OP + Ysel	-9	-3	6	35	31	35	37	31	35	36	32	36	94.7	95.2	94.7
OP + YZsel	-9	-3	6	35	31	35	37	31	35	36	32	36	94.8	95.6	94.7

Table C.4: Simulation results for the scenario with sparse models and sample size of 1000. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	167	125	-41	26	27	24	169	128	48	25	26	24	0	0.4	59.9
OAL	-1	0	1	27	24	26	27	24	26	27	24	26	95.0	94.7	95.2
LOGIS															
Confounder	0	1	1	28	26	27	28	26	27	28	26	27	95.1	94.6	95.2
Treatment	0	0	0	35	31	31	35	31	31	34	30	30	93.8	94.3	94.8
Outcome	0	0	1	27	24	26	27	24	26	27	24	26	94.8	94.7	95.0
All	55	42	-12	28	26	26	61	50	29	30	29	28	55.7	70.4	94.8
Ysel	-3	-1	1	27	24	26	27	24	26	27	24	26	95.1	94.8	94.9
YZsel	-2	-1	1	28	26	27	28	26	27	28	26	27	95.6	95.1	95.4
OP + All	13	11	-2	28	25	26	30	27	26	31	29	29	95.1	96.5	97.4
OP + Ysel	-3	-1	1	27	24	26	27	24	26	27	24	26	95.1	94.9	94.9
OP + YZsel	-2	-1	1	27	24	26	27	24	26	27	24	26	95.2	94.8	95.0
CART															
Confounder	34	26	-9	33	32	31	48	41	32	35	34	33	85.2	88.7	95.5
Treatment	74	56	-19	38	37	34	84	67	39	38	36	35	49.3	67.3	91.7
Outcome	44	32	-12	35	33	32	56	46	34	36	35	34	77.0	86.5	94.8
All	99	74	-25	42	40	38	108	84	46	41	40	38	32.0	55.2	89.6
Ysel	47	34	-13	36	33	32	59	47	35	37	36	35	74.4	86.0	94.5
YZsel	37	28	-9	34	33	31	51	43	32	35	34	33	82.7	87.4	95.2
OP + All	7	4	-3	37	34	37	38	35	37	38	35	37	94.4	95.2	95.2
OP + Ysel	13	7	-6	31	29	30	34	30	31	32	30	31	94.0	94.8	94.8
OP + YZsel	14	8	-5	30	28	29	33	29	30	30	28	29	93.0	93.8	94.8
Pruned CART															
Confounder	62	47	-15	33	31	27	71	56	31	30	29	28	45.6	61.3	91.9
Treatment	94	71	-23	36	33	28	100	78	36	33	32	29	21.1	40.3	87.6
Outcome	72	55	-18	35	32	26	80	63	32	29	29	27	34.8	53.1	90.3
All	116	88	-28	34	32	27	121	93	39	30	30	27	5.7	18.8	80.1
Ysel	75	57	-18	35	32	27	83	65	33	29	29	27	31.6	50.2	89.4
YZsel	65	49	-16	35	31	27	74	58	31	30	29	28	43.9	59.0	90.4
OP + All	-3	-2	2	26	25	27	27	25	27	27	25	26	95.3	95.0	95.0
OP + Ysel	-2	-1	1	26	24	25	26	24	25	26	24	25	95.6	94.6	95.0
OP + YZsel	-1	-1	0	26	24	26	26	24	26	26	24	26	95.2	95.3	94.8
Bagged CART															
Confounder	-89	-73	16	49	43	47	102	84	50	49	43	47	54.9	59.8	93.6
Treatment	11	6	-5	45	40	40	46	40	40	45	40	39	93.3	94.7	93.9
Outcome	-48	-37	11	41	34	38	63	50	39	42	37	39	81.5	84.2	95.6
All	92	69	-23	29	27	26	96	74	35	30	29	28	12.9	33.9	87.9
Ysel	-34	-26	8	42	35	36	54	43	37	40	36	38	86.4	89.8	95.3
YZsel	-80	-64	16	51	45	45	95	78	48	48	42	46	59.4	65.1	93.8
OP + All	-3	2	5	27	24	27	27	24	28	28	24	27	95.1	94.4	94.4
OP + Ysel	-6	-1	5	25	23	25	25	23	25	25	23	25	95.2	94.6	94.9
OP + YZsel	-6	0	5	24	23	25	25	23	25	25	23	25	95.2	94.8	94.9
Random Forests															
Confounder	-70	-51	18	46	36	41	83	63	45	44	36	41	64.4	70.8	94.4
Treatment	41	32	-9	33	29	30	53	44	31	33	31	31	75.4	84.5	94.9
Outcome	3	3	0	30	26	27	30	26	27	33	30	31	96.4	96.8	97.3
All	116	88	-28	25	25	24	119	92	37	26	27	25	0.9	8.1	80.6
Ysel	0	1	1	42	35	29	42	35	29	33	30	31	87.1	91.0	96.3
YZsel	-123	-91	31	88	68	56	151	114	64	53	42	49	40.2	42.9	91.0
OP + All	-3	1	5	29	25	28	29	25	28	29	25	27	94.0	94.8	94.7
OP + Ysel	-6	-1	5	25	23	25	25	23	25	26	23	25	95.0	94.8	94.9
OP + YZsel	-6	-1	5	25	23	25	25	23	25	25	23	25	95.2	95.2	94.8

Table C.5: Simulation results for the scenario with sparse models and sample size of 2000. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	141	108	-32	49	48	44	149	119	54	48	50	43	17.7	40.6	88.4
OAL	24	19	-5	53	48	48	58	52	48	50	47	47	90.2	92.6	94.1
LOGIS															
Confounder	0	2	1	54	48	52	54	48	52	53	49	51	94.4	95.3	95.0
Treatment	1	2	2	64	54	59	64	54	59	64	56	58	95.0	95.3	94.3
Outcome	0	2	1	54	46	52	54	46	52	54	48	52	94.8	95.9	94.4
All	94	73	-20	50	47	46	106	87	50	51	51	47	54.6	70.8	93.2
Ysel	3	3	0	53	47	51	53	48	51	55	50	53	95.2	95.9	95.4
YZsel	31	24	-7	58	51	50	65	56	50	53	50	50	88.4	92.4	94.4
OP + All	21	18	-4	48	44	47	53	47	47	52	51	51	93.6	96.7	96.4
OP + Ysel	3	3	0	53	47	51	53	48	51	55	50	53	95.2	95.9	95.3
OP + YZsel	5	5	-1	52	46	49	52	46	49	50	45	49	94.0	94.8	94.4
CART															
Confounder	72	54	-18	65	63	60	97	83	63	65	65	60	79.3	86.5	94.0
Treatment	98	76	-22	71	68	63	121	102	67	68	68	61	68.2	79.7	92.2
Outcome	84	65	-19	67	65	62	108	92	65	66	66	61	73.4	84.4	92.8
All	116	89	-26	75	73	66	138	116	71	72	74	65	62.4	76.6	92.4
Ysel	86	65	-21	67	64	62	109	91	66	67	67	62	73.1	84.2	92.6
YZsel	80	61	-19	65	62	60	103	87	63	64	64	59	73.8	84.7	92.7
OP + All	19	16	-3	72	67	73	75	69	73	72	68	73	92.6	94.6	93.9
OP + Ysel	21	15	-6	64	59	63	67	61	64	63	60	64	92.7	94.2	94.0
OP + YZsel	21	16	-6	60	56	60	64	58	60	59	56	59	92.0	93.1	93.8
Pruned CART															
Confounder	91	70	-21	60	57	52	109	90	56	56	56	51	60.1	74.2	91.8
Treatment	111	86	-25	62	58	53	127	104	59	58	58	52	50.9	67.1	90.9
Outcome	101	78	-22	59	55	52	117	96	57	56	56	50	54.3	70.2	91.3
All	124	95	-29	60	58	51	137	111	59	57	58	50	40.6	62.0	90.6
Ysel	103	79	-24	58	55	51	118	96	56	56	56	50	53.1	69.1	91.0
YZsel	95	73	-22	59	54	51	112	91	56	56	56	51	58.7	74.7	91.2
OP + All	13	10	-3	56	51	55	57	52	55	53	49	53	93.0	94.3	94.0
OP + Ysel	13	11	-3	53	48	52	54	49	52	51	47	51	92.8	94.5	94.3
OP + YZsel	15	12	-3	53	48	51	55	49	51	51	47	50	93.1	94.0	94.3
Bagged CART															
Confounder	19	15	-4	69	61	64	71	63	64	65	61	62	92.6	94.7	93.7
Treatment	73	57	-17	62	56	55	96	79	57	60	58	54	74.1	85.2	92.2
Outcome	54	41	-12	57	52	53	78	67	55	58	56	54	83.0	90.6	93.7
All	114	88	-26	51	49	46	125	100	53	51	52	47	40.3	61.6	91.4
Ysel	60	45	-15	56	51	52	82	68	54	57	55	53	80.1	89.8	94.5
YZsel	30	23	-7	74	68	69	79	71	69	70	66	65	90.5	93.7	94.2
OP + All	12	10	-2	49	44	47	51	46	47	47	43	46	92.2	94.4	94.5
OP + Ysel	10	9	-1	48	44	46	49	45	46	46	42	46	93.1	94.2	94.8
OP + YZsel	10	9	-1	48	44	47	49	45	47	46	42	46	93.0	93.9	94.7
Random Forests															
Confounder	50	39	-11	55	49	50	74	63	52	55	53	52	84.2	90.6	94.7
Treatment	84	66	-19	52	48	47	99	82	51	52	52	48	63.0	77.3	93.4
Outcome	73	57	-16	50	46	46	88	73	49	52	51	48	70.3	83.7	93.7
All	119	92	-27	49	47	44	129	103	52	49	50	45	32.9	55.4	91.3
Ysel	77	60	-18	50	46	46	92	75	49	51	51	48	66.9	81.5	94.0
YZsel	35	29	-6	78	67	64	85	73	64	63	59	59	84.4	90.3	94.7
OP + All	10	9	-1	50	44	48	51	45	48	47	43	47	92.4	94.4	94.4
OP + Ysel	8	8	0	48	44	47	49	44	47	46	42	46	93.1	93.9	94.9
OP + YZsel	8	8	0	49	44	47	50	45	47	46	43	46	92.9	94.2	94.8

Table C.6: Simulation results for the scenario with moderately sparse models and sample size of 500. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	142	109	-33	34	35	30	146	114	45	34	35	31	1.5	12.2	81.0
OAL	3	2	-1	35	32	34	35	32	34	36	32	34	95.9	95.3	95.6
LOGIS															
Confounder	1	0	-1	36	33	35	36	33	35	37	33	36	95.3	94.8	95.3
Treatment	1	0	-1	43	37	39	43	37	39	43	37	39	94.8	94.3	95.2
Outcome	0	0	0	35	32	34	35	32	34	36	32	35	96.4	95.4	95.8
All	70	54	-16	34	33	32	78	64	36	37	36	34	52.8	69.6	93.7
Ysel	-3	-3	0	35	32	35	36	32	35	37	33	36	96.6	95.0	95.9
YZsel	0	0	0	36	33	35	36	33	35	37	34	36	95.6	95.2	96.0
OP + All	12	9	-3	33	31	33	35	32	33	38	36	37	96.2	96.9	96.9
OP + Ysel	-3	-3	0	35	32	35	36	32	35	37	33	36	96.6	95.0	95.9
OP + YZsel	-2	-2	0	35	32	34	35	32	34	36	32	35	95.9	95.2	95.6
CART															
Confounder	66	49	-17	47	46	42	81	67	45	46	46	43	68.8	81.2	93.7
Treatment	94	71	-23	49	49	44	106	86	50	49	49	44	51.0	68.0	91.1
Outcome	76	58	-18	47	46	43	89	74	47	48	48	45	63.8	78.6	93.0
All	111	85	-27	52	54	48	123	101	55	52	53	48	43.4	63.8	90.4
Ysel	81	61	-20	50	48	44	95	77	48	49	49	45	60.8	76.8	92.6
YZsel	71	54	-17	47	45	42	86	71	45	47	47	44	65.7	79.1	94.5
OP + All	6	4	-2	52	48	53	53	48	53	52	48	53	94.2	94.4	94.8
OP + Ysel	10	5	-4	45	42	45	46	42	46	45	42	46	94.5	95.2	94.8
OP + YZsel	12	7	-5	43	40	43	45	41	43	43	40	43	93.7	94.0	95.1
Pruned CART															
Confounder	87	66	-21	43	42	35	97	78	41	39	39	36	39.1	59.9	90.6
Treatment	106	81	-25	43	42	37	114	92	44	41	41	36	28.4	48.2	88.4
Outcome	96	73	-23	41	40	35	104	83	42	39	39	35	31.6	52.1	90.0
All	121	93	-28	40	40	34	128	101	45	39	40	35	14.1	35.0	87.0
Ysel	99	75	-24	41	40	34	107	85	42	39	39	35	28.5	49.9	89.1
YZsel	91	69	-22	42	41	35	101	80	41	40	40	36	37.4	57.6	90.6
OP + All	0	0	0	36	33	36	36	33	36	37	33	37	95.5	95.4	95.2
OP + Ysel	1	0	-1	35	33	35	35	33	35	36	33	36	95.1	94.3	95.6
OP + YZsel	3	1	-1	36	33	35	36	33	35	36	33	36	95.4	94.7	95.8
Bagged CART															
Confounder	11	7	-3	47	43	44	48	43	44	47	43	44	94.0	95.1	95.1
Treatment	68	52	-16	42	40	37	80	65	40	42	41	38	63.5	75.1	93.3
Outcome	46	34	-11	38	36	36	59	50	38	41	39	38	80.2	87.4	95.5
All	108	83	-26	35	35	31	114	90	40	36	37	33	14.2	38.2	88.6
Ysel	57	43	-14	38	36	34	68	56	37	40	38	37	68.2	80.1	94.6
YZsel	25	19	-6	46	42	41	53	46	41	45	42	42	88.6	91.9	95.7
OP + All	-1	0	1	33	31	33	33	31	33	34	30	34	95.7	94.7	95.3
OP + Ysel	-3	-1	1	32	30	32	32	30	32	33	30	33	96.1	94.9	95.2
OP + YZsel	-2	-1	1	32	30	32	32	30	32	33	30	33	95.9	94.9	95.7
Random Forests															
Confounder	44	34	-10	36	35	34	57	49	36	39	37	37	79.4	86.7	95.3
Treatment	79	61	-19	35	35	32	87	70	37	37	37	34	43.1	62.6	92.7
Outcome	66	50	-16	33	33	31	74	60	35	37	36	34	56.5	73.0	94.2
All	115	88	-27	33	34	30	120	94	40	35	35	32	7.6	29.5	87.2
Ysel	76	58	-18	34	33	31	83	67	36	36	36	34	44.4	64.7	94.0
YZsel	47	36	-11	40	37	34	62	52	35	38	37	36	74.4	82.8	95.8
OP + All	-1	0	1	34	31	34	34	31	34	35	31	34	95.8	95.0	95.0
OP + Ysel	-3	-1	2	32	30	33	33	30	33	33	30	33	96.1	94.6	95.2
OP + YZsel	-3	-1	2	33	30	33	33	30	33	33	30	33	96.0	95.0	95.8

Table C.7: Simulation results for the scenario with moderately sparse models and sample size of 1000. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	113	87	-26	44	47	38	121	99	46	44	46	38	27.8	52.5	90.2
OAL	51	37	-14	48	47	41	70	60	43	44	43	41	76.4	84.1	93.4
LOGIS															
Confounder	2	1	-1	48	45	47	48	45	47	51	46	50	95.3	95.0	95.0
Treatment	2	1	-2	57	51	54	57	51	54	65	57	61	97.1	96.9	96.9
Outcome	1	0	-1	48	45	48	48	45	48	56	50	54	97.5	96.5	97.2
All	87	66	-20	44	45	39	97	80	44	46	46	41	50.5	70.3	92.5
Ysel	26	18	-8	48	46	43	55	49	44	48	45	46	90.1	91.4	95.5
YZsel	62	46	-17	48	47	41	78	66	44	46	45	41	71.8	82.8	93.5
OP + All	42	30	-12	45	45	40	62	54	42	46	45	43	84.8	89.9	95.5
OP + Ysel	26	18	-8	48	46	43	55	49	44	48	45	46	90.1	91.4	95.5
OP + YZsel	30	20	-10	45	44	41	54	49	42	43	41	41	88.3	90.9	94.2
CART															
Confounder	73	55	-17	61	61	56	95	82	59	60	60	54	77.2	84.0	92.4
Treatment	89	67	-22	65	66	56	111	94	60	63	64	55	68.8	80.2	92.2
Outcome	81	62	-19	64	63	58	103	88	61	62	62	55	71.1	82.2	91.5
All	96	74	-23	67	69	59	118	101	63	65	67	57	67.2	78.6	92.2
Ysel	81	60	-21	60	61	54	101	86	58	60	61	54	72.3	82.2	92.8
YZsel	85	63	-22	55	57	50	101	85	55	56	57	51	68.2	80.4	92.8
OP + All	44	32	-12	69	67	65	82	75	66	64	63	62	87.2	89.8	92.9
OP + Ysel	40	28	-12	59	58	55	72	64	57	57	56	56	86.8	90.5	94.1
OP + YZsel	38	25	-14	53	53	52	66	59	53	52	51	52	87.1	91.1	92.7
Pruned CART															
Confounder	88	67	-21	53	54	45	103	86	50	50	51	45	56.6	71.8	92.2
Treatment	99	75	-24	53	53	45	112	92	51	52	53	45	50.0	68.3	91.6
Outcome	93	71	-22	52	54	44	107	89	49	51	51	45	52.2	69.8	91.8
All	104	80	-24	52	54	45	116	96	51	51	52	45	45.8	65.6	91.6
Ysel	94	71	-23	50	51	43	107	87	49	49	50	44	51.1	68.8	91.8
YZsel	92	69	-23	50	51	43	104	86	49	50	51	44	53.8	71.4	91.7
OP + All	43	31	-12	54	52	47	70	61	48	49	47	47	82.7	87.6	93.9
OP + Ysel	38	25	-12	48	48	45	61	54	47	46	44	45	86.1	90.0	93.5
OP + YZsel	36	24	-12	47	47	44	59	53	46	45	44	45	87.3	90.0	93.8
Bagged CART															
Confounder	51	38	-14	50	49	45	72	62	47	50	50	46	82.7	87.9	93.8
Treatment	80	60	-20	48	49	42	93	78	47	48	49	43	62.1	75.8	92.4
Outcome	71	53	-17	46	46	41	84	71	45	48	48	43	69.7	80.8	93.6
All	94	72	-22	45	47	40	104	86	46	46	47	41	46.8	67.3	91.8
Ysel	64	46	-18	49	49	45	81	67	48	50	50	46	75.0	85.7	93.8
YZsel	57	42	-15	68	69	63	89	81	65	65	63	58	82.1	89.1	93.5
OP + All	41	29	-12	47	46	40	63	55	42	42	41	41	80.8	86.5	94.3
OP + Ysel	36	25	-11	45	44	40	57	50	42	41	40	40	84.9	88.7	93.3
OP + YZsel	34	23	-11	44	43	40	56	49	42	41	40	41	86.0	90.6	94.3
Random Forests															
Confounder	68	52	-17	45	46	40	82	69	44	46	46	42	69.2	80.0	93.4
Treatment	86	65	-21	44	46	39	96	80	44	45	46	41	53.0	70.2	92.5
Outcome	81	61	-19	43	45	39	91	76	43	45	46	41	56.7	73.6	93.2
All	97	75	-23	44	46	38	107	88	45	45	46	40	41.0	63.2	91.6
Ysel	72	54	-19	49	49	41	87	73	45	47	47	43	65.4	78.5	93.7
YZsel	49	36	-13	81	81	65	95	88	66	61	59	54	78.9	85.1	93.8
OP + All	37	26	-11	46	46	41	59	52	42	43	41	41	85.2	88.4	93.7
OP + Ysel	35	24	-11	45	44	41	57	51	42	41	40	41	85.2	88.7	93.9
OP + YZsel	33	22	-11	46	44	43	57	50	44	42	40	41	86.7	90.6	94.8

Table C.8: Simulation results for the scenario with dense models and sample size of 500. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	111	87	-24	32	32	27	116	93	36	31	32	27	5.8	21.6	86.3
OAL	12	10	-2	32	30	31	34	31	31	32	30	31	93.2	94.1	94.8
LOGIS															
Confounder	-1	0	1	33	30	33	33	30	33	34	31	33	94.5	95.1	94.5
Treatment	-1	0	0	38	33	36	38	33	36	39	34	36	95.0	95.1	94.3
Outcome	-1	0	1	32	29	33	32	29	33	34	30	33	95.4	95.5	94.6
All	61	48	-13	31	30	29	68	57	31	33	33	30	55.4	70.0	94.1
Ysel	1	1	0	32	29	33	32	29	33	34	31	34	96.2	95.8	94.7
YZsel	13	10	-2	33	30	32	36	32	32	34	31	32	93.0	94.1	94.5
OP + All	13	10	-3	30	28	30	33	30	30	34	32	32	95.0	95.8	96.0
OP + Ysel	1	1	0	32	29	33	32	29	33	34	31	34	96.2	95.8	94.7
OP + YZsel	2	2	0	31	29	32	31	29	32	33	29	32	95.3	94.9	94.7
CART															
Confounder	68	53	-15	44	43	39	81	68	42	43	43	39	65.0	76.8	92.3
Treatment	85	67	-18	46	45	41	97	81	44	46	46	40	52.4	68.7	92.0
Outcome	76	60	-17	45	46	41	89	75	45	45	45	40	59.4	72.7	92.0
All	93	73	-19	49	48	41	105	88	46	48	48	42	51.1	67.0	92.6
Ysel	77	60	-17	45	45	41	90	75	45	45	46	40	59.0	75.1	92.2
YZsel	72	57	-16	43	43	38	84	71	41	44	44	39	61.8	74.5	93.2
OP + All	11	10	-2	47	43	46	48	44	46	46	43	47	92.5	94.0	94.6
OP + Ysel	13	9	-4	42	40	44	44	41	44	42	40	43	94.2	93.6	93.8
OP + YZsel	13	10	-3	39	38	41	42	39	41	40	38	41	93.8	94.6	94.2
Pruned CART															
Confounder	82	65	-17	39	37	31	91	75	36	36	36	31	37.9	54.8	90.6
Treatment	94	74	-20	38	38	32	102	83	38	37	37	32	29.1	47.2	90.0
Outcome	88	69	-19	38	37	32	96	78	37	36	36	31	32.0	48.6	89.3
All	100	78	-22	38	37	31	107	87	38	36	37	31	22.4	42.0	88.8
Ysel	90	70	-20	37	36	32	97	79	37	36	36	31	30.0	48.0	89.6
YZsel	85	67	-18	38	36	32	93	76	37	36	36	32	34.7	53.2	90.3
OP + All	8	7	-1	34	31	33	34	31	33	33	31	33	93.2	94.7	95.0
OP + Ysel	7	6	-1	32	30	33	32	31	33	33	30	33	94.4	94.1	94.8
OP + YZsel	8	7	-2	33	31	33	34	31	33	33	30	33	93.5	94.6	94.7
Bagged CART															
Confounder	45	35	-9	34	33	32	56	48	33	35	34	33	75.3	82.8	94.7
Treatment	75	58	-16	34	33	30	82	67	34	34	34	30	42.2	60.6	91.6
Outcome	64	50	-14	32	31	30	72	59	33	34	34	31	52.1	69.2	92.5
All	89	70	-19	32	32	28	95	77	34	33	33	29	22.4	44.6	91.0
Ysel	67	52	-15	32	31	29	74	61	32	34	33	30	49.1	67.3	93.4
YZsel	53	41	-12	35	33	32	63	53	34	36	35	32	68.1	79.1	94.2
OP + All	7	5	-1	29	28	30	30	28	30	30	27	30	93.7	94.1	94.8
OP + Ysel	6	5	-1	29	27	29	29	28	30	29	27	29	94.4	94.5	94.4
OP + YZsel	5	5	-1	29	27	30	29	27	30	29	27	30	94.5	94.6	94.4
Random Forests															
Confounder	62	49	-13	32	31	29	70	58	32	33	32	30	52.8	67.6	93.1
Treatment	81	63	-18	31	31	28	87	70	33	32	32	29	29.8	51.9	90.8
Outcome	75	59	-16	30	30	28	81	66	32	32	32	29	35.3	56.5	92.0
All	93	73	-20	31	31	27	98	79	34	32	32	28	16.9	38.1	89.3
Ysel	77	60	-17	31	30	28	83	67	32	32	32	29	33.2	54.7	92.0
YZsel	65	51	-14	32	32	29	73	60	32	33	33	30	48.4	66.3	93.6
OP + All	6	4	-1	30	28	30	30	28	30	30	28	30	94.2	94.2	94.1
OP + Ysel	5	4	0	29	27	30	29	28	30	30	27	30	94.5	94.6	94.6
OP + YZsel	5	4	0	29	27	30	29	28	30	30	27	30	94.5	94.6	94.9

Table C.9: Simulation results for the scenario with dense models and sample size of 1000. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	38	26	-13	53	52	53	65	58	55	54	54	54	90.3	92.4	94.8
OAL	10	3	-7	47	46	47	48	47	48	55	55	55	97.4	97.2	97.2
LOGIS															
Confounder	10	7	-3	50	49	49	51	49	49	51	49	50	95.2	94.7	95.7
Treatment	-1	0	1	51	49	50	51	49	50	52	50	52	95.7	95.1	95.3
Outcome	-1	0	1	48	46	47	48	46	47	49	47	48	95.3	95.8	96.0
All	35	24	-11	52	52	53	63	57	54	55	54	54	91.1	92.7	95.0
Ysel	8	0	-8	47	46	48	48	46	49	56	56	56	98.1	97.8	97.5
YZsel	28	18	-9	53	54	51	60	57	52	54	54	54	92.0	93.8	94.6
OP + All	12	2	-11	45	44	45	47	44	47	55	54	55	98.1	98.2	97.7
OP + Ysel	8	0	-8	47	46	48	48	46	49	48	46	47	94.8	94.6	94.0
OP + YZsel	12	4	-8	46	47	45	48	47	45	46	44	46	94.6	95.4	94.6
CART															
Confounder	18	14	-5	68	64	67	70	65	67	70	69	70	94.6	95.6	95.6
Treatment	19	14	-5	72	69	71	75	71	71	73	72	72	94.4	95.0	95.3
Outcome	16	12	-4	69	67	69	71	68	69	72	71	72	94.2	95.5	95.0
All	28	19	-9	81	80	79	85	82	79	80	79	81	92.8	93.1	95.0
Ysel	20	10	-10	70	70	69	72	70	70	73	72	73	95.3	94.9	95.4
YZsel	31	19	-13	63	62	62	70	65	63	65	64	65	91.6	92.7	94.9
OP + All	12	5	-7	73	73	71	74	73	72	75	73	75	94.7	94.3	95.2
OP + Ysel	14	3	-11	64	64	63	65	64	64	66	64	66	94.8	94.2	95.1
OP + YZsel	15	4	-11	57	55	57	59	55	58	56	55	56	92.1	94.9	94.4
Pruned CART															
Confounder	22	14	-7	58	56	57	62	58	57	59	59	59	94.2	94.3	95.2
Treatment	26	18	-8	61	58	59	66	61	59	62	60	60	93.1	94.3	95.3
Outcome	24	16	-7	59	56	58	64	58	59	60	59	60	93.0	94.6	94.9
All	32	22	-11	62	62	61	70	65	62	62	61	62	91.9	92.9	94.8
Ysel	27	17	-10	58	57	57	64	59	58	60	59	60	93.2	94.8	95.3
YZsel	31	19	-12	57	56	53	65	59	54	58	58	59	92.1	89.9	93.8
OP + All	11	2	-10	53	53	53	55	53	53	54	52	53	94.4	94.9	94.2
OP + Ysel	12	2	-10	50	50	49	52	50	51	51	50	51	94.8	95.0	94.6
OP + YZsel	15	4	-11	48	48	47	50	48	49	49	48	50	93.8	94.4	94.9
Bagged CART															
Confounder	1	0	0	64	64	65	64	64	65	67	66	66	95.6	95.4	94.9
Treatment	7	5	-2	62	60	61	63	60	61	65	63	64	96.1	95.7	95.9
Outcome	3	2	-1	55	54	55	56	54	55	62	61	62	97.2	97.0	97.0
All	29	19	-10	53	52	53	60	55	53	57	56	57	93.8	95.3	96.3
Ysel	13	4	-9	53	51	53	54	52	54	60	60	61	97.6	97.4	97.2
YZsel	11	9	-2	106	103	98	106	104	98	89	90	89	91.0	92.1	91.0
OP + All	12	1	-11	45	44	46	47	44	47	46	45	46	94.8	95.0	93.8
OP + Ysel	12	1	-11	45	44	46	47	44	47	46	45	46	95.2	95.0	94.4
OP + YZsel	15	5	-10	47	47	47	49	48	48	47	46	48	93.3	94.4	94.9
Random Forests															
Confounder	11	8	-3	54	54	55	56	54	55	60	59	60	96.6	96.9	96.2
Treatment	15	11	-4	54	52	53	56	53	53	58	57	58	96.2	96.0	96.6
Outcome	13	9	-4	49	48	49	51	49	49	57	57	57	97.4	97.0	97.7
All	32	21	-10	51	51	52	60	55	53	55	54	55	92.9	94.2	95.7
Ysel	20	10	-10	48	48	49	52	49	50	56	56	57	96.9	97.0	97.1
YZsel	9	8	-1	148	144	137	148	144	137	89	88	85	84.3	82.6	86.5
OP + All	12	1	-11	46	45	46	47	45	47	46	45	46	94.6	94.8	93.7
OP + Ysel	12	1	-11	45	44	46	47	44	47	46	45	46	94.8	95.3	94.3
OP + YZsel	13	4	-9	59	59	49	60	59	50	50	48	49	94.9	94.9	94.9

Table C.10: Simulation results for the scenario with nonlinear main effects and no interactions. Each simulated dataset contained 500 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

Estimators	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	41	27	-15	38	37	38	56	45	41	38	38	38	81.0	89.6	93.8
OAL	6	2	-4	34	32	33	34	32	33	39	38	39	97.4	98.2	98.0
LOGIS															
Confounder	13	8	-5	36	34	35	38	35	35	36	34	35	93.7	93.8	94.6
Treatment	2	0	-2	36	34	35	36	34	35	36	35	35	95.0	94.8	95.0
Outcome	1	0	-2	34	32	33	34	32	33	34	32	33	95.2	95.3	94.9
All	35	22	-12	37	36	37	51	42	39	39	38	38	85.9	92.1	94.5
Ysel	5	0	-6	34	32	33	34	32	33	39	38	39	97.8	98.0	97.8
YZsel	21	13	-7	37	35	36	42	37	37	38	38	38	91.5	94.7	95.3
OP + All	9	3	-7	33	31	32	34	31	33	38	38	38	97.4	98.0	98.0
OP + Ysel	5	0	-6	34	32	33	34	32	33	33	32	33	94.7	95.0	94.6
OP + YZsel	8	2	-6	32	30	32	33	30	32	33	31	32	95.4	95.7	95.6
CART															
Confounder	18	13	-6	48	47	48	51	49	49	50	49	50	94.1	94.4	96.2
Treatment	23	13	-9	50	48	50	55	50	51	52	51	52	93.9	95.4	95.4
Outcome	16	11	-5	48	46	48	50	48	49	51	51	51	94.8	95.8	95.6
All	32	20	-12	59	57	59	67	60	60	58	57	58	91.0	92.8	93.6
Ysel	19	12	-8	49	48	51	52	49	51	52	51	52	94.3	95.4	95.0
YZsel	24	16	-8	43	42	43	49	45	44	46	46	46	94.9	93.8	95.3
OP + All	10	3	-7	54	52	53	55	52	53	52	51	52	93.4	94.2	94.3
OP + Ysel	11	4	-7	45	44	46	46	44	47	46	45	46	94.4	94.2	94.8
OP + YZsel	10	4	-7	38	36	38	40	36	39	40	39	39	95.1	94.5	94.5
Pruned CART															
Confounder	23	15	-8	41	39	40	47	42	41	41	41	41	90.4	94.2	95.1
Treatment	30	20	-11	43	40	41	52	45	43	43	42	42	88.8	93.1	94.7
Outcome	23	14	-9	40	39	40	46	41	41	42	41	41	91.8	94.7	95.3
All	36	23	-13	42	41	43	56	47	44	43	42	42	85.8	91.2	93.4
Ysel	26	16	-10	40	39	40	48	42	42	42	41	41	90.9	93.6	94.8
YZsel	27	17	-10	38	36	38	47	40	39	40	40	40	89.3	93.6	94.7
OP + All	10	3	-8	38	36	37	39	36	37	37	35	36	93.5	94.4	93.7
OP + Ysel	10	3	-7	35	34	35	37	34	36	36	34	35	94.8	94.8	94.6
OP + YZsel	8	2	-6	33	31	33	34	32	34	35	33	34	95.3	95.7	95.1
Bagged CART															
Confounder	1	0	-1	44	44	43	44	44	43	46	45	46	96.4	95.6	96.2
Treatment	9	4	-5	44	42	42	45	42	42	46	44	44	95.4	95.9	95.8
Outcome	4	1	-2	38	36	38	38	36	38	43	42	42	96.9	97.0	97.2
All	30	19	-11	37	36	37	48	40	39	40	39	40	90.3	94.4	95.5
Ysel	10	4	-6	37	35	36	38	35	37	42	41	42	96.7	97.6	97.7
YZsel	-4	-1	3	75	73	71	75	73	71	65	63	64	91.6	90.4	92.6
OP + All	10	2	-7	33	31	32	34	31	33	33	32	33	93.8	94.8	94.4
OP + Ysel	9	2	-7	33	31	32	34	31	33	33	31	32	94.0	95.0	94.4
OP + YZsel	10	3	-6	32	30	32	34	30	33	33	32	33	94.5	95.9	94.7
Random Forests															
Confounder	13	8	-4	38	37	38	41	38	38	42	41	41	95.3	96.1	97.3
Treatment	17	10	-7	38	36	37	41	38	38	41	40	40	94.7	95.6	96.2
Outcome	14	9	-5	35	33	34	37	34	35	40	39	40	95.7	97.8	97.9
All	34	22	-12	37	35	36	50	41	38	39	38	39	87.0	92.7	95.0
Ysel	18	10	-8	34	32	34	39	34	35	40	39	40	95.4	97.6	97.6
YZsel	-28	-23	5	106	103	92	110	106	92	71	67	63	79.1	85.2	89.5
OP + All	9	2	-7	33	31	33	34	31	33	33	32	33	94.1	95.2	94.6
OP + Ysel	9	2	-7	33	31	32	34	31	33	33	32	32	94.0	94.9	94.6
OP + YZsel	7	2	-5	35	33	36	36	33	36	34	33	34	95.5	96.1	94.9

Table C.11: Simulation results for the scenario with nonlinear main effects and no interactions. Each simulated dataset contained 1000 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

Estimators	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	128	62	-66	54	54	53	139	82	84	54	54	54	35.1	78.6	78.0
OAL	33	31	-2	51	48	48	61	57	48	56	55	55	92.4	93.3	96.8
LOGIS															
Confounder	8	31	23	52	50	50	52	59	55	53	51	51	95.6	91.1	92.2
Treatment	-6	23	29	56	52	53	56	57	60	58	54	54	95.4	94.0	91.1
Outcome	-3	24	27	49	46	47	49	52	55	51	48	49	95.6	92.8	91.3
All	95	53	-42	51	52	50	108	75	66	56	55	55	60.5	84.4	90.8
Ysel	7	24	17	49	47	48	50	52	51	56	56	56	97.0	95.5	96.4
YZsel	48	39	-10	58	52	54	75	65	55	55	53	53	83.6	89.9	94.4
OP + All	25	20	-5	46	45	45	52	49	45	56	55	55	96.1	96.6	98.0
OP + Ysel	7	24	17	49	47	48	50	52	51	48	47	47	94.3	91.8	92.2
OP + YZsel	13	22	9	48	46	47	50	51	47	48	46	47	94.8	92.1	94.5
CART															
Confounder	47	34	-14	72	68	65	86	76	67	73	72	70	90.0	92.8	95.7
Treatment	61	39	-22	73	73	70	95	82	74	75	73	73	86.9	91.3	94.2
Outcome	50	35	-15	73	70	69	89	78	71	74	73	72	90.1	92.0	94.8
All	89	52	-38	83	81	82	122	96	90	82	81	80	79.7	89.2	91.8
Ysel	56	36	-20	73	71	70	92	80	73	75	74	73	89.1	92.4	94.4
YZsel	70	42	-27	68	65	64	97	77	70	70	69	68	85.4	91.1	94.6
OP + All	29	19	-10	75	73	74	81	76	74	75	74	74	93.0	94.0	94.6
OP + Ysel	32	21	-11	66	64	64	73	68	65	67	66	65	92.7	93.8	94.6
OP + YZsel	29	22	-8	60	57	57	67	61	58	62	60	59	93.1	94.0	94.6
Pruned CART															
Confounder	57	37	-20	66	60	59	87	71	63	64	63	62	82.8	91.3	94.1
Treatment	76	44	-32	67	64	62	101	78	70	65	63	63	76.2	87.8	93.0
Outcome	65	40	-25	64	61	59	91	73	64	64	62	62	81.3	88.9	94.0
All	100	54	-46	65	61	63	119	82	78	64	62	62	63.0	85.5	87.8
Ysel	71	42	-29	64	62	60	96	74	67	64	62	61	78.4	88.8	92.6
YZsel	77	45	-33	62	58	57	99	73	66	62	61	60	76.0	89.0	91.7
OP + All	23	19	-4	54	52	52	59	55	52	55	53	53	93.8	93.8	95.0
OP + Ysel	24	20	-4	54	53	51	59	57	51	54	53	52	93.2	93.2	95.0
OP + YZsel	23	20	-3	51	49	49	56	53	50	53	51	50	94.0	92.9	95.3
Bagged CART															
Confounder	-4	14	18	74	70	69	74	72	71	73	71	70	94.3	94.0	93.4
Treatment	22	30	8	66	62	63	70	69	63	68	65	65	94.2	92.8	95.1
Outcome	12	23	12	61	56	58	62	61	59	65	64	64	95.7	94.9	96.0
All	85	51	-34	52	53	51	99	73	61	58	57	57	70.3	86.6	93.7
Ysel	28	28	0	57	53	54	64	60	54	63	62	62	94.6	95.9	97.0
YZsel	15	22	7	93	88	90	94	91	90	84	81	79	93.2	93.8	92.2
OP + All	20	19	-2	47	45	45	51	49	45	48	46	46	94.1	93.1	95.1
OP + Ysel	21	21	0	47	45	45	51	50	45	48	46	46	94.1	93.4	95.4
OP + YZsel	22	19	-3	47	45	45	52	49	45	48	46	46	93.6	93.2	95.3
Random Forests															
Confounder	29	27	-2	59	56	56	66	62	56	63	61	61	94.0	94.4	96.7
Treatment	53	39	-14	55	54	53	76	67	54	59	58	58	87.8	90.4	96.7
Outcome	46	36	-10	52	49	49	69	61	50	58	57	58	89.8	93.8	97.8
All	105	56	-49	52	52	50	117	76	70	55	55	55	52.3	82.9	88.3
Ysel	61	40	-21	50	49	48	79	63	53	57	56	57	84.9	92.6	96.9
YZsel	33	30	-3	102	97	92	108	101	92	74	71	70	90.4	92.0	94.0
OP + All	19	18	-1	47	45	45	51	49	45	48	46	46	94.0	93.0	95.0
OP + Ysel	20	21	1	47	45	45	51	49	45	48	46	46	94.1	93.0	95.0
OP + YZsel	21	19	-2	47	45	45	51	49	45	48	46	46	93.8	92.9	95.3

Table C.12: Simulation results for the scenario with linear main effects and linear interactions. Each simulated dataset contained 500 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

Estimators	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	128	62	-66	38	38	38	133	73	76	38	38	38	8.7	63.2	59.2
OAL	23	31	8	39	34	36	46	46	36	39	39	39	90.6	90.1	96.6
LOGIS															
Confounder	8	33	25	37	36	35	38	48	43	37	36	36	94.4	85.0	88.3
Treatment	-7	24	32	39	37	37	40	44	48	40	37	37	94.8	89.2	85.7
Outcome	-4	25	29	35	33	33	35	42	44	35	33	34	94.7	88.1	85.6
All	71	47	-24	36	36	35	80	60	42	40	39	39	57.5	78.4	93.4
Ysel	0	25	25	35	33	33	35	42	41	39	39	39	96.9	92.7	93.1
YZsel	16	32	16	40	36	37	43	48	40	37	36	36	91.6	86.9	91.0
OP + All	15	22	7	33	33	32	37	40	33	40	39	39	96.4	94.2	97.2
OP + Ysel	0	25	25	35	33	33	35	42	41	33	33	33	94.0	87.9	87.4
OP + YZsel	3	25	22	35	33	33	35	42	40	35	33	33	94.5	88.2	89.0
CART															
Confounder	46	34	-12	52	50	49	69	60	50	53	51	50	86.6	91.0	94.1
Treatment	56	38	-18	54	53	49	78	65	52	55	53	52	82.5	88.2	94.3
Outcome	47	34	-14	52	50	48	70	61	50	54	53	52	86.4	91.0	95.8
All	77	47	-30	59	58	58	97	75	65	60	58	58	74.1	87.1	92.2
Ysel	52	36	-16	53	52	50	74	63	52	55	53	52	85.1	89.7	94.8
YZsel	52	36	-16	51	48	47	73	60	50	52	51	50	83.9	89.3	94.7
OP + All	21	18	-3	52	52	51	57	55	51	53	52	52	93.6	93.4	94.6
OP + Ysel	29	23	-6	47	46	45	55	52	46	48	46	46	90.6	90.8	94.7
OP + YZsel	28	22	-6	45	43	43	53	48	44	45	43	43	90.4	91.4	94.2
Pruned CART															
Confounder	53	35	-18	46	43	42	70	55	46	45	44	43	76.2	86.7	93.2
Treatment	70	43	-28	49	45	43	86	62	51	46	44	44	63.5	82.6	89.8
Outcome	58	35	-22	46	42	41	74	55	47	44	43	43	71.6	86.6	91.8
All	92	50	-41	46	43	43	103	66	59	44	42	42	44.0	76.9	82.2
Ysel	63	38	-25	46	43	43	78	57	49	44	43	43	68.4	84.9	90.7
YZsel	60	37	-22	46	42	41	75	56	47	44	43	42	71.0	86.1	92.2
OP + All	15	19	4	37	36	35	40	41	36	37	36	36	93.6	91.3	94.7
OP + Ysel	18	20	2	38	36	36	42	42	36	37	36	36	92.1	90.6	94.8
OP + YZsel	18	20	2	38	36	36	42	41	36	37	36	35	91.7	90.8	94.0
Bagged CART															
Confounder	-5	12	16	53	51	48	53	52	50	53	51	49	93.9	93.8	94.0
Treatment	19	28	8	46	44	42	50	52	43	48	46	46	93.8	91.2	95.5
Outcome	9	20	11	43	42	39	44	46	41	47	45	45	95.8	94.0	96.2
All	74	48	-26	38	37	36	83	61	45	41	40	40	55.4	79.1	92.2
Ysel	21	25	4	42	41	39	47	48	39	45	44	43	94.2	92.8	96.7
YZsel	-1	14	15	55	52	50	55	54	52	55	52	51	94.7	94.3	93.7
OP + All	13	22	9	34	33	33	36	40	34	34	33	33	93.7	89.7	93.6
OP + Ysel	16	22	6	34	33	33	38	40	33	35	33	33	93.3	89.8	94.3
OP + YZsel	16	21	5	34	33	33	38	39	33	35	33	33	92.8	90.5	94.9
Random Forests															
Confounder	29	26	-3	42	41	39	51	48	39	44	43	43	91.2	91.1	96.1
Treatment	51	38	-13	39	38	37	64	54	39	42	41	41	78.6	86.2	96.0
Outcome	43	34	-9	36	35	34	56	49	35	41	40	41	85.2	90.3	97.4
All	99	55	-45	37	37	36	106	66	57	39	39	39	26.1	71.3	81.0
Ysel	54	38	-16	36	35	34	65	52	38	40	40	40	75.2	87.3	96.5
YZsel	30	27	-4	47	44	43	56	51	43	46	44	44	90.4	91.7	96.5
OP + All	11	21	10	34	34	33	36	40	34	34	33	33	94.2	89.4	92.7
OP + Ysel	14	22	7	34	34	33	37	40	33	35	33	33	93.3	90.0	93.7
OP + YZsel	15	20	5	34	33	33	37	39	33	35	33	33	93.2	90.5	94.6

Table C.13: Simulation results for the scenario with linear main effects and linear interactions. Each simulated dataset contained 1000 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

Estimators	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	33	16	-18	54	53	53	64	56	55	55	54	53	90.5	94.6	93.8
OAL	10	4	-6	48	47	45	49	47	45	56	55	54	97.6	98.2	97.8
LOGIS															
Confounder	9	9	-1	51	50	48	51	51	48	51	51	49	94.9	95.3	95.2
Treatment	-1	1	2	51	51	49	51	51	49	53	52	50	95.0	95.3	95.1
Outcome	-1	2	2	48	46	45	48	46	45	49	48	47	95.1	96.2	95.6
All	32	15	-17	54	53	52	62	55	55	55	55	53	91.1	95.1	94.2
Ysel	8	1	-6	48	47	45	49	47	46	56	56	55	97.7	98.7	98.2
YZsel	25	13	-12	53	55	56	59	56	57	55	54	53	92.8	96.2	94.5
OP + All	12	1	-11	46	45	43	47	45	44	55	55	54	97.7	98.6	98.2
OP + Ysel	8	1	-6	48	47	45	49	47	46	48	48	46	94.2	96.2	95.8
OP + YZsel	12	4	-8	47	46	47	48	47	48	46	46	44	94.5	96.2	92.8
CART															
Confounder	14	9	-5	68	68	64	70	69	64	71	71	69	94.9	95.2	96.0
Treatment	17	9	-8	74	71	69	76	72	69	73	73	71	94.0	94.0	95.5
Outcome	16	8	-8	71	69	67	73	69	68	73	72	71	94.5	95.4	95.3
All	24	12	-11	80	80	77	83	81	78	81	80	79	94.3	94.0	95.2
Ysel	19	9	-10	72	68	68	74	69	69	74	73	72	94.3	95.4	95.5
YZsel	26	15	-11	60	66	63	65	68	63	65	65	63	95.8	95.8	93.8
OP + All	10	1	-9	74	74	69	75	74	70	75	74	73	94.9	95.0	95.7
OP + Ysel	12	1	-10	66	64	62	67	64	63	66	65	64	94.6	95.1	94.9
OP + YZsel	15	5	-10	52	58	54	54	58	55	57	56	54	95.8	95.8	93.8
Pruned CART															
Confounder	20	11	-8	61	59	57	64	60	57	61	61	59	93.2	94.8	95.4
Treatment	22	11	-11	62	60	58	66	61	59	62	62	60	93.3	95.2	95.2
Outcome	22	11	-11	60	59	56	64	60	57	61	61	60	93.5	95.0	95.0
All	28	13	-15	62	61	60	69	62	62	62	62	60	92.4	95.0	94.3
Ysel	26	13	-13	60	57	56	65	58	58	61	61	59	93.1	95.9	94.4
YZsel	28	17	-12	55	57	59	62	60	60	59	59	57	92.7	94.8	95.8
OP + All	10	1	-10	54	52	50	55	52	51	54	53	52	94.8	96.4	94.7
OP + Ysel	12	2	-10	52	51	49	53	51	50	53	52	50	94.8	95.8	95.0
OP + YZsel	12	5	-7	48	49	49	50	49	49	51	49	47	94.8	97.9	93.8
Bagged CART															
Confounder	4	3	-1	66	64	64	66	64	64	67	67	65	95.2	95.4	95.4
Treatment	6	4	-2	62	61	57	62	61	57	64	64	61	95.2	95.4	95.6
Outcome	5	4	-1	55	53	52	55	54	52	61	62	60	97.1	97.1	97.7
All	25	12	-13	53	53	50	59	54	52	57	57	55	94.3	96.4	96.4
Ysel	13	5	-8	53	53	51	54	53	52	60	60	59	97.5	97.4	97.5
YZsel	15	2	-12	102	102	103	103	102	104	89	90	87	93.8	99.0	93.8
OP + All	12	1	-11	46	45	43	48	45	45	47	46	45	94.2	96.4	95.0
OP + Ysel	12	1	-11	46	45	44	48	45	45	47	46	45	94.4	96.3	94.8
OP + YZsel	14	3	-11	49	49	48	51	49	49	48	47	45	94.8	96.9	93.8
Random Forests															
Confounder	11	7	-4	57	55	54	58	55	54	60	60	59	95.5	96.2	96.4
Treatment	14	8	-6	54	53	51	56	54	51	58	58	56	95.5	96.3	96.4
Outcome	13	8	-5	50	49	47	52	50	47	57	57	56	97.5	98.0	98.2
All	28	13	-14	52	52	50	59	53	52	55	55	54	93.2	95.7	95.7
Ysel	18	7	-11	49	48	46	53	49	48	57	57	55	96.4	97.7	97.7
YZsel	16	7	-9	138	139	137	139	139	137	90	92	86	86.5	91.7	93.8
OP + All	12	0	-11	46	45	44	48	45	45	47	46	45	93.6	96.2	94.8
OP + Ysel	11	1	-10	46	45	44	47	45	45	47	46	45	94.6	96.5	94.6
OP + YZsel	7	3	-4	63	58	57	64	58	57	50	50	48	93.8	97.9	93.8

Table C.14: Simulation results for the scenario with nonlinear main effects and linear interactions. Each simulated dataset contained 500 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

Estimators	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	36	17	-19	39	39	38	53	42	42	39	38	38	83.9	92.8	92.8
OAL	6	5	-1	34	34	33	34	34	33	39	39	38	97.3	97.5	97.2
LOGIS															
Confounder	12	11	-1	36	36	34	38	38	34	36	36	34	94.2	93.4	94.5
Treatment	2	4	2	37	36	34	37	36	34	36	36	34	95.0	94.6	94.2
Outcome	1	4	2	34	33	33	34	34	33	34	33	32	95.1	94.0	93.7
All	32	15	-17	38	38	37	50	41	40	39	39	38	87.2	92.8	93.8
Ysel	5	3	-2	34	33	33	34	34	33	39	39	38	97.4	97.4	97.0
YZsel	23	12	-10	39	38	37	45	40	38	38	38	38	90.0	92.3	94.3
OP + All	9	4	-5	33	33	32	34	33	33	39	39	38	97.2	97.6	97.0
OP + Ysel	5	3	-2	34	33	33	34	34	33	33	33	32	94.8	94.4	93.2
OP + YZsel	8	7	-2	33	32	31	34	33	31	33	33	31	95.0	93.8	94.3
CART															
Confounder	17	12	-5	49	49	47	52	50	48	51	50	49	94.0	94.1	96.0
Treatment	18	11	-7	52	52	50	55	53	50	53	52	51	93.8	93.7	95.3
Outcome	15	8	-6	51	49	47	53	50	48	52	52	50	94.6	95.4	95.5
All	27	15	-13	58	58	55	64	60	56	58	58	57	92.8	93.6	95.0
Ysel	17	9	-8	51	50	49	54	51	50	53	53	51	94.3	95.0	95.3
YZsel	28	15	-13	45	45	44	53	48	46	46	46	45	92.7	95.8	95.2
OP + All	11	5	-6	52	52	51	53	53	51	53	52	51	94.8	94.9	94.3
OP + Ysel	10	4	-6	48	46	46	49	46	46	47	46	45	94.0	95.6	93.4
OP + YZsel	12	7	-4	39	40	38	41	41	38	40	39	38	94.1	96.2	95.2
Pruned CART															
Confounder	22	13	-9	44	43	41	49	45	42	44	44	42	91.2	94.0	95.1
Treatment	24	13	-11	45	43	42	51	45	43	44	43	42	90.8	93.4	95.1
Outcome	22	12	-9	44	42	41	49	43	42	44	43	42	91.6	94.4	95.2
All	32	16	-16	44	44	42	54	47	44	43	43	42	87.8	92.8	93.9
Ysel	23	12	-11	43	42	41	49	44	43	43	43	42	91.1	93.9	94.8
YZsel	29	15	-14	41	40	40	50	43	42	41	41	40	91.0	92.7	95.5
OP + All	9	4	-5	37	37	36	38	37	36	37	37	35	94.6	94.8	94.0
OP + Ysel	9	4	-5	39	37	37	40	37	37	38	37	36	94.4	95.3	94.0
OP + YZsel	9	6	-4	34	35	33	35	35	34	35	34	33	95.8	94.5	93.1
Bagged CART															
Confounder	5	6	1	46	44	43	46	45	43	47	47	45	95.4	95.6	95.2
Treatment	9	7	-2	43	41	39	44	42	39	45	44	42	95.4	95.9	96.1
Outcome	6	6	-1	38	38	37	39	38	37	43	43	41	97.0	97.2	97.5
All	26	14	-12	38	38	36	46	40	37	40	40	39	92.1	95.0	95.8
Ysel	12	7	-4	37	37	36	39	37	36	42	42	41	96.2	97.3	96.8
YZsel	6	7	1	73	73	73	73	74	73	62	62	60	90.3	91.0	90.0
OP + All	9	4	-5	33	33	32	34	33	33	33	33	32	94.2	94.3	93.3
OP + Ysel	9	4	-4	33	33	33	35	33	33	34	33	32	94.2	94.0	93.3
OP + YZsel	10	6	-4	32	32	32	34	33	32	33	33	32	96.2	95.5	95.5
Random Forests															
Confounder	13	10	-4	39	38	38	41	39	38	42	42	41	95.6	96.4	96.3
Treatment	15	9	-6	38	38	36	41	39	36	41	40	39	95.1	96.0	96.0
Outcome	14	9	-5	35	35	34	38	36	35	40	40	39	96.4	97.0	97.2
All	29	15	-14	37	37	35	47	40	38	39	39	38	89.8	94.2	95.2
Ysel	17	9	-8	35	35	34	39	36	35	40	40	39	95.0	96.8	97.4
YZsel	-11	-2	9	108	104	95	109	104	96	65	67	55	84.1	89.6	86.2
OP + All	9	4	-5	33	33	32	34	33	33	33	33	32	94.0	94.0	93.4
OP + Ysel	8	4	-4	33	33	33	34	33	33	34	33	32	94.8	94.4	93.6
OP + YZsel	9	4	-5	38	34	38	39	35	38	34	34	33	96.9	96.9	95.2

Table C.15: Simulation results for the scenario with nonlinear main effects and linear interactions. Each simulated dataset contained 1000 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	62	38	-24	56	54	54	84	67	59	55	54	54	79.0	88.6	92.6
OAL	13	6	-7	50	47	48	52	48	48	56	55	55	96.1	97.2	97.3
LOGIS															
Confounder	9	8	-1	52	51	50	53	51	50	52	50	50	94.2	94.0	94.7
Treatment	0	1	1	52	51	51	52	51	51	53	51	51	94.7	93.9	94.5
Outcome	0	2	2	49	47	48	49	47	48	50	47	48	94.0	94.3	94.6
All	48	30	-18	54	53	53	73	61	56	55	54	54	86.2	91.2	93.8
Ysel	8	1	-7	49	47	48	50	47	49	57	56	56	96.9	97.8	97.4
YZsel	21	13	-7	56	52	53	59	54	53	55	54	53	93.2	94.1	94.8
OP + All	14	4	-10	46	45	45	48	45	47	55	54	54	97.0	98.0	97.8
OP + Ysel	8	1	-7	49	47	48	50	47	49	48	47	47	93.4	94.7	93.3
OP + YZsel	11	2	-9	49	46	46	50	46	47	47	45	45	93.6	94.0	95.6
CART															
Confounder	23	16	-7	71	68	69	75	70	70	71	70	69	93.7	94.6	94.7
Treatment	22	15	-6	72	69	70	75	71	70	74	72	72	94.2	94.3	94.8
Outcome	23	15	-8	70	68	68	74	69	68	73	72	72	94.0	95.4	95.5
All	34	22	-13	84	81	81	90	84	81	81	80	80	92.0	93.0	93.9
Ysel	27	16	-11	71	69	70	76	71	71	75	73	73	94.2	94.7	94.8
YZsel	28	15	-13	67	64	65	73	65	66	68	67	66	93.9	95.2	95.4
OP + All	13	3	-10	75	74	73	76	75	74	75	73	74	93.8	94.2	94.1
OP + Ysel	14	5	-9	65	64	65	67	64	65	67	65	65	94.8	95.0	94.4
OP + YZsel	18	3	-15	59	56	55	62	56	57	60	58	58	93.9	96.2	95.0
Pruned CART															
Confounder	28	19	-10	63	60	59	69	63	60	61	60	59	91.3	93.2	94.0
Treatment	31	20	-11	63	60	59	70	64	60	62	61	60	91.1	93.4	94.5
Outcome	29	19	-11	62	60	58	68	63	59	62	60	59	90.8	93.8	94.4
All	40	26	-15	65	63	61	76	68	63	62	61	60	88.6	92.0	93.3
Ysel	31	20	-12	62	59	60	70	63	61	62	61	60	91.1	93.2	93.8
YZsel	32	20	-13	60	57	56	68	60	58	60	59	57	91.8	93.7	93.9
OP + All	12	3	-9	53	52	52	55	52	52	54	52	52	95.0	94.2	94.4
OP + Ysel	12	3	-9	53	51	52	54	51	53	53	51	51	94.6	93.8	93.8
OP + YZsel	14	3	-11	52	49	48	54	49	49	51	49	49	94.5	96.6	95.2
Bagged CART															
Confounder	-11	-6	5	76	70	70	76	71	70	74	71	69	94.0	95.4	94.0
Treatment	-1	0	0	66	63	61	66	63	61	67	65	63	94.8	95.6	95.2
Outcome	-3	0	2	63	60	58	63	60	58	67	65	63	95.6	95.9	96.6
All	37	23	-13	55	53	53	66	58	54	57	57	56	91.4	94.2	95.5
Ysel	9	3	-5	59	56	55	60	56	55	64	62	61	96.2	96.8	97.0
YZsel	-21	-13	8	105	99	98	107	100	98	90	88	86	88.0	94.5	92.9
OP + All	13	3	-10	47	45	46	49	46	47	47	45	46	93.4	94.1	94.2
OP + Ysel	12	3	-10	47	46	46	48	46	47	47	45	46	93.8	93.8	93.8
OP + YZsel	15	4	-11	48	45	46	50	45	47	47	45	46	93.5	96.6	95.2
Random Forests															
Confounder	11	8	-3	60	57	57	61	58	57	63	61	60	95.0	95.4	95.6
Treatment	20	13	-7	55	53	54	59	55	54	59	58	57	94.5	96.0	95.5
Outcome	18	13	-6	53	50	51	56	52	51	59	58	57	95.8	96.4	97.2
All	49	31	-19	54	52	52	73	61	55	55	55	54	86.4	92.0	94.5
Ysel	28	16	-12	52	49	50	59	52	51	57	57	56	94.8	96.6	97.0
YZsel	-20	-15	4	138	131	124	139	132	124	88	86	81	87.8	90.8	93.3
OP + All	13	3	-10	47	46	46	49	46	47	47	45	46	94.2	94.3	93.7
OP + Ysel	12	2	-9	47	46	46	48	46	47	47	45	46	94.0	94.3	94.3
OP + YZsel	15	3	-11	49	46	45	51	46	47	47	45	46	93.7	95.8	95.4

Table C.16: Simulation results for the scenario with nonlinear main effects and nonlinear interactions. Each simulated dataset contained 500 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

Estimators	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	61	38	-23	39	38	39	73	54	45	39	38	38	64.6	82.8	89.3
OAL	3	2	-1	35	33	33	35	33	33	39	39	38	97.2	97.4	97.4
LOGIS															
Confounder	8	7	-1	37	35	36	38	36	36	36	35	35	93.8	94.3	94.8
Treatment	0	1	2	37	36	35	37	36	35	37	35	35	95.0	94.4	95.1
Outcome	-1	1	1	35	33	33	35	33	33	34	33	33	94.6	95.2	94.3
All	36	24	-13	38	37	38	52	44	40	39	38	38	85.9	91.3	93.4
Ysel	3	0	-3	35	33	33	35	33	34	39	39	38	97.2	97.9	97.2
YZsel	10	8	-2	39	37	38	40	38	38	38	37	37	93.6	94.6	94.6
OP + All	8	3	-5	33	32	32	34	32	33	39	38	38	97.3	98.0	97.9
OP + Ysel	3	0	-3	35	33	33	35	33	34	33	32	32	93.4	95.2	93.9
OP + YZsel	5	2	-3	34	32	33	35	32	33	34	32	32	94.1	95.3	94.0
CART															
Confounder	20	13	-7	49	48	48	53	50	49	51	50	49	94.7	95.1	95.2
Treatment	21	15	-6	51	49	50	55	52	50	53	52	51	93.9	94.9	95.6
Outcome	20	14	-6	50	49	49	54	51	50	53	52	51	94.5	94.6	95.3
All	26	18	-9	57	56	58	62	59	58	58	58	57	93.3	94.5	94.8
Ysel	22	13	-9	51	50	50	55	51	51	53	52	52	94.4	95.9	95.1
YZsel	17	11	-6	46	46	46	49	47	46	48	48	47	94.6	95.0	94.5
OP + All	8	3	-4	52	50	52	53	50	52	53	51	51	94.8	94.4	94.2
OP + Ysel	11	4	-7	47	45	46	48	45	47	47	45	46	94.6	95.0	94.1
OP + YZsel	11	4	-6	42	40	40	43	40	41	42	40	40	93.6	95.3	93.9
Pruned CART															
Confounder	24	16	-8	42	41	41	48	44	42	43	42	41	90.6	93.2	94.2
Treatment	25	16	-9	43	41	42	50	44	42	43	42	41	91.0	93.0	94.0
Outcome	25	16	-9	42	41	41	49	44	42	42	42	41	91.0	93.0	94.2
All	32	20	-12	44	42	42	54	47	44	43	42	41	88.1	91.8	93.9
Ysel	26	17	-9	42	41	40	49	44	42	42	41	41	90.5	93.2	94.7
YZsel	22	14	-8	41	40	40	47	43	41	41	41	40	92.0	93.8	94.8
OP + All	7	1	-5	37	35	36	38	35	36	36	35	35	94.2	94.8	93.7
OP + Ysel	7	2	-5	37	35	35	37	35	35	36	35	35	94.4	95.2	94.2
OP + YZsel	6	1	-4	36	34	34	36	34	34	35	34	34	94.3	95.7	94.0
Bagged CART															
Confounder	-14	-7	7	53	50	47	54	50	48	53	51	48	94.2	95.8	94.8
Treatment	-4	-1	3	46	44	42	46	44	42	48	46	44	95.3	95.3	95.5
Outcome	-6	-3	3	44	41	40	44	41	40	47	46	44	96.2	96.9	96.7
All	26	17	-9	38	37	37	46	40	38	41	40	39	92.2	94.4	95.7
Ysel	2	0	-2	42	40	38	42	40	38	46	44	43	96.9	97.0	97.6
YZsel	-35	-21	14	70	65	66	78	69	68	64	62	60	87.9	91.9	91.5
OP + All	7	2	-5	34	32	33	34	32	33	34	32	32	94.0	94.8	94.0
OP + Ysel	6	2	-4	34	32	33	35	32	33	34	32	32	94.3	94.9	94.3
OP + YZsel	8	3	-4	34	32	33	34	32	33	33	32	32	93.4	95.3	93.8
Random Forests															
Confounder	9	8	-1	41	39	40	42	40	40	44	43	42	95.9	96.2	96.2
Treatment	17	12	-5	38	37	37	42	39	38	41	41	40	95.5	95.2	96.2
Outcome	16	11	-5	36	35	35	40	36	36	41	41	40	96.0	96.8	97.3
All	45	29	-16	37	36	37	58	46	40	39	39	38	80.8	89.6	93.9
Ysel	23	14	-9	36	34	35	42	37	36	40	40	39	94.2	96.2	97.2
YZsel	-34	-25	8	87	79	77	94	83	77	63	60	55	86.6	89.7	92.7
OP + All	6	2	-4	34	32	33	34	32	33	34	32	32	94.3	95.3	94.3
OP + Ysel	6	2	-4	34	32	33	35	32	33	34	32	33	94.4	95.0	94.4
OP + YZsel	7	3	-4	34	32	32	34	32	33	33	32	32	93.8	95.4	94.3

Table C.17: Simulation results for the scenario with nonlinear main effects and nonlinear interactions. Each simulated dataset contained 1000 subjects. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Bias×1000			MCSD×1000			RMSE×1000			SE×1000			Coverage (%; modified)		
Estimators	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3	1 vs 2	1 vs 3	2 vs 3
Naive	-217	-38	179	39	46	45	220	60	185	37	38	36	0	78.6	0.5
OAL	-24	-14	10	40	47	40	47	49	41	38	46	41	89	93.6	95.2
LOGIS															
Confounder	-3	3	6	36	46	42	36	46	42	36	46	42	94.7	95.1	94.8
Treatment	-3	4	7	46	53	44	46	53	45	43	52	44	93.3	94.6	94.7
Outcome	-3	2	5	34	43	40	34	43	40	34	43	40	94.5	95.0	95.2
All	-84	-61	23	38	46	42	92	76	48	40	50	46	45.2	78.6	94.0
Ysel	-16	-7	10	40	47	41	44	48	42	39	48	42	93.0	94.5	95.3
YZsel	-13	-4	8	41	49	42	43	49	43	40	49	43	93.3	94.6	95.3
OP + All	-62	-43	19	33	43	40	70	61	44	40	50	46	69.6	89.4	95.9
OP + Ysel	-16	-7	10	40	47	41	44	48	42	39	48	42	93.0	94.5	95.3
OP + YZsel	-15	-5	9	40	47	41	42	47	42	38	47	42	93.2	94.7	95.0
CART															
Confounder	-54	-35	19	46	61	56	71	70	59	48	61	56	82.6	91.4	93.6
Treatment	-98	-69	29	55	66	61	112	95	67	54	65	61	57.0	82.1	92.8
Outcome	-66	-45	21	48	62	56	82	76	60	50	63	59	76.3	89.6	94.6
All	-131	-89	42	61	76	71	145	117	82	60	73	69	40.8	76.0	89.8
Ysel	-100	-69	31	52	66	63	112	95	70	54	66	62	55.1	82.3	91.6
YZsel	-85	-59	26	51	63	59	99	87	64	52	64	59	64.1	85.2	93.2
OP + All	-71	-45	26	50	68	64	86	82	69	51	67	63	74.8	90.2	92.1
OP + Ysel	-61	-40	22	45	60	57	76	72	61	46	59	56	74.9	89.6	92.2
OP + YZsel	-57	-37	20	43	57	54	71	68	58	44	57	53	77.7	91.1	93.2
Pruned CART															
Confounder	-80	-57	24	50	54	48	94	78	54	43	52	48	54.4	79.1	92.2
Treatment	-117	-85	33	53	59	52	129	103	62	48	57	52	32.4	65.2	90.8
Outcome	-93	-65	28	49	53	47	105	84	55	43	52	48	42.5	74.8	91.2
All	-147	-106	41	49	55	51	155	119	66	45	53	50	12.0	47.5	86.0
Ysel	-122	-87	35	50	54	50	132	102	61	45	54	50	26.0	62.5	88.5
YZsel	-109	-78	32	51	56	50	121	96	59	46	54	50	35.7	69.0	90.6
OP + All	-76	-52	24	37	48	44	84	71	50	36	47	44	44.5	78.4	90.8
OP + Ysel	-67	-46	21	37	47	44	76	66	49	37	47	44	55.2	82.3	91.7
OP + YZsel	-63	-42	21	38	48	45	73	64	49	37	47	44	61.0	84.8	92.3
Bagged CART															
Confounder	67	63	-4	56	69	60	87	93	60	54	68	59	70.5	81.9	95.3
Treatment	-33	-18	15	55	63	55	65	66	57	54	63	55	90.9	94.4	95.3
Outcome	17	20	3	46	55	49	49	59	49	48	59	52	93.6	95.5	97.0
All	-127	-90	37	39	48	45	133	102	58	42	50	47	11.8	56.6	89.3
Ysel	-50	-31	19	51	56	48	72	64	52	48	57	51	80.6	91.6	94.7
YZsel	-7	4	10	64	68	56	64	68	57	55	65	56	90.8	93.3	95.4
OP + All	-60	-40	20	33	43	40	68	59	44	34	43	40	60.0	85.0	92.2
OP + Ysel	-54	-36	18	33	42	39	63	56	43	34	43	40	66.8	86.2	93.2
OP + YZsel	-52	-34	17	33	43	39	61	55	43	34	43	40	68.2	86.9	93.2
Random Forests															
Confounder	42	39	-3	48	57	50	64	69	50	47	59	52	83.4	89.7	95.4
Treatment	-61	-43	18	42	50	46	74	66	49	44	53	48	72.8	89.3	94.8
Outcome	-24	-14	10	36	46	42	43	48	43	41	52	47	94.6	96.7	96.8
All	-149	-107	42	36	45	44	153	116	60	39	48	46	1.8	37.5	86.4
Ysel	-83	-59	25	40	46	42	93	75	49	41	50	47	45.9	80.6	93.7
YZsel	-25	-14	11	66	65	48	70	66	49	45	55	49	79.1	90.6	95.3
OP + All	-42	-27	15	34	44	40	54	52	42	35	44	40	79.2	90.2	93.3
OP + Ysel	-44	-30	14	34	43	39	56	52	42	34	43	40	76.6	89.0	94.0
OP + YZsel	-43	-29	14	34	43	39	55	52	42	34	43	40	76.8	89.4	94.0

Table C.18: Simulation results for setting with censored observations. Standard errors were estimated based on 200 bootstrap replications. Results were obtained using 2000 simulated datasets.

	Total	# Selected by Outcome	# Selected by Treatment	# in the Intersection	Description of Phecodes in the Intersection
Circulatory system	130	8	8	2	Congestive heart failure (CHF) NOS, congestive heart failure; nonhypertensive.
Congenital anomalies	13	0	0	0	
Dermatologic	51	1	2	0	
Digestive	96	2	2	0	
Endocrine/metabolic	95	4	6	1	Type 2 diabetes.
Genitourinary	79	7	6	0	
Hematopoietic	35	5	5	1	Lymphadenitis.
Infectious diseases	32	1	0	0	
Injuries and poisonings	72	2	1	0	
Mental disorders	42	3	5	2	Delirium dementia and amnestic and other cognitive disorders, tobacco use disorder.
Musculoskeletal	82	0	7	0	
Neoplasms	89	8	15	4	Cancer of prostate, secondary malignancy of respiratory organs, secondary malignant neoplasm, secondary malignant neoplasm of liver.
Neurological	46	2	1	0	
Respiratory	64	6	4	1	Abnormal findings examination of lungs.
Sense organ	83	0	2	0	
Symptoms	36	5	5	1	Nausea and vomiting.

Table C.19: Number of Phecodes selected for each group of disease for 180-day risk of ER visits.

	Total	# Selected by Outcome	# Selected by Treatment	# in the Intersection	Description of Phecodes in the Intersection
Circulatory system	130	10	13	3	Atrial fibrillation and flutter, congestive heart failure (CHF) NOS, congestive heart failure; nonhypertensive.
Congenital anomalies	13	0	0	0	
Dermatologic	51	1	3	0	
Digestive	96	2	5	0	
Endocrine/metabolic	95	3	6	1	Type 2 diabetes.
Genitourinary	79	5	8	0	
Hematopoietic	35	3	5	1	Lymphadenitis.
Infectious diseases	32	1	1	0	
Injuries and poisonings	72	2	2	0	
Mental disorders	42	1	6	1	Tobacco use disorder.
Musculoskeletal	82	2	9	0	
Neoplasms	89	5	16	3	Cancer of bronchus; lung, secondary malignant neoplasm, secondary malignant neoplasm of liver.
Neurological	46	3	2	0	
Respiratory	64	6	6	1	Abnormal findings examination of lungs.
Sense organ	83	0	2	0	
Symptoms	36	4	6	1	Malaise and fatigue.

Table C.20: Number of Phecodes selected for each group of disease for 360-day risk of ER visits.

	Total	# Selected by Outcome	# Selected by Treatment	# in the Intersection	Description of Phecodes in the Intersection
Circulatory system	130	7	10	2	Congestive heart failure (CHF) NOS, congestive heart failure; nonhypertensive.
Congenital anomalies	13	0	0	0	
Dermatologic	51	0	3	0	
Digestive	96	2	5	0	
Endocrine/metabolic	95	6	5	0	
Genitourinary	79	10	9	0	
Hematopoietic	35	6	5	1	Lymphadenitis.
Infectious diseases	32	2	0	0	
Injuries and poisonings	72	2	1	0	
Mental disorders	42	2	5	0	
Musculoskeletal	82	1	8	0	
Neoplasms	89	11	15	6	Cancer of prostate, cancer of stomach, malignant neoplasm of head, face, and neck, secondary malignancy of respiratory organs, secondary malignant neoplasm, secondary malignant neoplasm of liver.
Neurological	46	2	1	0	
Respiratory	64	6	5	0	
Sense organ	83	1	2	0	
Symptoms	36	5	5	1	Nausea and vomiting

Table C.21: Number of Phecodes selected for each group of disease for 180-day risk of hospitalization.

	Total	# Selected by Outcome	# Selected by Treatment	# in the Intersection	Description of Phecodes in the Intersection
Circulatory system	130	10	14	2	Atrial fibrillation and flutter, congestive heart failure (CHF) NOS.
Congenital anomalies	13	0	0	0	
Dermatologic	51	4	3	1	Chronic ulcer of skin.
Digestive	96	4	7	0	
Endocrine/metabolic	95	6	6	1	Type 2 diabetes.
Genitourinary	79	12	10	0	
Hematopoietic	35	4	5	1	Lymphadenitis.
Infectious diseases	32	1	1	0	
Injuries and poisonings	72	4	2	0	
Mental disorders	42	1	6	0	
Musculoskeletal	82	2	9	0	
Neoplasms	89	12	15	6	Cancer of connective tissue, cancer of esophagus, cancer of prostate, chemotherapy, secondary malignant neoplasm, secondary malignant neoplasm of liver.
Neurological	46	5	2	0	
Respiratory	64	7	7	2	Abnormal findings examination of lungs, pneumonia
Sense organ	83	2	2	0	
Symptoms	36	6	6	2	Malaise and fatigue, nausea and vomiting.

Table C.22: Number of Phecodes selected for each group of disease for 360-day risk of hospitalization.

C.2 Supplementary Figures

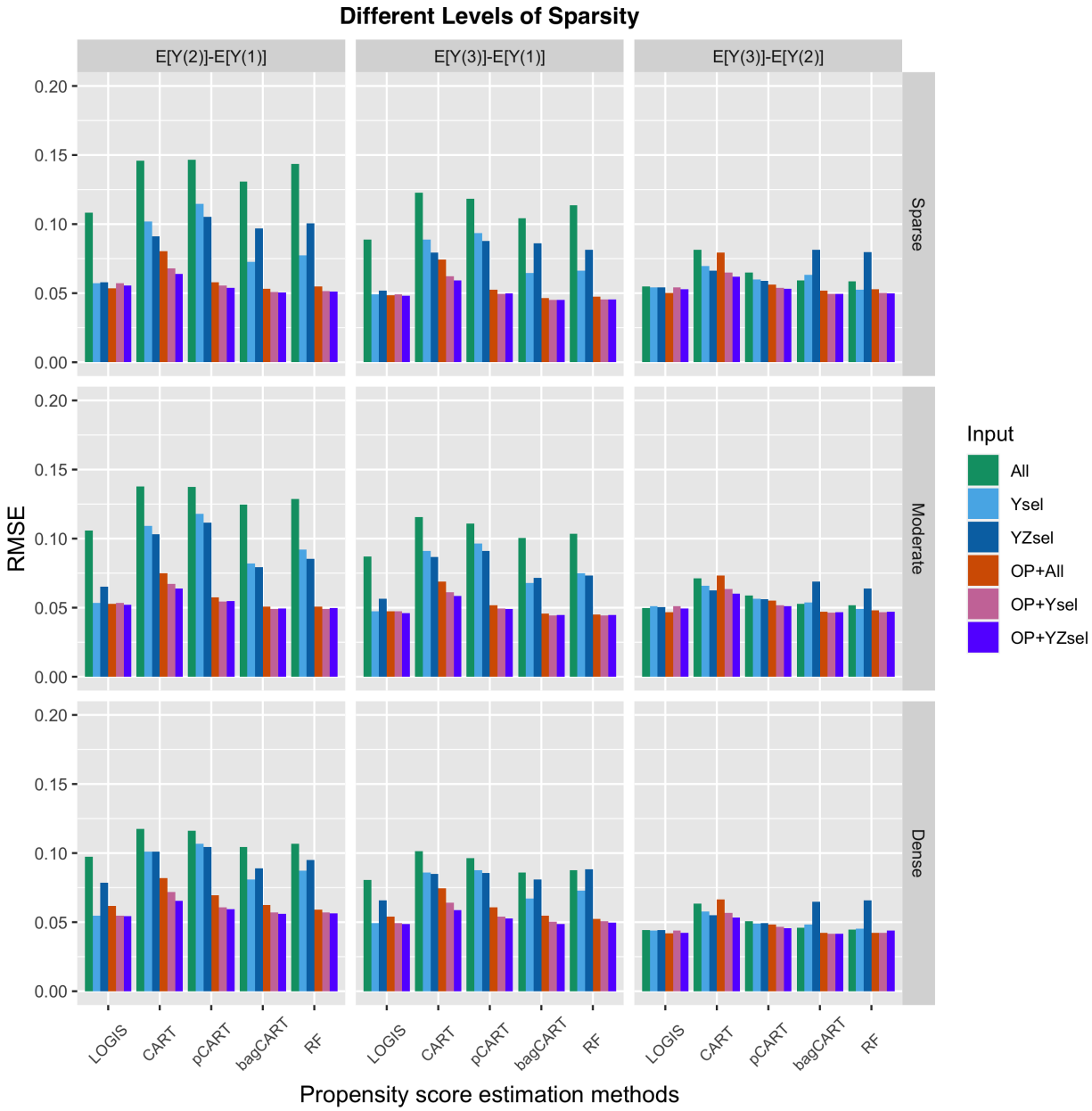

Figure C.1: RMSE for 2000 inverse probability weighted estimates for the average treatment effects under scenarios with different levels of sparsity. The rows represent scenarios and columns represent treatment pairs. Each simulated dataset contained 500 samples.

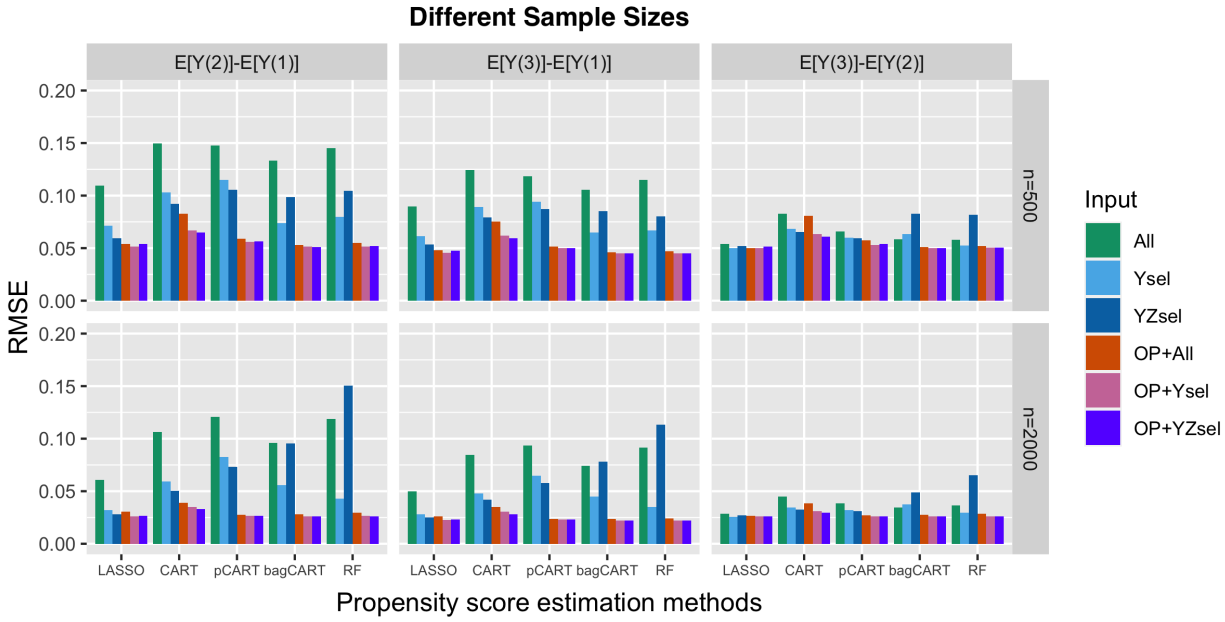

Figure C.2: RMSE for 2000 inverse probability weighted estimates for the ATE for different sample sizes. The rows represent scenarios and columns represent treatment pairs.

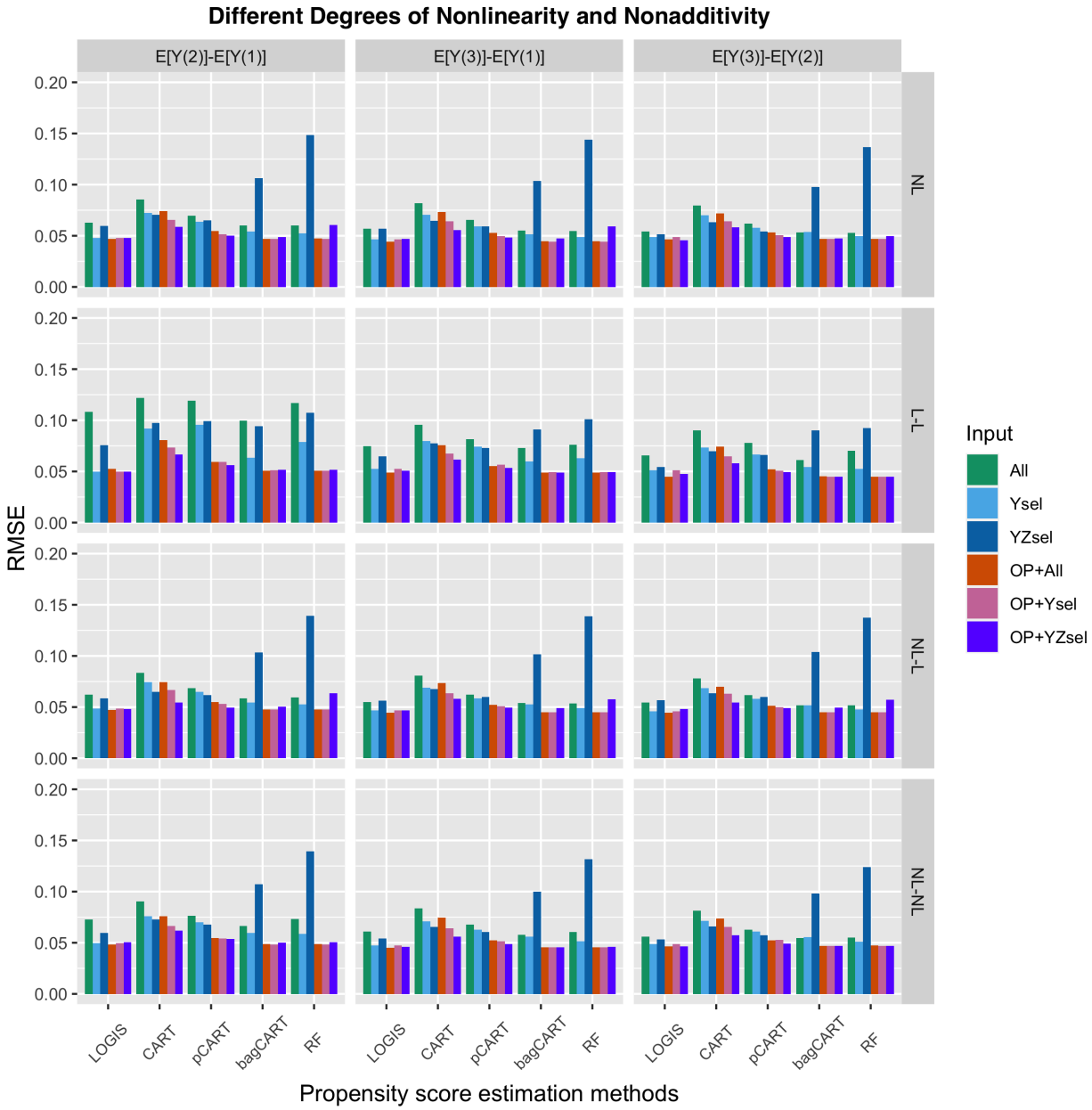

Figure C.3: RMSE for 2000 inverse probability weighted estimates for the ATE under scenarios with various degrees of nonlinearity and nonadditivity in the treatment generating model. The rows represent scenarios and columns represent treatment pairs. Each simulated dataset contained 500 samples.

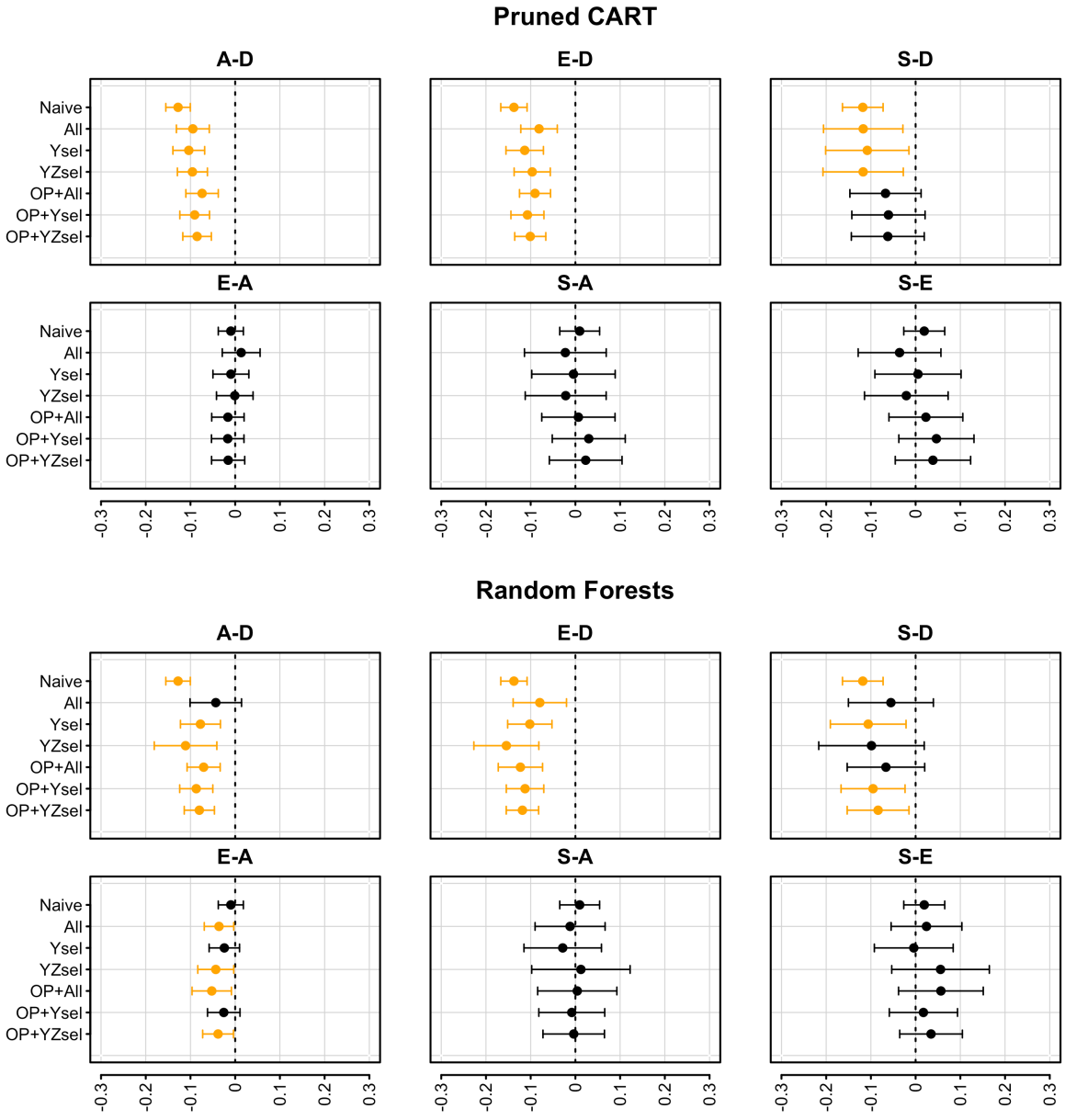

Figure C.4: Average treatment effects for ER visits within 180 days of treatment initiation for pruned CART and random forests. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7678$ ($N_A = 2757$, $N_D = 2311$, $N_E = 2043$, $N_S = 567$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

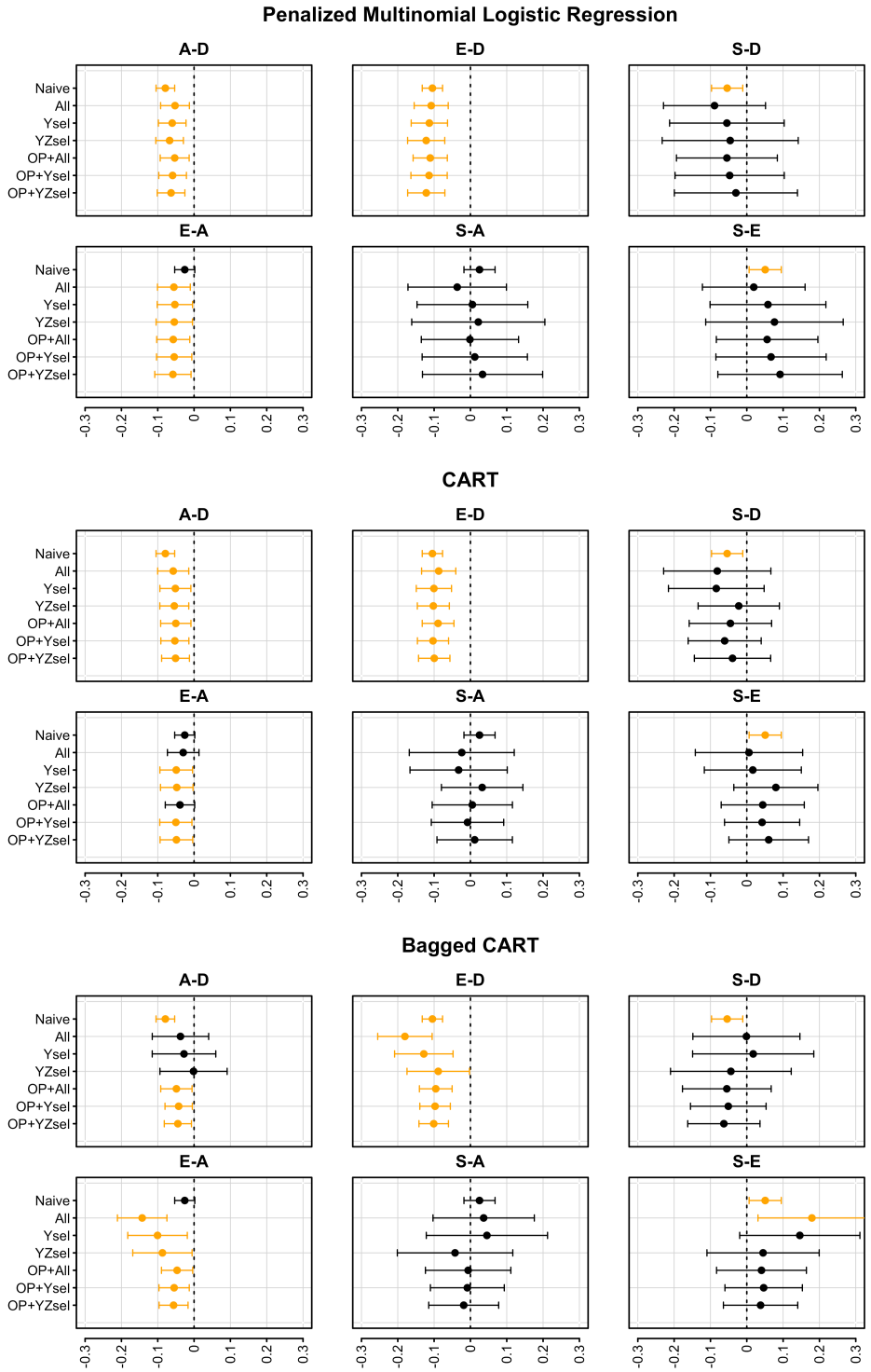

Figure C.5: Average treatment effects for ER visits within 360 days of treatment initiation for LOGIS, CART and bagged CART. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7678$ ($N_A = 2757$, $N_D = 2311$, $N_E = 2043$, $N_S = 567$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

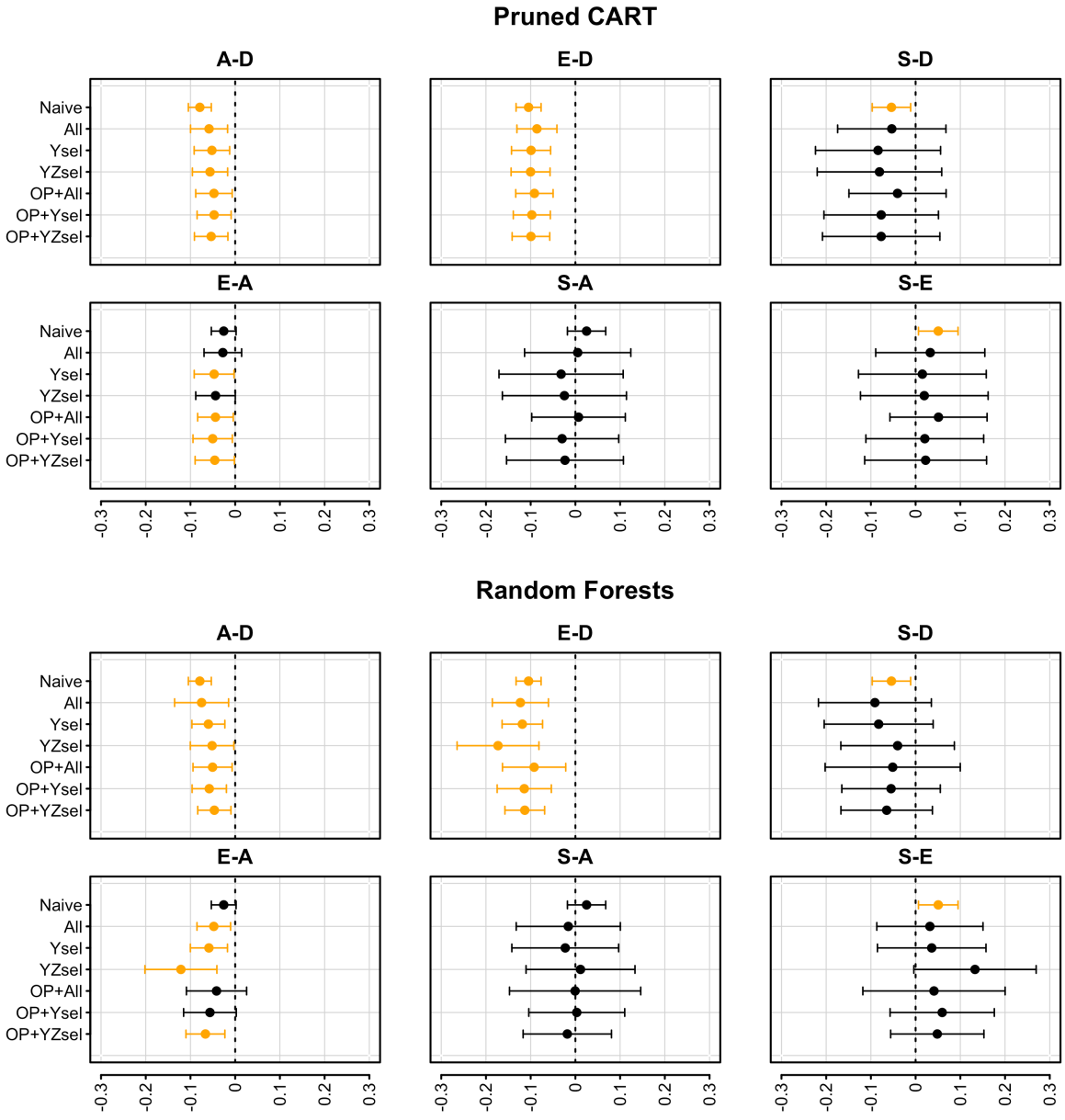

Figure C.6: Average treatment effects for ER visits within 360 days of treatment initiation for pruned CART and random forests. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7678$ ($N_A = 2757$, $N_D = 2311$, $N_E = 2043$, $N_S = 567$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

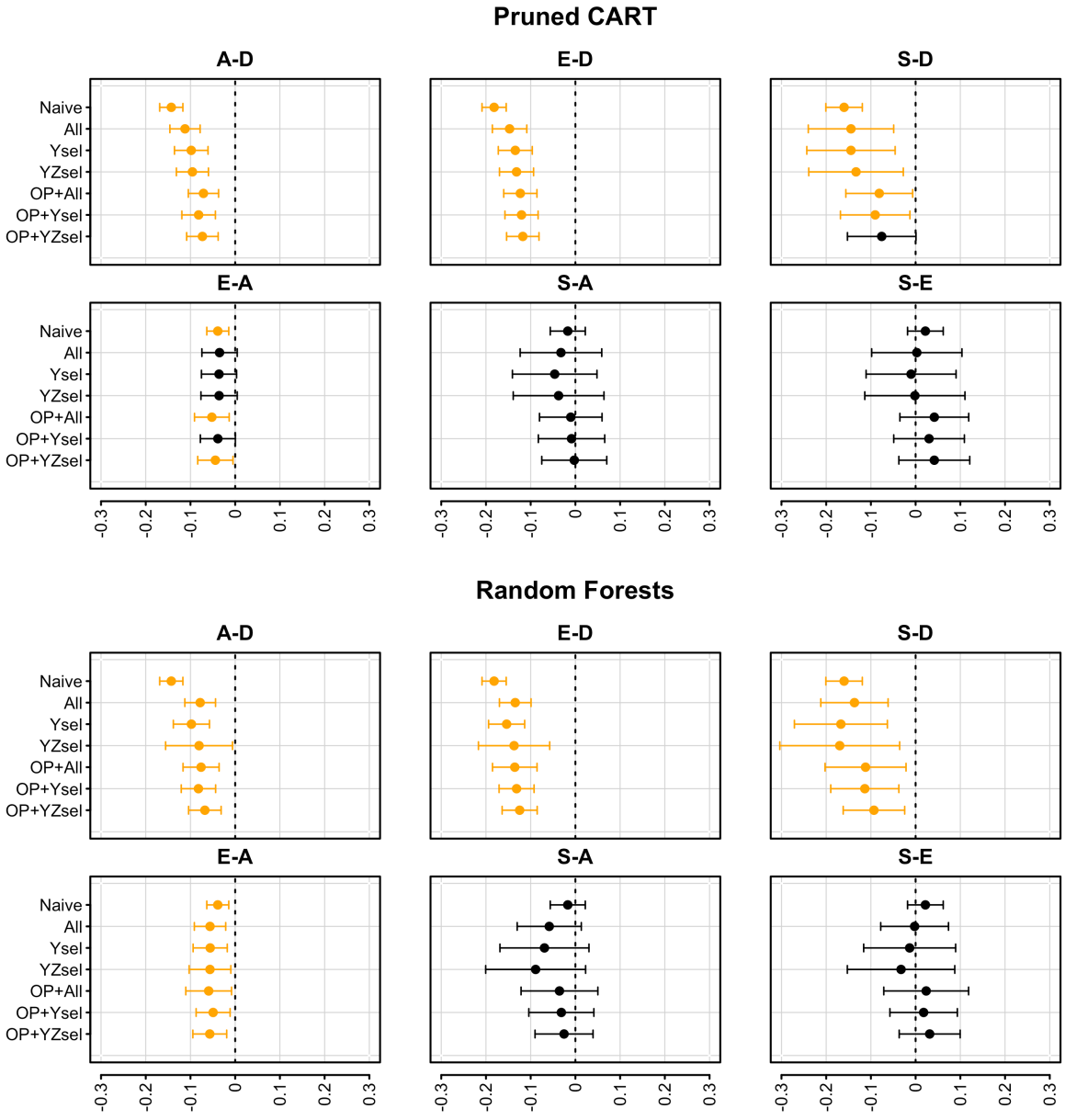

Figure C.7: Average treatment effects for hospitalization within 180 days of treatment initiation for pruned CART and random forests. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7709$ ($N_A = 2766$, $N_D = 2320$, $N_E = 2051$, $N_S = 572$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

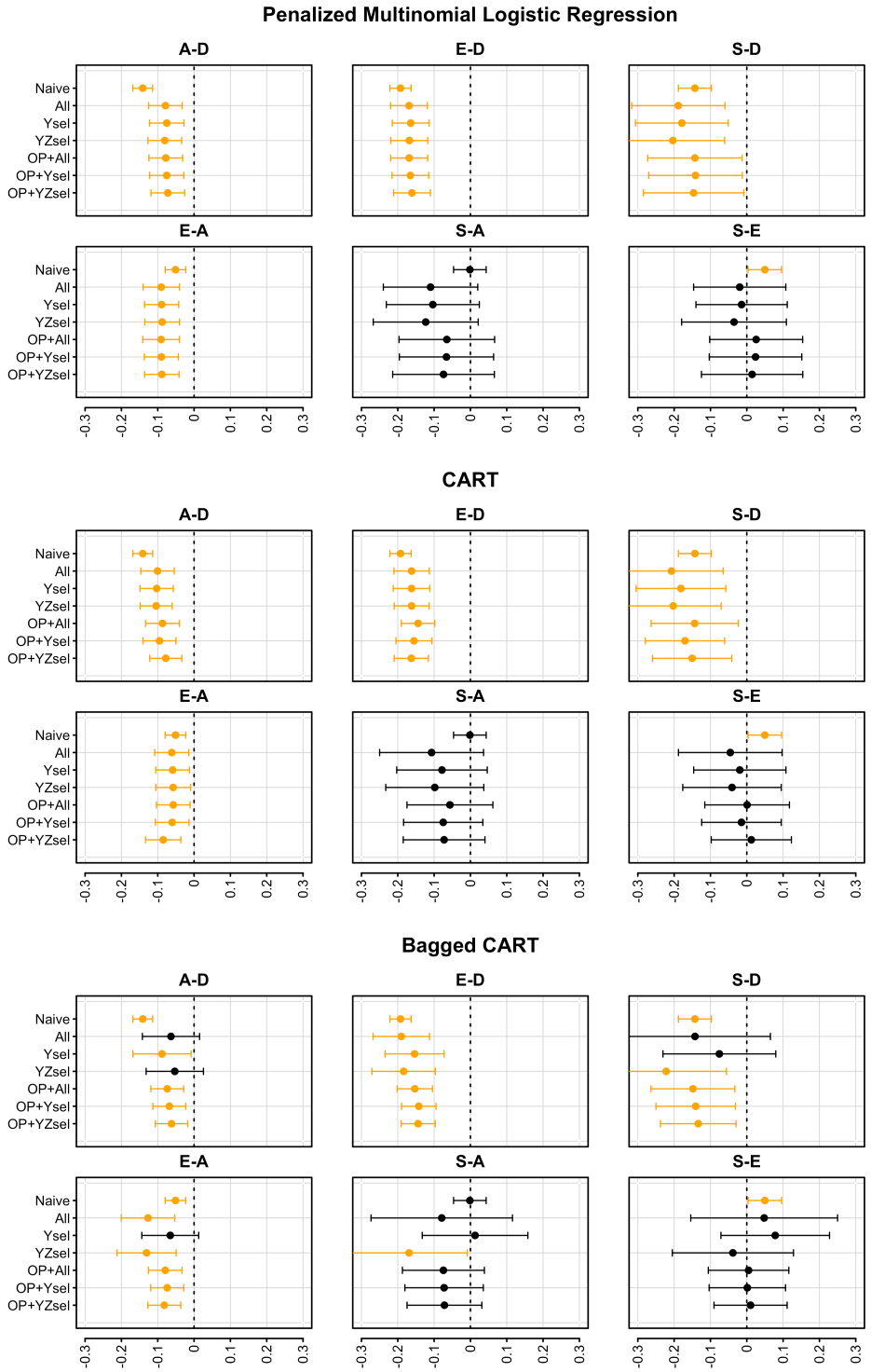

Figure C.8: Average treatment effects for hospitalization within 360 days of treatment initiation for LOGIS, CART, and bagged CART. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7709$ ($N_A = 2766$, $N_D = 2320$, $N_E = 2051$, $N_S = 572$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

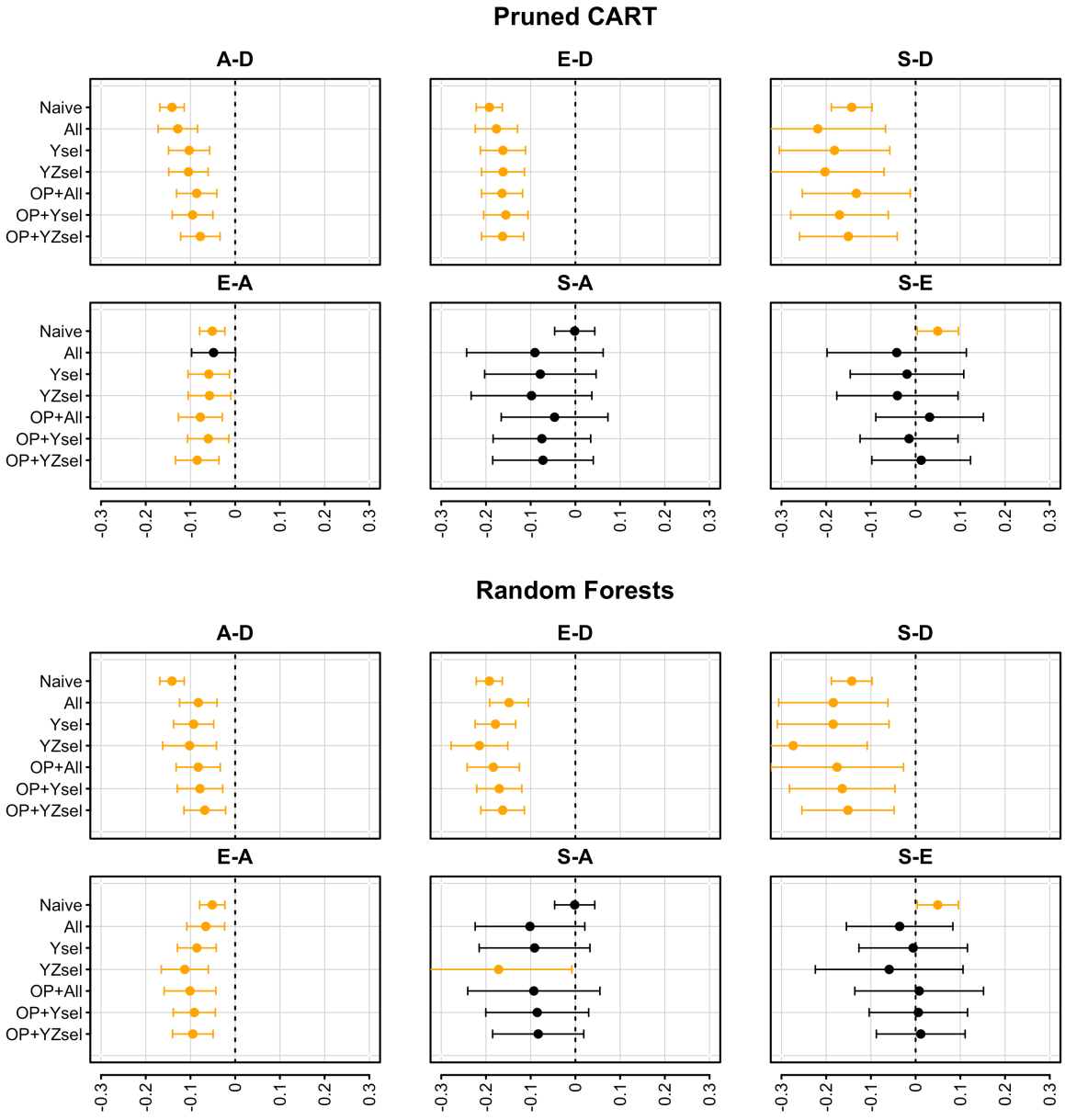

Figure C.9: Average treatment effects for hospitalization within 360 days of treatment initiation for pruned CART and random forests. Data were obtained from Optum Clinformative Data Mart. Total sample size was $N = 7709$ ($N_A = 2766$, $N_D = 2320$, $N_E = 2051$, $N_S = 572$). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T.

BIBLIOGRAPHY

[1] Harold C. Sox. Comparative Effectiveness Research: A Report From the Institute of Medicine. *Annals of Internal Medicine*, 151(3):203, August 2009. ISSN 0003-4819. doi: 10.7326/0003-4819-151-3-200908040-00125.

[2] Office of the Commissioner. Real-World Evidence. http://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence, 2019.

[3] Jared K. Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960, October 2004. ISSN 1097-0258. doi: 10.1002/sim.1903.

[4] Peter C. Austin. Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations. *Biometrical Journal*, 51(1):171–184, 2009. ISSN 1521-4036. doi: 10.1002/bimj.200810488.

[5] Seo Young Kim and Daniel H Solomon. Use of administrative claims data for comparative effectiveness research of rheumatoid arthritis treatments. *Arthritis Research & Therapy*, 13 (5):129, 2011. ISSN 1478-6354. doi: 10.1186/ar3472.

[6] Ian F. Tannock, Ronald de Wit, William R. Berry, Jozsef Horti, Anna Pluzanska, Kim N. Chi, Stephane Oudard, Christine Théodore, Nicholas D. James, Ingela Turesson, Mark A. Rosenthal, and Mario A. Eisenberger. Docetaxel plus Prednisone or Mitoxantrone plus Prednisone for Advanced Prostate Cancer. *New England Journal of Medicine*, 351(15): 1502–1512, October 2004. ISSN 0028-4793. doi: 10.1056/NEJMoa040720.

[7] Philip W. Kantoff, Celestia S. Higano, Neal D. Shore, E. Roy Berger, Eric J. Small, David F. Penson, Charles H. Redfern, Anna C. Ferrari, Robert Dreicer, Robert B. Sims, Yi Xu, Mark W. Frohlich, and Paul F. Schellhammer. Sipuleucel-T Immunotherapy for Castration-Resistant Prostate Cancer. *New England Journal of Medicine*, 363(5):411–422, July 2010. ISSN 0028-4793. doi: 10.1056/NEJMoa1001294.

[8] Johann S. de Bono, Christopher J. Logothetis, Arturo Molina, Karim Fizazi, Scott North, Luis Chu, Kim N. Chi, Robert J. Jones, Oscar B. Goodman, Fred Saad, John N. Staffurth, Paul Mainwaring, Stephen Harland, Thomas W. Flaig, Thomas E. Hutson, Tina Cheng, Helen Patterson, John D. Hainsworth, Charles J. Ryan, Cora N. Sternberg, Susan L. Ellard, Aude Fléchon, Mansoor Saleh, Mark Scholz, Eleni Efstathiou, Andrea Zivi, Diletta Bianchini, Yohann Loriot, Nicole Chieffo, Thian Kheoh, Christopher M. Haqq, and Howard I. Scher. Abiraterone and Increased Survival in Metastatic Prostate Cancer. *New*

England Journal of Medicine, 364(21):1995–2005, May 2011. ISSN 0028-4793. doi: 10.1056/NEJMoa1014618.

[9] Howard I. Scher, Karim Fizazi, Fred Saad, Mary-Ellen Taplin, Cora N. Sternberg, Kurt Miller, Ronald de Wit, Peter Mulders, Kim N. Chi, Neal D. Shore, Andrew J. Armstrong, Thomas W. Flaig, Aude Fléchon, Paul Mainwaring, Mark Fleming, John D. Hainsworth, Mohammad Hirmand, Bryan Selby, Lynn Seely, Johann S. de Bono, and AFFIRM Investigators. Increased survival with enzalutamide in prostate cancer after chemotherapy. *The New England Journal of Medicine*, 367(13):1187–1197, September 2012. ISSN 1533-4406. doi: 10.1056/NEJMoa1207506.

[10] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, April 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41.

[11] Paul R. Rosenbaum and Donald B. Rubin. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, 39(1):33–38, February 1985. ISSN 0003-1305. doi: 10.1080/00031305.1985.10479383.

[12] Liang Li and Tom Greene. A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*, 9(2):215–234, July 2013. ISSN 1557-4679. doi: 10.1515/ijb-2012-0030.

[13] Kazuki Yoshida, Sonia Hernández-Díaz, Daniel H. Solomon, John W. Jackson, Joshua J. Gagne, Robert J. Glynn, and Jessica M. Franklin. Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology (Cambridge, Mass.)*, 28(3):387–395, May 2017. ISSN 1531-5487. doi: 10.1097/EDE.0000000000000627.

[14] Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400, January 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2016.1260466.

[15] Paul R. Rosenbaum and Donald B. Rubin. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516–524, 1984. ISSN 0162-1459. doi: 10.2307/2288398.

[16] S. Vansteelandt and R. M. Daniel. On regression adjustment for the propensity score. *Statistics in Medicine*, 33(23):4053–4072, October 2014. ISSN 1097-0258. doi: 10.1002/sim.6207.

[17] Tingting Zhou, Michael R. Elliott, and Roderick J. A. Little. Penalized Spline of Propensity Methods for Treatment Comparison. *Journal of the American Statistical Association*, 114 (525):1–19, January 2019. ISSN 0162-1459. doi: 10.1080/01621459.2018.1518234.

[18] Bijan J Borah, James P Moriarty, William H Crown, and Jalpa A Doshi. Applications of propensity score methods in observational comparative effectiveness and safety research:

Where have we come and where should we go? *Journal of Comparative Effectiveness Research*, 3(1):63–78, November 2013. ISSN 2042-6305. doi: 10.2217/cer.13.89.

[19] François Laliberté, Michel Cloutier, Winnie W. Nelson, Craig I. Coleman, Dominic Pilon, William H. Olson, C. V. Damaraju, Jeffrey R. Schein, and Patrick Lefebvre. Real-world comparative effectiveness and safety of rivaroxaban and warfarin in nonvalvular atrial fibrillation patients. *Current Medical Research and Opinion*, 30(7):1317–1325, July 2014. ISSN 0300-7995. doi: 10.1185/03007995.2014.907140.

[20] Prasanna Sooriakumaran, Tommy Nyberg, Olof Akre, Leif Haendler, Inge Heus, Mats Olsson, Stefan Carlsson, Monique J. Roobol, Gunnar Steineck, and Peter Wiklund. Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: Observational study of mortality outcomes. *BMJ*, 348:g1502, February 2014. ISSN 1756-1833. doi: 10.1136/bmj.g1502.

[21] Guido W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, September 2000. ISSN 0006-3444. doi: 10.1093/biomet/87.3.706.

[22] Kosuke Imai and David A. van Dyk. Causal Inference With General Treatment Regimes. *Journal of the American Statistical Association*, 99(467):854–866, September 2004. ISSN 0162-1459. doi: 10.1198/016214504000001187.

[23] Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697, 2012. ISSN 1097-0258. doi: 10.1002/sim.4168.

[24] Shu Yang, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065, 2016. ISSN 1541-0420. doi: 10.1111/biom.12505.

[25] Michael J. Lopez and Roee Gutman. Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas. *Statistical Science*, 32(3):432–454, August 2017. ISSN 0883-4237, 2168-8745. doi: 10.1214/17-STS612.

[26] Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415, December 2019. ISSN 1932-6157, 1941-7330. doi: 10.1214/19-AOAS1282.

[27] Judea Pearl. The Foundations of Causal Inference. *Sociological Methodology*, 40(1):75–149, 2010. ISSN 1467-9531. doi: 10.1111/j.1467-9531.2010.01228.x.

[28] Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1):1–21, February 2010. ISSN 0883-4237. doi: 10.1214/09-STS313.

[29] Jeremy A. Rassen, Daniel H. Solomon, Robert J. Glynn, and Sebastian Schneeweiss. Simultaneously assessing intended and unintended treatment effects of multiple treatment options: A pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety*, 20(7):675–683, 2011. ISSN 1099-1557. doi: 10.1002/pds.2121.

[30] Jeremy A. Rassen, Abhi A. Shelat, Jessica M. Franklin, Robert J. Glynn, Daniel H. Solomon, and Sebastian Schneeweiss. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology (Cambridge, Mass.)*, 24(3):401–409, May 2013. ISSN 1531-5487. doi: 10.1097/EDE.0b013e318289dedf.

[31] Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, 2006. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x.

[32] Zhanglin L Cui, Lisa M Hess, Robert Goodloe, and Doug Faries. Application and comparison of generalized propensity score matching versus pairwise propensity score matching. *Journal of Comparative Effectiveness Research*, 7(9):923–934, June 2018. ISSN 2042-6305. doi: 10.2217/cer-2018-0030.

[33] Xiaoning He, Yumei Wang, Hongliang Cong, Chengzhi Lu, and Jing Wu. Impact of Optimal Medical Therapy at Discharge on 1-year Direct Medical Costs in Patients with Acute Coronary Syndromes: A Retrospective, Observational Database Analysis in China. *Clinical Therapeutics*, February 2019. ISSN 1879-114X. doi: 10.1016/j.clinthera.2019.01.005.

[34] Kiran Gupta, Jeffrey Trocio, Allison Keshishian, Qisu Zhang, Oluwaseyi Dina, Jack Mardekian, Lisa Rosenblatt, Xianchen Liu, Shalini Hede, Anagha Nadkarni, and Tom Shank. Real-World Comparative Effectiveness, Safety, and Health Care Costs of Oral Anticoagulants in Nonvalvular Atrial Fibrillation Patients in the U.S. Department of Defense Population. *Journal of Managed Care & Specialty Pharmacy*, 24(11):1116–1127, November 2018. ISSN 2376-1032. doi: 10.18553/jmcp.2018.17488.

[35] Shervin M. Shirvani, Jing Jiang, Joe Y. Chang, James W. Welsh, Daniel R. Gomez, Stephen Swisher, Thomas A. Buchholz, and Benjamin D. Smith. Comparative Effectiveness of 5 Treatment Strategies for Early-Stage Non-Small Cell Lung Cancer in the Elderly. *International Journal of Radiation Oncology*Biology*Physics*, 84(5):1060–1070, December 2012. ISSN 0360-3016. doi: 10.1016/j.ijrobp.2012.07.2354.

[36] Laura Mauri, Treacy S. Silbaugh, Pallav Garg, Robert E. Wolf, Katya Zelevinsky, Ann Lovett, Manu R. Varma, Zheng Zhou, and Sharon-Lise T. Normand. Drug-Eluting or Bare-Metal Stents for Acute Myocardial Infarction. *New England Journal of Medicine*, 359(13): 1330–1342, September 2008. ISSN 0028-4793. doi: 10.1056/NEJMoa0801485.

[37] Leonard A Stefanski and Dennis D Boos. The Calculus of M-Estimation. *The American Statistician*, 56(1):29–38, February 2002. ISSN 0003-1305. doi: 10.1198/000313002753631330.

[38] Huzhang Mao, Liang Li, and Tom Greene. Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research*, page 0962280218781171, June 2018. ISSN 0962-2802. doi: 10.1177/0962280218781171.

[39] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994. ISSN 0162-1459. doi: 10.1080/01621459.1994.10476818.

[40] Marshall M Joffe, Thomas R Ten Have, Harold I Feldman, and Stephen E Kimmel. Model Selection, Confounder Control, and Marginal Structural Models. *The American Statistician*, 58(4):272–279, November 2004. ISSN 0003-1305. doi: 10.1198/000313004X5824.

[41] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 1939-2176(Electronic),0022-0663(Print). doi: 10.1037/h0037350.

[42] Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189, 2003. ISSN 1468-0262. doi: 10.1111/1468-0262.00442.

[43] Donald B. Rubin. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. ISSN 0162-1459. doi: 10.2307/2287653.

[44] Joseph D. Y. Kang and Joseph L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539, November 2007. ISSN 0883-4237, 2168-8745. doi: 10.1214/07-STS227.

[45] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12027.

[46] Christian Fong, Marc Ratkovic, Kosuke Imai, Chad Hazlett, Xiaolin Yang, and Sida Peng. CBPS: Covariate Balancing Propensity Score, March 2019.

[47] Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, March 2009. ISSN 0006-3444. doi: 10.1093/biomet/asn055.

[48] Alberto Abadie and Guido W. Imbens. Matching on the Estimated Propensity Score. *Econometrica*, 84(2):781–807, March 2016. ISSN 1468-0262. doi: 10.3982/ECTA11293.

[49] M. P. Wand. Smoothing and mixed models. *Computational Statistics*, 18(2):223–249, July 2003. ISSN 1613-9658. doi: 10.1007/s001800300142.

[50] Simon Wood. Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, November 2019.

[51] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Ltd, first edition, 1987. doi: 10.1002/9780470316696.

[52] Jasjeet S. Sekhon. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software*, 42(1): 1–52, June 2011. ISSN 1548-7660. doi: 10.18637/jss.v042.i07.

[53] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15:199–236, 2007.

[54] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(1):1–28, June 2011. ISSN 1548-7660. doi: 10.18637/jss.v042.i08.

[55] Tianhui Zhou, Guangyu Tong, Fan Li, Laine Thomas, and Fan Li. PSweight: Propensity Score Weighting for Causal Inference with Observational Studies and Randomized Trials, September 2020.

[56] Peter C. Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29(20):2137–2148, September 2010. ISSN 1097-0258. doi: 10.1002/sim.3854.

[57] Megan E. V. Caram, Jason P. Estes, Jennifer J. Griggs, Paul Lin, and Bhramar Mukherjee. Temporal and geographic variation in the systemic treatment of advanced prostate cancer. *BMC cancer*, 18(1):258, March 2018. ISSN 1471-2407. doi: 10.1186/s12885-018-4166-3.

[58] Megan E. V. Caram, Ryan Ross, Paul Lin, and Bhramar Mukherjee. Factors Associated With Use of Sipuleucel-T to Treat Patients With Advanced Prostate Cancer. *JAMA network open*, 2(4):e192589, April 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.2589.

[59] Megan E. V. Caram, Shikun Wang, Phoebe Tsao, Jennifer J. Griggs, David C. Miller, Brent K. Hollenbeck, Paul Lin, and Bhramar Mukherjee. Patient and Provider Variables Associated with Variation in the Systemic Treatment of Advanced Prostate Cancer. *Urology Practice*, January 2019.

[60] Kazuki Yoshida, Daniel H. Solomon, Sebastien Haneuse, Seoyoung C. Kim, Elisabetta Patorno, Sara K. Tedeschi, Houchen Lyu, Jessica M. Franklin, Til Stürmer, Sonia Hernández-Díaz, and Robert J. Glynn. Multinomial Extension of Propensity Score Trimming Methods: A Simulation Study. *American Journal of Epidemiology*, 188(3):609–616, March 2019. ISSN 1476-6256. doi: 10.1093/aje/kwy263.

[61] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science*, 22(4):544–559, November 2007. ISSN 0883-4237, 2168-8745. doi: 10.1214/07-STS227D.

[62] Daniel F. McCaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F. Burgette. A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in medicine*, 32(19):3388–3414, August 2013. ISSN 0277-6715. doi: 10.1002/sim.5753.

[63] Liangyuan Hu, Chenyang Gu, Michael Lopez, Jiayi Ji, and Juan Wisnivesky. Estimation of causal effects of multiple treatments in observational studies with a binary outcome:. *Statistical Methods in Medical Research*, May 2020. doi: 10.1177/0962280220921909.

[64] K. John McConnell and Stephan Lindner. Estimating treatment effects with machine learning. *Health Services Research*, 54(6):1273–1282, 2019. ISSN 1475-6773. doi: 10.1111/1475-6773.13212.

[65] Megan S. Schuler and Sherri Rose. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American Journal of Epidemiology*, 185(1):65–73, January 2017. ISSN 0002-9262. doi: 10.1093/aje/kww165.

[66] Xuan Wang, Lauren A. Beste, Marissa M. Maier, and Xiao-Hua Zhou. Double robust estimator of average causal treatment effect for censored medical cost data. *Statistics in Medicine*, 35(18):3101–3116, August 2016. ISSN 1097-0258. doi: 10.1002/sim.6876.

[67] Min Zhang and Douglas E. Schaubel. Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics*, 68(4):999–1009, December 2012. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2012.01759.x.

[68] Kevin J. Anstrom and Anastasios A. Tsiatis. Utilizing Propensity Scores to Estimate Causal Treatment Effects with Censored Time-Lagged Data. *Biometrics*, 57(4):1207–1218, December 2001. ISSN 0006341X. doi: 10.1111/j.0006-341X.2001.01207.x.

[69] Youfei Yu, Min Zhang, Xu Shi, Megan E. V. Caram, Roderick J. A. Little, and Bhramar Mukherjee. A comparison of parametric propensity score-based methods for causal inference with multiple treatments and a binary outcome. *Statistics in Medicine*, January 2021. ISSN 1097-0258. doi: 10.1002/sim.8862.

[70] Per K. Andersen, Elisavet Syriopoulou, and Erik T. Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in Medicine*, 36(17):2669–2681, July 2017. ISSN 1097-0258. doi: 10.1002/sim.7297.

[71] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer-Verlag, New York, 2006. ISBN 978-0-387-32448-7. doi: 10.1007/0-387-37345-4.

[72] Daniel F. Heitjan and Donald B. Rubin. Ignorability and Coarse Data. *The Annals of Statistics*, 19(4):2244–2253, December 1991. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176348396.

[73] Richard D. Gill, Mark J. van der Laan, and James M. Robins. Coarsening at Random: Characterizations, Conjectures, Counter-Examples. In D. Y. Lin and T. R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, Lecture Notes in Statistics,

pages 255–294, New York, NY, 1997. Springer US. ISBN 978-1-4684-6316-3. doi: 10.1007/978-1-4684-6316-3_14.

[74] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1972.tb00899.x.

[75] George Casella and Roger L. Berger. *Statistical Inference*. Thomson Learning, 2002. ISBN 978-0-534-24312-8.

[76] D. Y. Lin and L. J. Wei. The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, 84(408):1074–1078, December 1989. ISSN 0162-1459. doi: 10.1080/01621459.1989.10478874.

[77] Denni D. Boos and L. A. Stefanski. M-estimation (estimating equations). In *Essential Statistical Inference: Theory and Methods*, pages 297–337. Springer New York, New York, NY, 2013. ISBN 978-1-4614-4818-1. doi: 10.1007/978-1-4614-4818-1_7.

[78] Maja Pohar Perme, Mette Gerster, and Kevin Rodrigues. Pseudo: Computes Pseudo-Observations for Modeling, July 2017.

[79] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005. ISSN 1097-0258. doi: 10.1002/sim.2059.

[80] Senol Tonyali, Hakan Bahadir Haberal, and Emrullah Sogutdelen. Toxicity, Adverse Events, and Quality of Life Associated with the Treatment of Metastatic Castration-Resistant Prostate Cancer. *Current Urology*, 10(4):169–173, November 2017. ISSN 1661-7649. doi: 10.1159/000447176.

[81] Susan M. Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017. ISSN 1541-0420. doi: 10.1111/biom.12679.

[82] Susan Athey, Guido W. Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018. ISSN 1467-9868. doi: 10.1111/rssb.12268.

[83] Soko Setoguchi, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555, June 2008. ISSN 1099-1557. doi: 10.1002/pds.1555.

[84] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, February 2010. ISSN 0277-6715. doi: 10.1002/sim.3782.

[85] Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4):512–522, July 2009. ISSN 1531-5487. doi: 10.1097/EDE.0b013e3181a663cc.

[86] Cheng Ju, Mary Combs, Samuel D. Lendle, Jessica M. Franklin, Richard Wyss, Sebastian Schneeweiss, and Mark J. van der Laan. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *Journal of Applied Statistics*, 46(12):2216–2236, September 2019. ISSN 0266-4763. doi: 10.1080/02664763.2019.1582614.

[87] Chunhao Tu. Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation*, 89(4):708–719, March 2019. ISSN 0094-9655. doi: 10.1080/00949655.2019.1571059.

[88] Xavier De Luna, Ingeborg Waernbaum, and Thomas S. Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, December 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr041.

[89] M. Alan Brookhart, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, June 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj149.

[90] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, December 2006. ISSN 0162-1459. doi: 10.1198/016214506000000735.

[91] Cheng Ju, David Benkeser, and Mark J. van der Laan. Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*, 76(1):109–118, 2020. ISSN 1541-0420. doi: 10.1111/biom.13121.

[92] Corwin Matthew Zigler and Francesca Dominici. Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model Averaged Causal Effects. *Journal of the American Statistical Association*, 109(505):95–107, January 2014. ISSN 0162-1459. doi: 10.1080/01621459.2013.869498.

[93] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.

[94] Nicolai Meinshausen. Relaxed Lasso. *Computational Statistics & Data Analysis*, 52(1): 374–393, September 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2006.12.019.

[95] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010. ISSN 1548-7660.

[96] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC Press, January 1984. ISBN 978-0-412-04841-8.

[97] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Tree-Based Methods. In Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors, *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics, pages 327–365. Springer US, New York, NY, 2021. ISBN 978-1-07-161418-1. doi: 10.1007/978-1-0716-1418-1_8.

[98] Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and Dana C. Crawford. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, May 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq126.

[99] Lucie-Marie Scailteux, Fabien Despas, Frédéric Balusson, Boris Campillo-Gimenez, Romain Mathieu, Sébastien Vincendeau, André Happe, Emmanuel Nowak, Sandrine Kerbrat, and Emmanuel Oger. Hospitalization for adverse events under abiraterone or enzalutamide exposure in real-world setting: A French population-based study on prostate cancer patients. *British Journal of Clinical Pharmacology*, July 2021. ISSN 1365-2125. doi: 10.1111/bcp.14972.

[100] Jenny Häggström and Xavier de Luna. Targeted smoothing parameter selection for estimating average causal effects. *Computational Statistics*, 29(6):1727–1748, December 2014. ISSN 1613-9658. doi: 10.1007/s00180-014-0515-0.